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

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

Abstract

In panoptic segmentation, individual instances must be separated within semantic classes. As state-of-the-art methods rely on a pre-defined set of classes, they struggle with novel categories and out-of-distribution (OOD) data. This is particularly problematic in safety-critical applications, such as autonomous driving, where reliability in unseen scenarios is essential. We address the gap between outstanding benchmark performance and reliability by proposing Prior2Former (P2F), the first approach for segmentation vision transformers rooted in evidential learning. P2F extends the mask vision transformer architecture by incorporating a Beta prior for computing model uncertainty in pixel-wise binary mask assignments. This design enables high-quality uncertainty estimation that effectively detects novel and OOD objects, enabling state-of-the-art anomaly instance segmentation and open-world panoptic segmentation. Unlike most segmentation models addressing unknown classes, P2F operates without access to OOD data samples or contrastive training on void (i.e., unlabeled) classes, making it highly applicable in real-world scenarios where such prior information is unavailable. Additionally, P2F can be flexibly applied to anomaly instance and panoptic segmentation. Through comprehensive experiments on the Cityscapes, COCO, SegmentMeIfYouCan, and OoDIS datasets, P2F demonstrates state-of-the-art performance across the board.

1. Introduction

Semantic understanding is essential for any autonomous agent navigating in the environment or interacting with it. State-of-the-art methods rely on comprehensive datasets that cover a data distribution of interest [10, 12] but cannot operate beyond that, limiting their applicability in real-world settings where unknown scenarios are the norm [20]. In safety-critical applications, such as mobile robotics or autonomous driving, safe operation in environments outside the training data distribution is crucial [52]. For example, an autonomous car must detect objects on the street regardless of whether their semantic class belongs to the training set and adapt its trajectory accordingly. Consequently, *open-world* segmentation [28] and *anomaly* segmentation [44] recently emerged to study these OOD settings systematically.

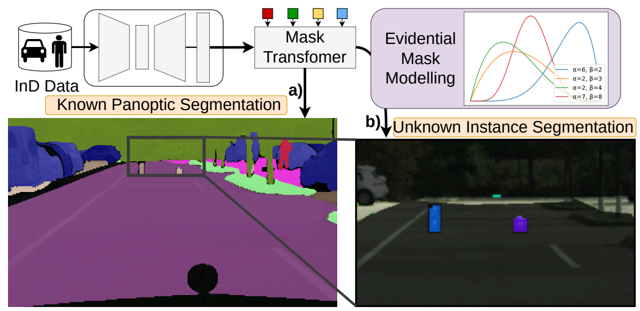

Figure 1. The proposed Prior2Former allows **a)** state-of-the-art panoptic segmentation performance and **b)** identifies unknown instances by uncertainty-based evidential learning without any knowledge beyond the standard in-domain (InD) training data.

Existing methods often approach this problem through exposure to OOD data [28] or *void* categories (i.e., unlabeled) [63] during training, or with powerful external models like CLIP [45] familiar with such OOD data [42]. Often used as black boxes, these external models can be large foundation models with extensive vocabularies that enable the wrapping methods to significantly expand the pool of objects they can handle [42]. However, assuming that their knowledge covers all possible objects is unrealistic, so they fail whenever the input lies outside of their vocabulary.

As a dense task, segmentation requires a prediction for every point or pixel of the input. Some datasets annotate pixels not containing relevant elements as *void* [14]. Prior works utilize this by learning the *void* areas, to which un-

* Equal Contribution
Corresponding author: sebastian95.schmidt@tum.de
Project page: www.cs.cit.tum.de/daml/prior2former

known classes are associated during inference [28, 56, 63]. However, they assume that the training data's *void* areas are diverse enough to represent all unknowns, which is often unrealistic, limiting their ability to handle OOD data [28]. Others model the generative process of unknowns [15].

These works assume knowledge about the distribution of unknown and OOD objects, incorporating known OOD data into the training distribution. However, datasets only represent a fraction of the real-world diversity, significantly restricting their applicability with unseen objects (i.e., OOD) in the real world [20], despite their exceptional benchmark performance [43]. In contrast, U3HS [20] segments completely unseen, unknown objects without relying on external knowledge or assumptions, but it compromises performance for known classes, creating a trade-off between known and unknown capabilities.

In this paper, we overcome these issues by proposing P2F (Fig. 1), the first vision transformer for segmentation rooted in evidential learning. It combines the Mask2Former (M2F) architecture [12] with prior networks [39] to provide reliable uncertainty estimates that enable the segmentation of unseen objects without compromising performance.

In particular, we explicitly model pixel-wise prior distributions for each binary segmentation mask query in the mask transformer architecture. The concentration parameters of this distribution, the so-called *evidence*, indicate the model's confidence and inversely correlate with its uncertainty. By learning Bayesian updates to these prior distributions, P2F is highly effective in various open-world applications. Notably, it does *neither require any exposure to OOD data* nor make *any other assumptions* about the unknown categories while maintaining strong performance on known classes. We summarize our main contributions as follows:

- We introduce P2F, the *first* evidential mask transformer that *does not require OOD data or knowledge* and quantifies uncertainty with negligible computational overhead.
- We introduce a novel loss function for training this architecture and improve existing training procedures by evidential sampling.
- We show that P2F achieves state-of-the-art performance among models that do not rely on OOD across the tasks of anomaly segmentation, open-world semantic and panoptic segmentation, and anomaly instance segmentation.

