DenseHybrid: Hybrid Anomaly Detection for Dense Open-set Recognition

Matej Grcić, Petra Bevandić, and Siniša Šegvić

University of Zagreb
Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
{matej.grcic,petra.bevandic,sinisa.segvic}@fer.hr

Abstract. Anomaly detection can be conceived either through generative modelling of regular training data or by discriminating with respect to negative training data. These two approaches exhibit different failure modes. Consequently, hybrid algorithms present an attractive research goal. Unfortunately, dense anomaly detection requires translational equivariance and very large input resolutions. These requirements disqualify all previous hybrid approaches to the best of our knowledge. We therefore design a novel hybrid algorithm based on reinterpreting discriminative logits as a logarithm of the unnormalized joint distribution $\hat{p}(\mathbf{x}, \mathbf{y})$. Our model builds on a shared convolutional representation from which we recover three dense predictions: i) the closed-set class posterior $P(\mathbf{y}|\mathbf{x})$, ii) the dataset posterior $P(d_{in}|\mathbf{x})$, iii) unnormalized data likelihood $\hat{p}(\mathbf{x})$. The latter two predictions are trained both on the standard training data and on a generic negative dataset. We blend these two predictions into a hybrid anomaly score which allows dense open-set recognition on large natural images. We carefully design a custom loss for the data likelihood in order to avoid backpropagation through the untractable normalizing constant $Z(\theta)$. Experiments evaluate our contributions on standard dense anomaly detection benchmarks as well as in terms of open-mIoU - a novel metric for dense open-set performance. Our submissions achieve state-of-the-art performance despite neglectable computational overhead over the standard semantic segmentation baseline. Official implementation: `https://github.com/matejgrcic/DenseHybrid`

Keywords: Dense anomaly detection, Dense open-set recognition, Out-of-distribution detection, Semantic segmentation

1 Introduction

High accuracy, fast inference and small memory footprint of modern neural networks steadily expand the horizon of downstream applications. Many exciting applications require advanced image understanding functionality provided by semantic segmentation [17]. These models associate each pixel with a class from a predefined taxonomy. They can accurately segment two megapixel images in

real-time on low-power embedded hardware [11,43,26]. However, the standard training procedures assume the closed-world setup which may raise serious safety issues in real-world deployments. For example, if a segmentation model misclassifies an unknown object (e.g. lost cargo) as road, the autonomous car may experience a serious accident. Such hazards can be alleviated by complementing semantic segmentation with dense anomaly detection. The resulting dense open-set recognition models are more suitable for real-world applications due to ability to decline the decision in anomalous pixels.

Previous approaches for dense anomaly detection either use a generative or a discriminative perspective. Generative approaches are based on density estimation [6] or image resynthesis [36,4]. Discriminative approaches use classification confidence [23], a binary classifier [3] or Bayesian inference [29]. These two perspectives exhibit different failure modes. Generative detectors inaccurately disperse the probability volume [41,47,38,53] or rely on risky image resynthesis. On the other hand, discriminative detectors assume training on full span of the input space, even including unknown unknowns [25].

In this work we combine the two perspectives into a hybrid anomaly detector. The proposed approach complements a standard semantic segmentation model with two additional predictions: i) unnormalized dense data likelihood $\hat{p}(\mathbf{x})$ [6], and ii) dense data posterior $P(d_{in}|\mathbf{x})$ [3]. Both predictions require training with negative data [25,3,4,10]. Joining these two outputs yields an accurate yet efficient dense anomaly detector which we refer to as DenseHybrid.

We summarize our contributions as follows. We propose the first hybrid anomaly detector which allows end-to-end training and operates at pixel level. Our approach combines likelihood evaluation and discrimination with respect to an off-the-shelf negative dataset. Our experiments reveal accurate anomaly detection despite minimal computational overhead. We complement semantic segmentation with DenseHybrid to achieve dense open-set recognition. We report state-of-the-art dense open-set recognition performance according to a novel performance metric which we refer to as *open-mIoU*.

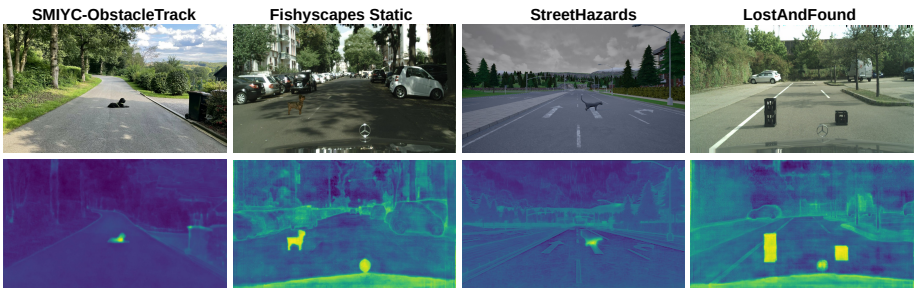

Fig. 1. Qualitative performance of the proposed DenseHybrid approach on standard datasets. Top: input images. Bottom: dense maps of the proposed anomaly score

2 Related Work

Detecting samples which deviate from the generative process of the training data is a decades old problem [22]. In the machine learning community this task is also known as anomaly detection or out-of-distribution (OOD) detection [24]. Early image-wide approaches utilize max-softmax probability [24], input perturbations [34] ensembling [31] or Bayesian uncertainty [40]. More encouraging performance has been reported by discriminative training against a broad negative dataset [14,25,3,37] or an appropriately trained generative model [32,21,54].