2. Related Work

Closed-World Segmentation: While early semantic segmentation models employ encoder-decoder CNN architectures [8, 9, 47], recently, transformer-based models [11, 12, 62] have leveraged attention to enhance performance. Panoptic segmentation [31] extends semantic segmentation by categorizing classes into "*stuff*" (amorphous regions) and "*things*" (countable objects), distinguishing also individual instances within semantic "*thing*" classes. Methods such as Panoptic-DeepLab [10] provide a proposal-free approach to identify the different instances, while the masked approach of Maskformer [11] or M2F [12] directly uses masks to predict instances. They leverage attention to query and assign masks to classes, predicting both instances and semantic *stuff* classes with the same architecture.

Uncertainty Estimation: Uncertainty estimation [21] is important in various domains including active learning [40, 49, 50, 53] or OOD detection [51, 61] and particularly in safety-critical applications like autonomous driving [19]. Approximate Bayesian [18] and ensemble [33] methods offer high-quality estimates but come at a high computational cost as they require multiple forward passes at inference time. Instead, sampling-free deterministic approaches estimate uncertainty in a single pass: SNGP [36] uses spectral normalized Gaussian processes, while DDU [40] fits a Gaussian Mixture Model (GMM) based uncertainty estimation of DUQ [2]. GMMSeg [34] directly incorporates a GMM into the training process. Prior networks [39, 54] rely on evidential learning and learn a conjugate prior to the predictive distribution from which uncertainty can be estimated. Posterior networks [6, 7, 57] extend this concept by learning Bayesian updates to this conjugate prior from a flow-based data density estimator. Our work builds on this principle by computing a conjugate prior for each pixel in each of the masks of the M2F architecture.

Open-set and Anomaly Segmentation: Compared to closed-set standard segmentation, where only a pre-defined set of object categories is considered, in open-world or open-set settings, models encounter unknown and novel object categories that are not part of the training data [20, 28, 51]. Open-set segmentation [28] aims to classify known classes pixel-wise while identifying unknown classes , contrasting anomaly segmentation [44], which only performs a binary distinction between known objects and unknowns. Benchmarks for anomaly segmentation include Lost & Found (L&F) [44], Fishyscapes (FS) [3], and SegmentMeIfYouCan (SMIYC) [5]. Recent works combine the open-world concept with the individual instance segmentation concept and propose open-set panoptic [28] and holistic segmentation [20] as well as the binary anomaly instance segmentation [43]. Recently, [55] proposed open-world panoptic segmentation with the PANIC benchmark, requiring distinguishing among unknown categories.

Open-set and anomaly segmentation approaches can be roughly categorized into three groups: 1) using additional models, 2) relying on external data and therefore making assumptions about OOD data, and 3) models without additional requirements. While in the context of OOD detection for classification, the use of OOD data is explicitly prohibited [64, 66], in anomaly segmentation, using OOD data or external models to ensure generalization is rather common.

1) Approaches using external foundation models can

leverage their knowledge from vast datasets with numerous classes. For example, UGainS [42] uses SAM [32] for instance-level anomaly segmentation. This paradigm implicitly relies on the foundation model to cover all potential anomaly types, even though this is unrealistic in practice, as objects unknown to the larger model would still exist.

2) Most approaches for anomaly segmentation employ OOD data or *void* regions that contain various elements that do not belong to any of the other labeled semantic classes. Here, OOD data is used by directly training the model to assign these instances to the *void* class and an additional OOD class [13, 37, 59]. Maskanomaly [1] and RbA [41] propose mask rejection after tuning on OOD data. Mask2Anomaly (M2A) [46] aims to learn an additional background mask from OOD samples. EAM [22] generates OOD artifacts through patch cutting during training, while Uno [15] generates synthetic OOD data with a normalizing flow. The use of real and synthetic OOD data during training introduces a distributional prior knowledge about the anomalies that helps the model detect them. ContMav [56] employs contrastive and objectosphere losses to *void* regions for anomaly and open-world segmentation. EOPSN [28] and DDOSP [63] re-identify unlabeled unknown categories that are within the training data's *void* regions. These methods relying on OOD data or *void* regions perform well on existing benchmarks. However, these benchmarks often satisfy the distribution assumptions that these works rely on [28]. This makes them particularly weak whenever the OOD data comes from different distributions. Therefore, *in line with the holistic segmentation setting [20], we consider any use of (pseudo-) OOD data – including post-training heuristics, external foundation models, or contrastive learning on void regions – to violate the open-set paradigm, as it leaks knowledge about OOD categories into the model.*

3) Only a few approaches are assumption-free in this regard. SML [29] achieves good unknown segmentation performance using standardized logits, while MSP [26] leverages softmax scores. U3HS [20] employs a Dirichlet Prior Network based on DeeplabV3+ and clustering of uncertain embeddings to segment and distinguish unknown instances.

In this work, we develop an uncertainty-aware transformer architecture that does not rely on assumptions about OOD data through evidential learning. While this paradigm has shown potential in previous segmentation models [11, 12, 20, 62], it remains unexplored for transformer-based segmentation. Unlike prior works, this enables our models to segment any type of unknown object while performing strongly on known categories.

3. Preliminaries

We consider a dataset of RGB images $X \in \mathbb{R}^{H\times W\times 3}$ and panoptic labels $Y \in \mathbb{N}^{H\times W\times C}$.