Another line of work detects anomalies by estimating the likelihood with a generative model. Surprisingly, this research revealed that anomalies may give rise to higher likelihood than inliers [41,47,53]. Further works suggest that better performance can be hoped for group-wise anomaly detection [27], however, this case has less practical importance. Generative models can be encouraged to assign low likelihood in negative training data [25]. This practice may mitigate sub-optimal dispersion of the probability volume [38].

Image-wide anomaly detection approaches can be adapted for dense prediction with variable success. None of the existing generative approaches can deliver dense likelihood estimates. On the other hand, concepts such as max-softmax and discriminative training with negative data are easily ported to dense prediction. Many dense anomaly detectors are trained on mixed-content images obtained by pasting negatives (e.g. ImageNet, COCO, ADE20k) over regular training images [3,10,4]. Discriminative anomaly detections may be produced by a dedicated OOD head which shares features with the standard classification head. Shared features improve OOD performance and incur neglectable computational overhead with respect to the baseline semantic segmentation model [3]. Recent approach [10] encourages large softmax entropy in negative pixels.

Anomalies can also be recognized in feature space [6]. However, this approach complicates the detection of small objects due to subsampled feature represenations and feature collapse [38,1]. Orthogonally to previous approaches, anomaly detector can be implemented according to dissimilarity between the input and a resynthesised image [36,4,50]. The resynthesis is performed by a generative model conditioned on the predicted labels. However, this approach is suitable only for uniform backgrounds such as roads [36]. Furthermore, it adds significant computational overhead making it inapplicable for real-time applications.

Our approach to dense anomaly detection is a hybrid combination of discriminative detection and likelihood evaluation. Discriminative OOD detection has been introduced in [3,25,14]. Contrary to all these approaches, we improve discriminative OOD detection through synergy with likelihood testing. Dense likelihood evaluation has been accomplished by fitting a generative model to discriminative features [6]. However, their approach is vulnerable to feature collapse[38,1] due to two-phase training. Moreover, detection of small outliers is jeopardized due to subsampling. Contrary to their approach, our method allows joint training with the standard dense prediction model and anomaly detection at full resolution.

We perform dense likelihood evaluation by reinterpreting logits as unnormalized joint likelihood [20]. However, the method [20] is completely unsuitable for dense prediction due to intractability of Langevin sampling at large resolutions. We reformulate their method in order to allow training on mixed-content images and show that such adaptation dramatically simplifies the training by precluding backpropagation through intractable normalizing constant $Z(\theta)$. To the best of our knowledge, the proposed design offers the first approach for dense likelihood evaluation that is suitable for end-to-end training.

We build an open-set recognition model by thresholding our hybrid anomaly score and combining it with the standard semantic segmentation predictions [7]. The resulting model is suitable for simultaneous anomaly detection and recognition of inlier scenery. We note that standard metrics for dense recognition performance [16] do not take into account the accuracy in anomalous samples. This is not surprising since outlier pixels have been introduced only in recent dense prediction benchmarks [52,5,9]. Also, previous work on discrimination in presence of anomalous pixels was more focused on robustness of algorithms rather than on recognition performance [52]. Hence, we propose a novel anomaly-aware metric (open-mIoU) which measures the prediction quality both in inliers and the outliers, similarly to previous image-wide metrics [48,46].

3 Dense Recognition with Hybrid Anomaly Detector

We propose a hybrid algorithm for dense anomaly detection based on unnormalized data likelihood and dataset posterior (Sec. 3.1). The proposed hybrid anomaly detector extends the standard dense classifier to form dense open-set recognition model (Sec. 3.2). The resulting recognition model trains on mixed content images.

3.1 Hybrid Anomaly Detection for Dense Prediction

We represent RGB images with a random variable $\underline{\mathbf{x}}$. Variable $\underline{\mathbf{y}}$ denotes the corresponding pixel-level predictions, while the binary random variable $\underline{d}$ models whether a given pixel belongs to the inliers or outliers. We denote a realization of a random variable without the underline. Thus, $P(\mathbf{y}|\mathbf{x})$ is a shortcut for $P(\underline{\mathbf{y}} = \mathbf{y}|\underline{\mathbf{x}} = \mathbf{x})$. We write d_{in} for inliers and d_{out} for outliers. Thus, $P(d_{in}|\mathbf{x})$ denotes a dense posterior probability that a given pixel is an inlier [25,3]. Conversely, $p(\mathbf{x})$ denotes dense likelihoods of patches centered at a given pixel.