Mask-Based Prediction: In contrast to CNN architectures, which predict class logits for each pixel independently, mask-based architectures [11, 12] generate binary pixel-wise masks, which are then each classified into the C segmentation labels. First, a backbone extracts image features from which a pixel decoder computes per-pixel embeddings $F_E \in \mathbb{R}^{E\times H\times W}$. A transformer then uses masked attention to compute embeddings $F_M \in \mathbb{R}^{E\times N_M}$ for all N_M masks and unnormalized per-mask class assignments $\tilde{P}_M \in \mathbb{R}^{N_M\times C}$. The similarity between each pixel's embedding and the mask embeddings determines its (unnormalized) mask correspondence $\tilde{M} = F_M \cdot F_E \in \mathbb{R}^{H\times W\times N_M}$. The per-mask class assignments are normalized with softmax $P_M = \sigma_{\textbf{SM}}(\tilde{P}_M)$, while mask correspondences are normalized with a sigmoid $M = \sigma(\tilde{M})$. Finally, the per-pixel segmentation logits are obtained by weighting the per-mask class assignment for each pixel individually by its correspondence to each respective mask:

$$L(x)[h,w,c] = \sum_{i=1}^{N_M} P(x)[i,c] \cdot M(x)[h,w,i]. \quad (1)$$

The loss is calculated using mask class assignments and per-pixel correspondences. Labels are represented as l_X binary masks $Y_M \in \{0,1\}^{H\times W\times l_X}$, which correspond to class labels for each mask $Y_C \in \{1,\ldots,C\}^{l_X}$. The masks $M(x)$ are matched to the best-fitting ground-truth mask in Y_M using a Hungarian matcher. The pixel-to-mask correspondence is trained using a binary cross-entropy $\mathcal{L}_{\text{BCE}}$ and a Dice loss $\mathcal{L}_{\text{Dice}}$ [58] between $M(x)$ and Y_M. Similarly, the per-mask class assignments are trained with a cross-entropy loss $\mathcal{L}_{\text{CE}}$ between $P(x)$ and Y_C. For panoptic segmentation, the label set Y consists of one binary mask for every *stuff* class and one for each instance per image. Following M2F [12], the loss is computed only for a subset of pixels $\mathcal{P}$. Masks in $M(x)$ not matching any mask in Y_M are ignored in the mask loss, and the mask classification loss is adjusted to assign them to an additional *no-object* class.

Evidential Learning: Traditional models for semantic segmentation apply a softmax σ_{SM} over the class logits predicted for each pixel. Despite the ubiquity of this paradigm, recent work has found shortcomings in accurately representing model uncertainty due to a tendency toward overconfidence [23, 25, 30]. A promising alternative is Evidential Learning [54], which is grounded in the Dempster–Shafer Theory [16]. Rather than directly modeling a categorical distribution $p(y \mid \mathbf{x}) \sim \text{Cat}(\sigma_{\text{SM}}(L(\mathbf{x})))$ for an instance with features $\mathbf{x}$, evidential models parametrize its conjugate prior by predicting the *evidence* for each possible class. For a categorical distribution over C classes, the conjugate prior is a Dirichlet distribution parametrized by K concentration parameters $\boldsymbol{\kappa} = (\kappa_1, \kappa_2, \ldots, \kappa_C)$:

$$\text{Dir}(\boldsymbol{p}|\boldsymbol{\kappa}) = \frac{\Gamma(\sum_{k=1}^{K}\kappa_k)}{\prod_{k=1}^{K}\Gamma(\kappa_k)} \prod_{k=1}^{K} p_k^{\kappa_k - 1} \quad (2)$$

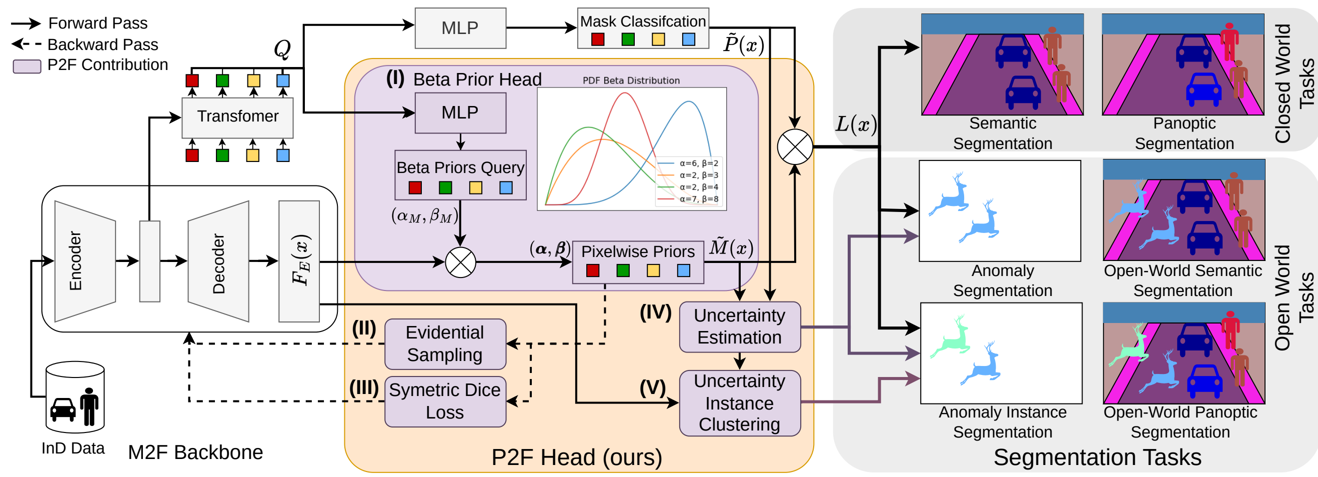

Figure 2. Overview of the Prior2Former (P2F) architecture. P2F is composed of a **beta-prior head (I)**, **evidential sampling (II)**, **symmetric Dice loss (III)**, **uncertainty estimation (IV)**, and **uncertainty instance clustering (V)**. Only standard, In-distribution data is used to train our P2F **(I)** using our novel symmetric Dice loss **(II)** and evidential sampling **(III)**. During inference, our uncertainty estimation **(IV)** detects anomalous objects, while our uncertainty instance clustering **(V)** can distinguish the instances of novel classes.