We build upon reinterpretation of logits $\mathbf{s}$ produced by a discriminative model $P(\mathbf{y}|\mathbf{x}) = \mathrm{softmax}(f_{\theta_2}(q_{\theta_1}(\mathbf{x})))$ [20]. We reinterpret the logits as unnormalized joint log-density of input and labels:

$$p(\mathbf{y}, \mathbf{x}) = \frac{1}{Z}\hat{p}(\mathbf{y}, \mathbf{x}) := \frac{1}{Z}\exp \mathbf{s}, \quad \mathbf{s} = f_{\theta_2}(q_{\theta_1}(\mathbf{x})). \tag{1}$$

Note that q_{θ_1} produces pre-logits $\mathbf{t}$ based on which f_{θ_2} computes logits $\mathbf{s}$. Hence, q_{θ_1} and f_{θ_2} form the standard discriminative model. $\hat{p}(\mathbf{y}, \mathbf{x})$ denotes unnormalized joint density across data $\underline{\mathbf{x}}$ and labels $\underline{\mathbf{y}}$, while Z denotes the corresponding

normalization constant. As usual, computing Z is intractable since it requires evaluating the unnormalized distribution for all realizations of $\underline{\mathbf{y}}$ and $\underline{\mathbf{x}}$. Throughout this work we conveniently eschew the evaluation of Z in order to enable efficient training and inference.

Standard discriminative predictions are easily obtained through Bayes rule:

$$P(\mathbf{y}|\mathbf{x}) = \frac{p(\mathbf{y},\mathbf{x})}{\sum_{\mathbf{y}} p(\mathbf{y},\mathbf{x})} = \frac{\exp \mathbf{s}}{\sum_i \exp \mathbf{s}_i} = \text{softmax}(\mathbf{s}). \tag{2}$$

Hence, we can recover the unnormalized joint density (1) through the standard closed-world discriminative learning over K classes. Moreover, we can share the logits with the primary discriminative task and even exploit pretrained classifiers.

We can express the dense likelihood $p(\mathbf{x})$ by marginalizing out $\underline{\mathbf{y}}$:

$$p(\mathbf{x}) = \sum_y p(\mathbf{y},\mathbf{x}) = \frac{1}{Z} \sum_y \hat{p}(\mathbf{y},\mathbf{x}) = \frac{1}{Z} \sum_i \exp \mathbf{s}_i. \tag{3}$$

One could argue for detecting anomalies with $p(\mathbf{x})$ directly: if a given input is unlikely under the $p(\mathbf{x})$, it should likely be an anomaly. However, this approach may not work very well in practice due to tendency of maximum likelihood optimization towards over-generalization [38]. In simple words, some outliers will have higher likelihood than the inliers [47,41]. We discourage such behaviour by minimizing the likelihood of negatives during the training [25].

Besides logit reinterpretation, we define the dataset posterior $P(d_{in}|\mathbf{x})$ as a non-linear transformation based on pre-logit activations $q_{\theta_1}(\mathbf{x})$ [3]:

$$P(d_{in}|\mathbf{x}) := \sigma(g_\gamma(q_{\theta_1}(\mathbf{x}))). \tag{4}$$

In our case, the function g is BN-ReLU-Conv1x1 of pre-logits, followed by a sigmoid non-linearity. Anomalies can be detected solely with $P(d_{in}|\mathbf{x})$ [13]: inlier samples should give rise to high posterior of the inlier dataset. However, our experiments show that this is suboptimal compared to our hybrid approach.

Fig. 2 illustrates shortcomings of generative and discriminative anomaly detectors on a toy problem. Blue dots designate inlier data. Green triangles designate the negative data used for training. Red squares denote anomalous test data. Discriminative detectors which model $P(d_{in}|\mathbf{x})$ can't differentiate inliers if the negative data seen during the training insufficiently covers the sample space (left). On the other hand, generative detectors which model $p(\mathbf{x})$ tend to inaccurately distribute probability volume over sample space [38] (center). Joining discriminative and generative approach into a hybrid detector we mitigate the aforementioned limitations (right).

We build our hybrid anomaly detector upon the discriminative dataset posterior $P(d_{in}|\mathbf{x})$ and the generative data likelihood $p(\mathbf{x})$. We express a novel hybrid anomaly score as log-ratio between $P(d_{out}|\mathbf{x}) = 1 - P(d_{in}|\mathbf{x})$ and $p(\mathbf{x})$:

$$s(\mathbf{x}) := \ln \frac{P(d_{out}|\mathbf{x})}{p(\mathbf{x})} = \ln P(d_{out}|\mathbf{x}) - \ln \hat{p}(\mathbf{x}) + \ln Z \tag{5}$$

$$\cong \ln P(d_{out}|\mathbf{x}) - \ln \hat{p}(\mathbf{x}). \tag{6}$$

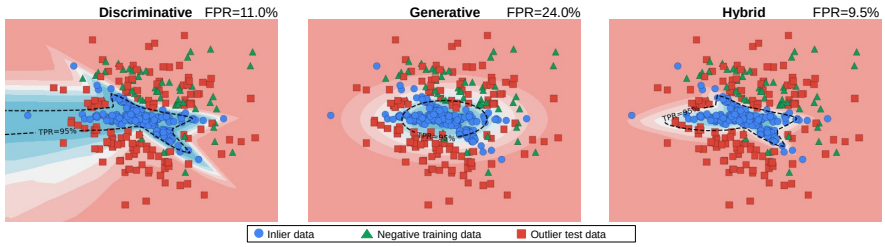

Fig. 2. Anomaly detection on a toy dataset. The discriminative approach (left) models $P(d_{in}|\mathbf{x})$. It fails if the negative training dataset does not cover all modes of the test anomalies. The generative approach (middle) models $p(\mathbf{x})$. It often assigns high likelihoods to test anomalies due to over-generalization [38]. The hybrid approach achieves a synergy between discriminative and generative modelling

We can neglect Z since ranking performance [24] is invariant to monotonic transformations such as taking a logarithm or adding a constant. Other formulations of $s(\mathbf{x})$ may also be effective which is an interesting direction for future work.

3.2 Dense Open-set Recognition based on Hybrid Anomaly Detection

Figure 3 illustrates the inference with the proposed open-set recognition setup. RGB input is fed to a hybrid dense model which produces pre-logit activations $\mathbf{t}$ and logits $\mathbf{s}$. Then, we obtain the closed-set class posterior $P(\mathbf{y}|\mathbf{x}) = \text{softmax}(\mathbf{s})$ (designated in yellow) and the unnormalized data likelihood $\hat{p}(\mathbf{x})$ (designated in green). A distinct head g transforms pre-logits $\mathbf{t}$ into the dataset posterior $P(d_{out}|\mathbf{x})$. The anomaly score $s(\mathbf{x})$ is a log-ratio between latter two distributions. The resulting anomaly map is thresholded and fused with the discriminative output into the final dense open-set recognition map.

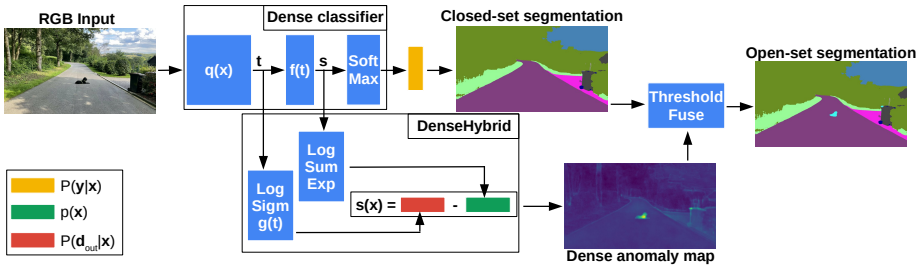

Fig. 3. The proposed dense open-set recognition approach. Our anomaly score is a log-ratio of outputs derived from the hybrid model. We fuse the thresholded anomaly score with the closed-set segmentation map to obtain the open-set segmentation map

The developed hybrid model aims at achieving a synergy between generative and discriminative modelling. However, the proposed hybrid interpretation requires specific training objectives. Dense class posteriors require a discriminative loss over the inlier data D_{in}:

$$L_{\text{cls}}(\theta) = \mathbb{E}_{\mathbf{x},\mathbf{y}\in D_{in}}[-\ln P(\mathbf{y}|\mathbf{x})] \tag{7}$$

$$= -\mathbb{E}_{\mathbf{x},\mathbf{y}\in D_{in}}[\mathbf{s}_y] + \mathbb{E}_{\mathbf{x},\mathbf{y}\in D_{in}}[\ln \sum_i \exp \mathbf{s}_i]. \tag{8}$$

The discriminative loss (7) corresponds to the standard training in the closed world. We introduce the negative data D_{out} into the training procedure to ensure the desired behaviour of $P(d_{in}|\mathbf{x})$ and $p(\mathbf{x})$ [25,3]. Both distributions should yield low probability in negative pixels. We propose to train $p(\mathbf{x})$ to maximize the likelihood in inliers and to minimize the likelihood in outliers. We derive the upper bound of the desired loss as follows:

$$L_{\mathbf{x}}(\theta) = \mathbb{E}_{\mathbf{x}\in D_{in}}[-\ln p(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\in D_{out}}[-\ln p(\mathbf{x})] \tag{9}$$

$$= \mathbb{E}_{\mathbf{x}\in D_{in}}[-\ln \hat{p}(\mathbf{x})] + \cancel{\ln Z} - \mathbb{E}_{\mathbf{x}\in D_{out}}[-\ln \hat{p}(\mathbf{x})] - \cancel{\ln Z} \tag{10}$$