Here, Γ is the gamma function. Intuitively, the (second-order) Dirichlet is a distribution over the (first-order) categorical distributions $\boldsymbol{p}$. Predicting the conjugate prior to the predictive distribution allows a model to represent different types of uncertainty [48]. Low evidence priors indicate high model uncertainty as the corresponding first-order predictive distributions are less concentrated. Conversely, high evidence priors, i.e., concentrated first-order distributions, correspond to high confidence. This enables the model to predict high-entropy first-order distributions with high confidence to represent data-inherent uncertainty. Notably, first-order distributions alone cannot differentiate between inherent and model-related uncertainty [48], a distinction crucial for downstream anomaly-related applications [40].

4. Prior2Former

We propose to leverage the benefits of evidential learning in the context of mask transformers for segmentation to enable them to faithfully represent uncertainty by designing P2F, the first evidential mask transformer architecture. P2F is trained to predict a second-order prior to its binary per-pixel segmentation masks. This induces uncertainty-aware embeddings that can be used to tackle various closed-world and open-world tasks, including closed-world and open-world panoptic segmentation, anomaly segmentation, and anomaly instance segmentation. Remarkably, all of these tasks can be addressed with our same architecture and training framework, *without* requiring additional data or any other assumption about potential anomalies.

Depicted in Fig. 2, P2F predicts the parameters of the conjugate Beta distribution for each segmentation mask through a **beta-prior head (I)** to accurately describe model uncertainty. Additionally, we introduce a novel **evidential** uncertainty **sampling** strategy **(II)** and propose a **symmetric Dice loss (III)** to train P2F. Both innovations are explicitly tailored toward Beta distributions to increase the quality of the associated mask prediction. Lastly, we convert the predicted mask-based evidence scores into accurate **uncertainty estimates (IV)**. We finally leverage these with an **uncertainty instance clustering (V)** to distinguish between individual unknown instances.

(I) Beta-Prior Head: As described in Sec. 3, the conjugate prior in classification problems is a Dirichlet distribution $\mathrm{Dir}(\boldsymbol{p}|\boldsymbol{\kappa})$ over all possible categorical distributions. In binary classification, the Beta distribution is used as the conjugate prior to the predictive Bernoulli distribution. It is parameterized by concentration parameters $\alpha, \beta > 0$:

$$\mathrm{Beta}(y \mid \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}. \quad (3)$$

Here, Γ is the gamma function. The concentration parameters α and β can be interpreted as pseudo-counts for positive and negative observations of the corresponding Bernoulli variable. Their magnitude, therefore, serves as *evidence* and can be used for uncertainty quantification.

Instead of letting the M2F architecture predict the (unnormalized) mask assignments $F_M(x)$ directly, we extend it to estimate the corresponding concentration parameters $\boldsymbol{\alpha}_M(x), \boldsymbol{\beta}_M(x) \in \mathbb{R}^{N_M \times E}$. Similarly to how M2F computes $\tilde{M}(x)$, we use multiplication to obtain per-pixel evidences $\boldsymbol{\alpha}(x), \boldsymbol{\beta}(x) \in \mathbb{R}^{H \times W \times N_M}$:

$$\boldsymbol{\alpha}(x) = \boldsymbol{\alpha}_M(x) \cdot F_E(x) \qquad \boldsymbol{\beta}(x) = \boldsymbol{\beta}_M(x) \cdot F_E(x). \quad (4)$$

While the M2F architecture computes the soft pixel-to-mask assignments with an element-wise sigmoid as

$M(x) = \sigma(\tilde{M}(x))$, in P2F, we propose to use the expected first-order pixel-to-mask assignment instead, as:

$$M(x) = \mathbb{E}_{\mathbf{p}\sim \mathrm{B}(\boldsymbol{\alpha}(x), \boldsymbol{\beta}(x))}[\mathbf{p}] = \frac{\boldsymbol{\alpha}(x)}{\boldsymbol{\alpha}(x) + \boldsymbol{\beta}(x)}. \tag{5}$$

For each pixel, the concentration parameters α and β indicate the evidence the model assigns to the positive and negative classes, respectively. Therefore, they induce an evidence-based uncertainty score:

$$\mathbf{U}_{\text{evi}}(x) = -(\boldsymbol{\alpha}(x) + \boldsymbol{\beta}(x)). \tag{6}$$

Intuitively, for a given pixel and mask, the P2F model assigns high uncertainty only if it does not predict any evidence for either the positive or negative class.

(II) Evidential Sampling: During training, P2F predicts the concentration parameters of the Beta prior $\boldsymbol{\alpha}_M, \boldsymbol{\beta}_M$. We compute them as the similarity scores between the output of a mask query MLP and the pixel-wise representations of the decoder. By first applying a softplus function and then incrementing all values by 1, we satisfy the positivity constraint of the concentration parameters. Consequently, the resulting Beta prior is always non-degenerate, and its expected values – the mask probabilities $M(x)$ – are always in the open interval $(0, 1)$. We adapt the mask sampling procedure of M2F [12] that is outlined in Sec. 3 and represent the segmentation masks through l_X binary ground-truth binary targets $Y_M \in \{0, 1\}^{H\times W\times l_X}$. In panoptic tasks, we use an individual mask for each instance within the image and each *stuff* class. Each of the N_M masks i predicted by P2F is then matched to a mask $\hat{i}$ in Y_M. We refer to masks that were matched as *predicted masks* and index them with $\mathcal{M}(x) \subseteq \{1, \dots, N_M\}$. For each mask $i \in \mathcal{M}(x)$, only a subset of pixels $\mathcal{P}_i(x)$ is sampled to calculate the loss with respect to the matched target mask $\hat{i}$.

We utilize the evidential uncertainty of P2F to improve the training efficiency through an evidence-based sampling strategy for each mask $\mathcal{P}_i(x)$. For a given budget of pixels for which the training loss is to be computed, we assign 75% of it to pixels with the highest evidential uncertainty according to Eq. (6) and randomly select the remaining 25%. This matches the sampling of M2F [12] and steers the training to focus on regions with high evidential uncertainty. The corresponding loss maximizes the log-likelihood of the ground-truth masks under the predicted Beta prior:

$$\mathcal{L}_{\text{evi}}(Y_M, x) = \frac{1}{|\mathcal{M}(x)|} \sum_{i\in\mathcal{M}(x)} \mathcal{L}_{\text{evi}}(Y_M(x)[\dots, \hat{i}], \boldsymbol{\alpha}(x)[\dots, i], \boldsymbol{\beta}(x)[\dots, i]) \tag{7}$$

where $\mathcal{L}_{\text{evi}}(Y, \boldsymbol{\alpha}, \boldsymbol{\beta}) =$

$$\sum_{h,w\in\mathcal{P}_i(x)} -\log \mathrm{Beta}(Y[h, w] \mid \boldsymbol{\alpha}[h, w], \boldsymbol{\beta}[h, w]). \tag{8}$$

(III) Symmetric Dice Loss: The task of predicting binary segmentation masks is highly imbalanced. Therefore, we introduce a novel symmetric Dice loss to prevent updates that predominantly focus on $\boldsymbol{\beta}$:

$$\begin{aligned}\mathcal{L}_{\text{sDice}}(Y_M, x) = \frac{1}{2}\Bigg(&\mathcal{L}_{\text{Dice}}\left(\frac{\boldsymbol{\alpha}(x)}{\boldsymbol{\alpha}(x) + \boldsymbol{\beta}(x)}, Y_M\right) \\ &+ \mathcal{L}_{\text{Dice}}\left(\frac{\boldsymbol{\beta}(x)}{\boldsymbol{\alpha}(x) + \boldsymbol{\beta}(x)}, 1 - Y_M\right)\Bigg).\end{aligned} \tag{9}$$

Here, $\mathcal{L}_{\text{Dice}}$ refers to the standard Dice loss between the predicted probabilities and the true segmentation. The total loss to train P2F combines the evidential and the symmetric Dice losses with the cross-entropy loss $\mathcal{L}_{\text{CE}}$ of Sec. 3 as:

$$\mathcal{L}_{P2F} = \lambda_{CE}\mathcal{L}_{CE} + \lambda_{\text{sDice}}\mathcal{L}_{\text{sDice}} + \lambda_{\text{evi}}\mathcal{L}_{\text{evi}}. \tag{10}$$

We compute all losses on the same pixel sets $\mathcal{P}_i$. In contrast to other approaches, P2F is not explicitly trained to reject anomalous instances by assigning them to an additional *void* class. Therefore, similarly to [20], we do not require any form of OOD data or surrogates that (implicitly) make assumptions about its distribution to train P2F.

(IV) Uncertainty Estimation: We now outline how the evidence-based mask uncertainty of Eq. (6) can be combined with the mask classification $P(x)$. First, we disregard all masks $\mathcal{M}_\emptyset \subseteq \{1, \dots, N_M\}$ that are classified as *no-object*; this step is referred to as mask filtering and rejects non-primary masks from the M2F training. For each of the remaining masks, P2F represents its confidence in a pixel being assigned to the mask through the concentration $\boldsymbol{\alpha}$. The confidence with which P2F assigns a pixel to any class is therefore represented by the highest evidence over all masks. For the corresponding mask i^* (Eq. (11a)), we obtain the probability of a pixel being assigned to this mask as the expectation under the predicted Beta prior p^*_M (Eq. (11b)). We use this probability to re-weight the confidence p^*_C with which the corresponding mask is classified (Eq. (11c)). Finally, the uncertainty U associated with the segmentation of a pixel is quantified as inversely correlated with this weighted classification confidence (Eq. (11d)). Effectively, our procedure adjusts the uncertainty arising from mask classification by re-weighting it according to the evidential pixel-wise uncertainty in the predicted mask.

$$i^*(h, w) = \operatorname*{arg\,max}_{i\in\{1,\dots,N_M\}\setminus\mathcal{M}_\emptyset} \boldsymbol{\alpha}(x)[h, w, i] \tag{11a}$$

$$p^*_M(h, w) = \left(\frac{\boldsymbol{\alpha}(x)}{\boldsymbol{\alpha}(x) + \boldsymbol{\beta}(x)}\right)[h, w, i^*(h, w)] \tag{11b}$$

$$p^*_C(h, w) = \max_c \sigma_{\text{SM}}(\tilde{P}_M(x)[i^*(h, w), c]) \tag{11c}$$

$$U(h, w) = -p^*_C(h, w) \cdot p^*_M(h, w) \tag{11d}$$

(V) Uncertainty Instance Clustering: To distinguish between different uncertain instances, we utilize the embeddings from the pixel decoder F_E and our uncertainty U from