$$= -\mathbb{E}_{\mathbf{x}\in D_{in}}\left[\ln \sum_i \exp(\mathbf{s}_i)\right] + \mathbb{E}_{\mathbf{x}\in D_{out}}\left[\ln \sum_i \exp(\mathbf{s}_i)\right] \tag{11}$$

$$\leq -\mathbb{E}_{\mathbf{x},\mathbf{y}\in D_{in}}[\mathbf{s}_y] + \mathbb{E}_{\mathbf{x}\in D_{out}}[\ln \sum_i \exp(\mathbf{s}_i)]. \tag{12}$$

Note that we eschew the backpropagation into the normalization constant Z, and derive the upper bound according to the following inequality:

$$\ln \sum_i \exp \mathbf{s}_i \geq \max_i \mathbf{s}_i \geq \mathbf{s}_y. \tag{13}$$

Proof of inequality (13) can be easily derived by recalling that log-sum-exp is a smooth upper bound of the max function. By comparing the standard classification loss (7) and the upper bound (12) we realize that minimizing the standard classification loss increases $p(\mathbf{x})$ for inlier pixels. Indeed, minimizing the negative logarithm of softmax output increases the value of logit for the correct class.

Alternatively, $p(\mathbf{x})$ could be trained only on inliers [45,15,20]. This would require sample hallucination via MCMC sampling and back-propagation into the corresponding approximation of Z. Such procedure is infeasible for large images. Consequently, we choose to deal with negative samples instead of hallucinated ones and optimize the proposed loss $L_{\mathbf{x}}(\theta)$.

We train the dataset posterior $P(d_{in}|\mathbf{x})$ with the standard discriminative loss [3]:

$$L_{\mathbf{d}}(\theta, \gamma) = \mathbb{E}_{\mathbf{x}\in D_{in}}[-\ln P(d_{in}|\mathbf{x})] + \mathbb{E}_{\mathbf{x}\in D_{out}}[-\ln(P(d_{out}|\mathbf{x}))]. \tag{14}$$

By joining losses L_{cls}, $L_{\mathbf{x}}$ and $L_{\mathbf{d}}$ we obtain the final loss:

$$L(\theta, \gamma) = -\mathbb{E}_{\mathbf{x},\mathbf{y}\in D_{in}}[\ln P(\mathbf{y}|\mathbf{x}) + \ln P(d_{in}|\mathbf{x})] \\ - \beta \cdot \mathbb{E}_{\mathbf{x}\in D_{out}}[\ln(P(d_{out}|\mathbf{x})) - \ln \hat{p}(\mathbf{x})]. \tag{15}$$

Hyperparameter β controls the impact of negative data to the primary classification task. Note that the final loss (15) omits the first term from $L_{\mathbf{x}}$ (12) in positive pixels. We choose to do so since $\hat{p}(\mathbf{x})$ is implicitly optimized through L_{cls}.

Figure 4 illustrates the described training procedure of the proposed open-set recognition model. We prepare the training images by pasting the negative instances atop the standard training images. The resulting mixed-content image [3] is fed to the hybrid model. We obtain the classification output $P(\mathbf{y}|\mathbf{x})$ with softmax. The unnormalized likelihood $\hat{p}(\mathbf{x})$ is obtained through sum-exp operator. We recover $p(d_{in}|\mathbf{x})$ by branching from pre-logit activations. The model outputs are trained by applying the dicriminative loss L_{cls} (7), likelihood loss $L_{\mathbf{x}}$ (12) and dataset posterior loss $L_{\mathbf{d}}$ (14). As proposed, these losses are conveniently joined into a single loss $L(\theta, \gamma)$ (15).

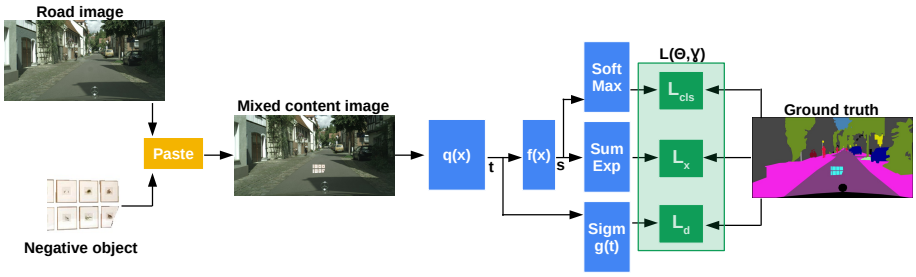

Fig. 4. The training procedure of the proposed open-set recognition model. Mixed-content images are fed to the open-set model with three outputs. Each output is optimized according to the compound loss (15)

4 Measuring Dense Open-set Performance

Test datasets for anomaly segmentation either exclusively measure the performance of anomaly detectors [44,9] or simply report the classification performance [5]. In the latter case, the reported drop in segmentation performance is usually negligible and is explained away by allocation of model capacity for the anomaly detection. We will show that the real impact of anomaly detector on the segmentation performance can be clearly seen only in the open world. Also, the impact is more severe than the small performance drop visible in the closed world.