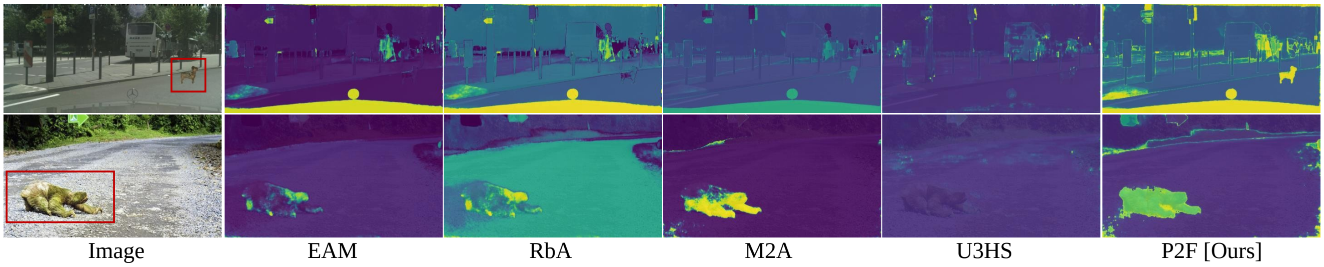

Figure 3. Qualitative comparison of uncertainty for binary anomaly segmentation on FS [3] (top) and SMIYC Road Anomaly [5] (bottom).

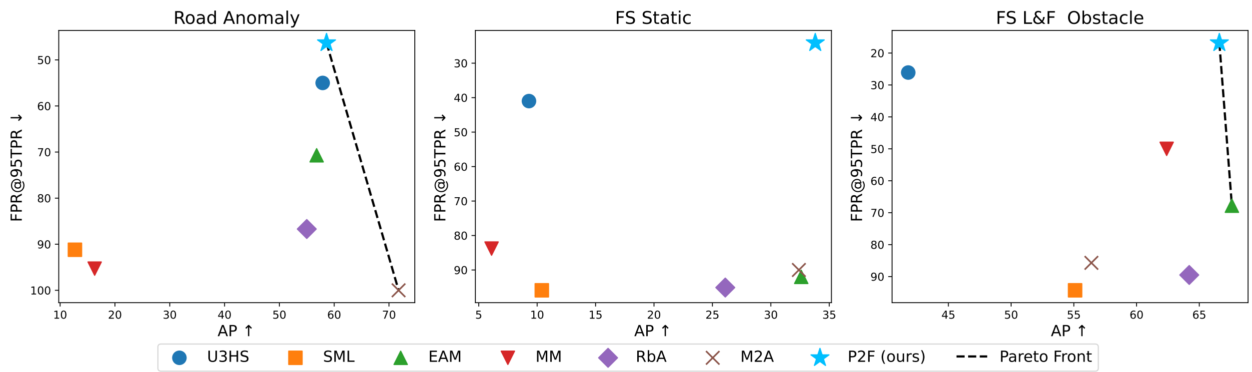

Figure 4. Anomaly Segmentation Benchmarks using the official SMIYC benchmark code [5] on the validation sets for SMIYC Road Anomaly, FS Static, and the FS L&F obstacle track. P2F shows the best trade-off between AP and FPR for all three benchmarks.

Eq. (11). By thresholding the uncertainty scores Eq. (11), we obtain candidates for anomaly segmentation. The corresponding pixel embeddings from the pixel decoder F_E are then clustered using DBSCAN [17]. This supports the instance assignment as the masks instance pairing can be unreliable for OOD data. Details on the threshold are given in Appendix A. The outlier pixels that are not assigned to any cluster are reassigned to the original semantic class and instance, respectively. In contrast to U3HS [20], which also employs clustering-based anomaly instance segmentation, P2F avoids contrastive losses, simplifying the training.

5. Experiments

We comprehensively evaluate P2F across various tasks and settings, primarily focusing on four tasks: **anomaly segmentation** and **anomaly instance segmentation** (i.e., known classes are ignored), **open-set semantic** and **open-set panoptic segmentation** (i.e., segmenting known classes, too) as in holistic segmentation [20]. To highlight that P2F addresses *all tasks with the same architecture and training procedure*, we adopt the same panoptic training strategy outlined in Sec. 3 for all experiments. Details on the experimental settings are provided in Appendix A.

Anomaly Segmentation: First, we study the anomaly segmentation capabilities of P2F by comparing ours with state-of-the-art anomaly segmentation approaches. In this binary segmentation task, anomalous objects must be identified. We evaluate on the benchmarks L&F [44], FS [3], and SMIYC [5]. To compare fairly, we train all models with the same ResNet-50 [24] backbone and use the same panoptic training procedure without OOD data. We use the official SMIYC benchmark code to compute the Average Precision (AP) and False Positive Rate (FPR) for each model. As the driving motivation for our work is *avoiding assumptions about OOD data*, we focus on baselines that do not rely on OOD data. Specifically, we employ as baselines **RbA** [41], **EAM** [22] and **M2A** [46] based on M2F and consider the general baselines **SML** [29] and **MM** as mask variants of **MSP** [27] on M2F. While **RbA** explicitly reports results for an OOD-free version, we use the OOD-free versions of **EAM** and **M2A** presented in their corresponding ablation studies. Further, we use the prior assumptions-free U3HS [20], which is based on DeeplabV3+ [9]. Finally, although ContMAV [56] shows promising results in anomaly segmentation, it relies on contrastive supervision on *void* classes during training (which we exclude), so it is omitted here. Implementation details are provided in Appendix A.1.