To properly measure open-set recognition performance, we first select threshold at which the anomaly detector achieves TPR of 95%. This ensures high safety standards for the recognition model. Then, we override the classification in pixels which raise concern according to the thresholded anomaly map. The resulting recognition map has $K + 1$ labels. We compute the recognition performance in

open-world using open intersection over union (open-IoU). For the k-th class we can compute the proposed open-IoU as:

$$\text{open-IoU}_k = \frac{\text{TP}_k}{\text{TP}_k + \text{FP}_k^{\text{ow}} + \text{FN}_k^{\text{ow}}}, \; \text{FP}_k^{\text{ow}} = \sum_{\substack{i \neq k \\ i=1}}^{K+1} \text{FP}_k^i, \; \text{FN}_k^{\text{ow}} = \sum_{\substack{i \neq k \\ i=1}}^{K+1} \text{FN}_k^i \quad (16)$$

Different that the standard IoU formulation, open-IoU also takes into account false positives and false negatives caused by imperfect anomaly detector. However, we still average open-IoU over K inlier classes. This means that a recognition model which uses a perfect anomaly detector would match segmentation performance in the closed world. This property would not be preserved if we averaged IoU over K+1 classes.

Figure 5 (right) shows the open world confusion matrix. Imperfect anomaly detection impacts recognition performance through increased false positives (designated in yellow) and false negatives (designated in red). Difference between closed mIoU and averaged open-IoU over K inlier classes reveals the performance hit due to inaccurate anomaly detection.

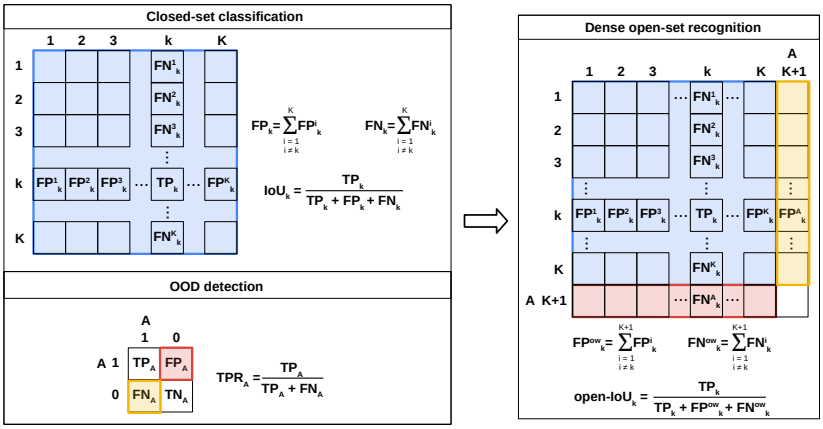

Fig. 5. The proposed open intersection over union (open-IoU) takes into account misclassifications in anomalous pixels to accurately measure dense recognition performance in open world

Measuring performance using the proposed open-IoU requires datasets with K+1 labels. Creating such taxonomy requires substantial resources. Currently, only StreetHazards [23] offers appropriate taxonomy for measuring open-IoU.

5 Experiments

We report dense anomaly detection and open-set recognition performance of the proposed DenseHybrid approach, and compare them with the state of the art.

We also explore influence of distance, show computational requirements of the proposed module, and ablate the design choices.

5.1 Benchmarks and Datasets

We evaluate performance on standard benchmarks for dense anomaly detection. Fishyscapes [5] considers urban scenarios on a subset of LostAndFound [44] and on Cityscapes validation images with pasted anomalies (FS Static). SegmentMeIfYouCan (SMIYC) [9] moves away from anomaly injection. Instead, appropriate images are collected from the real world and grouped based on the anomaly size into AnomalyTrack (large) and ObstacleTrack (small). Additionally, the benchmark encapsulates all LostAndFound images. Unfortunately, both benchmarks only have binary labels which makes them insufficient for measuring the recognition performance as proposed in Sec. 4. StreetHazards [23] is a synthetic dataset created by CARLA virtual environment. The simulated environment enables smooth anomaly injection and low-cost label extraction. Consequently, the dataset contains $K+1$ labels which makes it suitable for measuring both anomaly detection and dense recognition.

5.2 Dense Anomaly Detection

Table 1 shows performance of the proposed hybrid anomaly detector on the SMIYC benchmark [9]. DenseHybrid outperforms contemporary approaches on both AnomalyTrack and ObstacleTrack by a wide margin. Also, the proposed anomaly detector achieves the best FPR on LostAndFound.

Table 1. Performance evaluation on the SMIYC benchmark [9]. DenseHybrid outperforms contemporary approaches on Anomaly and Obstacle track by a wide margin, while also achieving the best FPR on LostAndFound

Method	Aux data	Img rsyn.	AnomalyTrack AP	AnomalyTrack FPR_{95}	ObstacleTrack AP	ObstacleTrack FPR_{95}	LAF-noKnown AP	LAF-noKnown FPR_{95}
SynBoost [4]	✓	✓	56.4	61.9	71.3	3.2	81.7	4.6
Image Resyn. [36]	✗	✓	52.3	25.9	37.7	4.7	57.1	8.8
JSRNet [50]	✗	✓	33.6	43.9	28.1	28.9	74.2	6.6
Road Inpaint. [35]	✗	✓	-	-	54.1	47.1	**82.9**	35.8
Embed. Dens. [5]	✗	✗	37.5	70.8	0.8	46.4	61.7	10.4
ODIN [34]	✗	✗	33.1	71.7	22.1	15.3	52.9	30.0
MC Dropout [29]	✗	✗	28.9	69.5	4.9	50.3	36.8	35.6
Max softmax [24]	✗	✗	28.0	72.1	15.7	16.6	30.1	33.2
Mahalanobis [33]	✗	✗	20.0	87.0	20.9	13.1	55.0	12.9
Void Classifier [5]	✓	✗	36.6	63.5	10.4	41.5	4.8	47.0
DenseHybrid (ours)	✓	✗	**78.0**	**9.8**	**87.1**	**0.2**	78.7	**2.1**

Table 2 shows performance of the proposed DenseHybrid on Fishyscapes [5]. Our anomaly detector achieves the best results on FS LostAndFound, and the best FPR on FS Static. We achieve these results while having negligible impact on classification task in closed-world. However, in the next section we show that the impact of anomaly detection to recognition performance is much more significant than in the closed world.

Table 2. Performance evaluation on the Fishyscapes benchmark [5]. DenseHybrid achieves the best performance on FS LostAndFound and the best FPR on FS Static

Method	Aux data	Img rsyn.	LostAndFound AP	LostAndFound FPR_{95}	Static AP	Static FPR_{95}	Closed world Cityscapes mIoU
SynBoost [4]	✓	✓	43.2	15.8	72.6	18.8	**81.4**
Image Resyn. [36]	✗	✓	5.7	48.1	29.6	27.1	81.4
Standardized ML [28]	✗	✗	31.1	21.5	53.1	19.6	80.3
Embed. Dens. [5]	✗	✗	4.7	24.4	62.1	17.4	80.3
Max softmax [24]	✗	✗	1.77	44.9	12.9	39.8	80.3
Dirichlet prior [39]	✓	✗	34.3	47.4	**84.6**	30.0	70.5
OOD Head [3]	✓	✗	30.9	22.2	84.0	10.3	77.3
Void Classifier [5]	✓	✗	10.3	22.1	45.0	19.4	70.4
Mutual information [40]	✓	✗	9.8	38.5	48.7	15.5	73.8
DenseHybrid (ours)	✓	✗	**43.9**	**6.2**	72.3	**5.5**	**81.0**

Table 3 explores sensitivity of anomaly detection with respect to distance from the camera. We perform all these experiments on LostAndFound since it includes disparity maps. Still, due to errors in available disparities, we limit our analysis to the first 50 meters from the camera. The proposed DenseHybrid approach achieves accurate results even at large distances from the vehicle.

Table 3. Anomaly detection performance at different distances from camera. Our DenseHybrid based on DeeplabV3+ with WRN38 backbone [55] accurately detects anomalies at different ranges

Method	Metric	Range in meters 5-10	10-15	15-20	20-25	25-30	30-35	35-40	40-45	45-50
Max-softmax [24]	AP	28.7	28.8	26.0	25.1	29.0	26.2	29.6	31.7	33.7
	FPR_{95}	16.4	29.7	28.8	44.2	41.3	47.8	44.7	43.2	45.3
Max-logit [23]	AP	76.1	73.9	78.2	69.6	72.6	70.2	71.0	74.0	73.9
	FPR_{95}	5.4	16.2	5.9	12.8	9.5	10.0	9.8	9.8	11.0
SynBoost [4]	AP	**93.7**	78.7	76.9	70.0	65.6	58.5	59.8	60.0	53.3
	FPR_{95}	**0.2**	17.7	25.0	23.3	18.8	27.4	25.4	25.8	29.9
DenseHybrid (ours)	AP	90.7	**89.8**	**92.9**	**89.1**	**89.5**	**87.7**	**85.0**	**85.6**	**82.1**
	FPR_{95}	0.3	**1.1**	**0.6**	**1.4**	**1.4**	**2.5**	**3.7**	**4.7**	**6.3**

5.3 Dense Open-set Recognition

By fusing a properly thresholded anomaly detector with the dense classifier, we obtain a dense open-set recognition model (Fig. 3). The resulting model detects anomalous scene parts, while correctly classifying the rest of the scene.