In Fig. 3, we compare the pixel-wise uncertainty scores predicted by different methods. While EAM, RbA, and M2A assign high uncertainties to parts of the animal only in the bottom sample, P2F successfully identifies the OOD animals in both inputs. In Fig. 4, we display the results for SMIYC Road Anomaly, FS Static, and L&F obstacle track of FS, highlighting the trade-off between AP and FPR. P2F consistently balances the two metrics favorably while outperforming in both metrics for FS. For L&F, EAM performs slightly better than P2F in terms of AP, but significantly

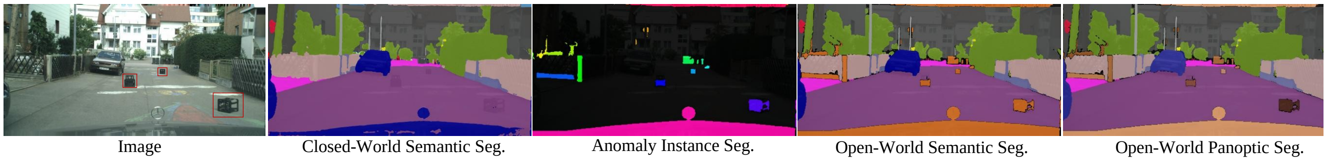

Figure 5. P2F model prediction on L&F [44], part of the OoDIS [43]. Red boxes mark unseen instances. P2F effectively distinguished unknown anomalies, even the ones outside the L&F region of interest (i.e., road), such as the trash cans at the end of the street.

Method	No Extra Models	No OOD Data	L&F AP ↑	L&F AP50 ↑	Anomaly AP ↑	Anomaly AP50 ↑	Obstacles AP ↑	Obstacles AP50 ↑	Mean AP ↑	Mean AP50 ↑
UGainS [42]	✗	✗	27.14	45.82	11.42	19.15	27.22	46.54	25.19	42.81
M2A [46]	✓	✗	11.73	23.64	4.78	9.03	17.23	28.44	13.73	24.3
U3HS [20]	✓	✓	0.19	0.73	0	0	0.22	0.62	0.19	0.58
P2F [ours]	✓	✓	**5.60**	**14.55**	**0.59**	**1.25**	**2.25**	**6.88**	**3.21**	**8.85**

Table 1. Comparison of anomaly instance segmentation methods on the **official OoDIS**[43] **benchmark**, including the three subsets L&F, Road Anomaly and Road Obstacles. Best scores without additional requirements are in bold.

worse in FPR. For SMIYC Road Anomaly, M2A realizes a good AP at the cost of a very high FPR. These results underline the effectiveness of our uncertainty estimation in the context of anomaly segmentation. Since P2F detects instances, in the following, we focus on the newer and more challenging anomaly **instance** segmentation task.

Anomaly Instance Segmentation: We further evaluate P2F on anomaly instance segmentation by submitting to the recent OoDIS benchmark [43]. OoDIS extends the previously used L&F [44], SMIYC [5] Road Anomaly, and Road Obstacles track by adding instance annotations, allowing a more complex evaluation. Because of its small and diverse objects, we consider this task as the most challenging.

In Tab. 1, we report the results of the official OoDIS benchmark evaluation compared to other works listed on the benchmark's official website. While M2A and UGainS show strong performances, they rely on knowledge about OOD data and additional external models to specifically target the anomalies that the benchmark evaluates. Together with P2F, U3HS [20] is the only one that does not rely on any sort of OOD data. P2F outperforms it by a great margin. In Fig. 5, we show a prediction of P2F on a L&F sample of OoDIS. Besides excellent closed-world performance, P2F detects multiple anomalous obstacles, including the trash bins that are not part of the Cityscapes semantic classes on which the model was trained. Additionally, panoptic L&F results are shown in Appendix C. In Appendix B.4, we supply the results of the OoDIS benchmark on the publicly available validation set comprising 100 L&F samples of the FS [3] split. In this setting, we again focus on U3HS [20] as it does not require assumptions about the OOD data distribution. Still, P2F demonstrates large margins.

Open-Set Segmentation: Lastly, we evaluate open-set panoptic segmentation. We provide an evaluation of open-set semantic segmentation in Appendices B.1 and B.2. In the panoptic setting, on top of the standard, closed-set panoptic segmentation [31], unknown areas must be segmented as such, and each unknown instance must be distinguished by individual IDs. Here, the number of available baselines is limited, especially when requiring models that do not rely on prior assumptions about OOD data, e.g., by explicitly training on the *void* class. EOSPN [28] and DDOPS[63] target this panoptic setting but require the *void* class supervision on the unknown classes. U3HS [20] generalizes to all evaluated tasks and is again the only other method without any OOD assumptions.

In Tab. 2, we evaluate P2F on COCO using as unseen, unknown the commonly 16 left-out classes [20, 28, 63] (see Appendix A.3). As evaluation metrics in open-world panoptic segmentation, we consider Panoptic Quality (PQ), Segmentation Quality (SQ), and Recognition Quality (RQ) scores for known and unseen classes, which do not include

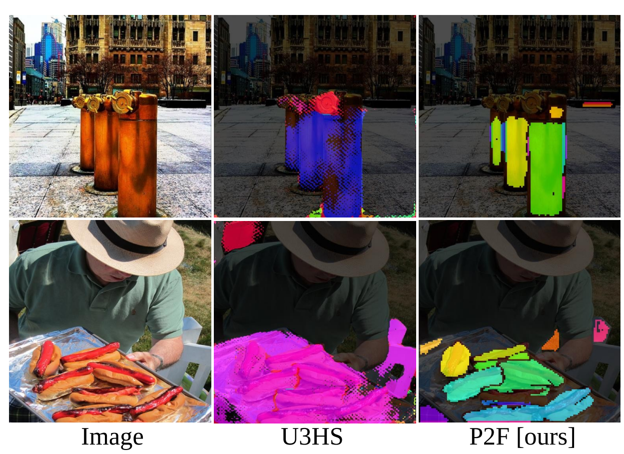

Figure 6. Comparison of anomaly instance segmentation on unseen instances on COCO [35].

Method	No *void*	Closed-World PQ↑	SQ ↑	RQ↑	Unseen Classes PQ↑	SQ ↑	RQ↑
EOPSN	✗	36.50	76.32	45.06	0.01	74.55	0.01
DDOPS	✗	37.93	77.28	46.80	8.39	81.52	10.29
U3HS [20]	✓	22.03	69.03	27.92	9.62	72.84	13.20
P2F [ours]	✓	**41.34**	**78.60**	**50.79**	**10.71**	**79.96**	**13.39**

Table 2. Panoptic comparison on the COCO val. set [35] in closed-world settings and on the unseen, unknown class at full resolution.

Method	Closed-World PQ↑	SQ ↑	RQ↑	Open-World PQ ↑	SQ ↑	RQ↑
U3HS [20]	46.53	78.87	58.99	41.21	79.77	51.67
P2F [ours]	**59.40**	**80.78**	**72.34**	**45.25**	**80.69**	**54.68**

Table 3. Closed- and open-world evaluation on Cityscapes [14].

the set of unknown classes used for *void* supervision of EOPSN and DDOPS.

While EOPSN and DDOPS leverage additional training *void* classes to predict the left-out classes as *void*, the uncertainty-based approaches P2F and U3HS can eclipse both of them in PQ and RQ, but not on SQ. Nevertheless, our P2F outperforms U3HS in all three metrics, further showcasing its performance.

In Fig. 6, we show predictions on COCO [35] for unseen classes of P2F and U3HS. For the top image, U3HS correctly identifies the hydrants as unknown, but fails to separate them into instances. In contrast, P2F distinguishes well single objects despite their similarity and overlap. In the lower image, U3HS provides a single instance for the whole tray (unknown, too), while P2F manages to segment the individual unknown hotdog instances within the tray. In general, the P2F manages to extract more fine-grained structures compared to U3HS. Further results and visual examples are presented in Appendices B.2 and C.

In Tab. 3, we report results on the Cityscapes open- and closed-world setting from [20], where we investigate how much the uncertainty thresholding influences the prediction in the absence of any anomalies. In Tab. 4, we further evaluate on L&F in the setting introduced by [20]. For

Method	PQ ↑	SQ ↑	RQ ↑
U3HS [20]	7.94	64.11	12.37
RbA [41] + P2F post-proc.	7.30	71.85	10.17
EAM [22] + P2F post-proc.	8.79	72.09	12.20
M2A [46] + P2F post-proc.	9.91	73.45	13.49
P2F [ours]	**11.22**	**74.47**	**15.06**

Table 4. Panoptic on the unseen L&F classes in the setting of [20] after training on Cityscapes [14]. Semantic works need P2F.

Mask Filtering	(III) Sym. Dice	(II) Evid. Sampl.	Road Anomaly AP ↑	FPR95 ↓	FS L&F Obstacle AP ↑	FPR95 ↓
✗	✓	✓	40.5	46.4	66.4	**16.8**
✗	✗	✓	31.7	62.3	18.5	28.4
✓	✗	✓	53.1	72.7	45.3	95.9
✓	✓	✗	47.2	58.9	23.1	22.2
✓	✓	✓	**58.6**	**46.3**	**66.6**	**16.8**

Table 5. Ablation study of the main components of P2F.

Cityscapes, P2F outperforms U3HS in all categories, but with a smaller gap in open-world settings. Especially for L&F, P2F significantly improves in PQ and RQ score. Further experiments, including the recent PANIC benchmark [55], can be found in Appendices B.3 to B.5, and visual results of our mask prediction are shown in Appendix C.4.

Ablation Study: In Tab. 5, we study the influence of no-object mask filtering, the evidential sampling (II), and the symmetric Dice loss (III). While including all masks in Eq. (11a) (i.e., $\mathcal{M}_\emptyset = \emptyset$) has only a minor influence on the FPR, the AP for Road Anomaly is negatively impacted by this ablation. Similarly, the symmetric Dice loss mainly boosts the FPR, and evidential sampling increases the AP. We report more detailed ablations in Appendix B.6.

Limitations: While our P2F brings remarkable improvements to various assumption-free segmentation settings of unknown objects, the gap to methods relying on OOD data and external models remains large (see Tab. 1). Furthermore, the performance degradation between closed- and open-world settings (Tab. 3) highlights the difficulties in calibrating the uncertainty estimates and the known-unknown thresholding solely on known classes. These points can be improved with better uncertainty estimates.

6. Conclusion

We introduced Prior2Former (P2F), the first evidential approach for mask-based vision transformers. P2F quantifies uncertainty by predicting the evidence of a Beta prior to the binary segmentation masks. This informs its embeddings to enable high-quality clustering of anomalous instances. P2F addresses many open-set segmentation tasks (anomaly segmentation, semantic and panoptic open-world segmentation, and anomaly instance segmentation) with the same architecture and training procedure. In contrast to most prior work, P2F does not rely on assumptions about OOD data. This makes it highly applicable to real-world use cases where these assumptions are often unrealistic. Still, compared to existing assumption-free works, P2F delivers far superior performance, as demonstrated in our extensive experiments, effectively narrowing the gap between reliable uncertainty estimates and strong closed-set predictions.

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, 12

[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. 12

[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, 12, 13

[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, 12

[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, 12, 14, 16

[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, 12, 14, 16, 19

[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, 14

[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, 12, 13, 15, 17, 18, 20

[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, 12, 13, 15

[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, 12

[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, 12

[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, 13

[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, 15

[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, 12

[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, 13, 15, 18, 20

[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. 14

[39] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2018. 2, 14

[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, 12, 13, 15

[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, 13, 15

[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, 13, 15

[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, 12, 13, 15

[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, 15, 16

[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. 15

[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, 13

[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. 13, 15, 17, 20

[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