To measure the dense recognition performance, we create two test folds based on towns t5 and t6 from StreetHazards test. Then, we select anomaly threshold on t6 and use it to measure the proposed open-mIoU on t5. We switch the folds and repeat the procedure. We compute the weighted average based on image count to obtain the final test set open-mIoU.

Table 4 shows performance of our dense recognition models on StreetHazards. The left part of the table considers anomaly detection where DenseHybrid achieves the best performance. The right part of the table considers dense recognition performance. Our model outperforms other contemporary approaches despite lower classification performance in the closed world. Note that the performance drop between the closed and the open set is significant. The models achieve over 60% mIoU in closed world while the open world performance peeks at 46%. Hence, we conclude that even the best anomaly detectors are still insufficient for matching the closed world performance in open-world. Researchers should strive to close this gap in order to improve the safety of recognition systems in the real world.

Table 4. Performance evaluation on StreetHazards [23]. DenseHybrid achieves the best anomaly detection performance. The corresponding open-set recognition model yields the best performance measured by open-mIoU (Sec. 4)

Method	Aux. data	Anomaly detection			Closed world	Open world		
		AP	FPR_{95}	AUC	$\overline{IoU}$	o-$\overline{IoU}$-t5	o-$\overline{IoU}$-t6	o-$\overline{IoU}$
SynthCP [51]	✗	9.3	28.4	88.5	-	-	-	-
Dropout [29][51]	✗	7.5	79.4	69.9	-	-	-	-
TRADI [19]	✗	7.2	25.3	89.2	-	-	-	-
OVNNI [18]	✗	12.6	22.2	91.2	54.6	-	-	-
SO+H [21]	✗	12.7	25.2	91.7	59.7	-	-	-
DML [8]	✗	14.7	17.3	93.7	-	-	-	-
MSP [24]	✗	7.5	27.9	90.1	65.0	32.7	40.2	35.1
ML [23]	✗	11.6	22.5	92.4	65.0	39.6	44.5	41.2
ODIN [34]	✗	7.0	28.7	90.0	65.0	26.4	33.9	28.8
ReAct [49]	✗	10.9	21.2	92.3	62.7	33.0	36.2	34.0
Energy [37]	✓	12.9	18.2	93.0	63.3	41.7	44.9	42.7
Outlier Exposure [25]	✓	14.6	17.7	94.0	61.7	43.7	44.1	43.8
OOD-Head [2]	✓	19.7	56.2	88.8	**66.6**	33.7	34.3	33.9
OH*MSP [3]	✓	18.8	30.9	89.7	**66.6**	43.3	44.2	43.6
DenseHybrid (ours)	✓	**30.2**	**13.0**	**95.6**	63.0	**46.1**	**45.3**	**45.8**

Figure 6 visualises dense anomaly and recognition maps on StreetHazards. Our recognition model significantly outperforms the max-logit baseline [23].

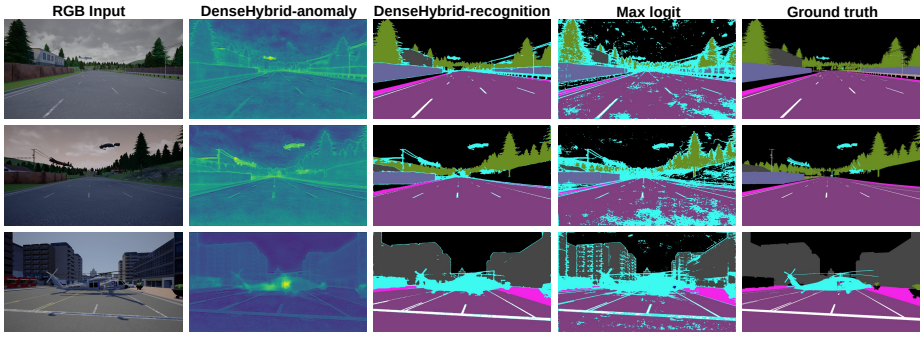

Fig. 6. Visualisation of dense open-set recognition performance on StreetHazards. DenseHybrid significantly outperforms the max-logit baseline [23]

5.4 Inference speed

Table 5 shows computational overhead of the proposed DenseHybrid anomaly detector over the baseline segmentation model on two megapixels images. DenseHybrid has negligible computational overhead of 0.1 GFLOPs and 2.8ms. Our results are averaged over 200 runs on NVIDIA RTX3090. These experiments also suggest that image resynthesis is not applicable for real-time inference.

Table 5. Computational overhead of the proposed DenseHybrid anomaly detector when inferring with RTX3090 on two megapixel images

Method	Resynth.	Infer. time (ms)	Frames per sec.	GFLOPs
SynBoost [4]	✓	1055.5	<1	-
SynthCP [51]	✓	146.9	<1	4551.1
LDN-121 [30]	✗	60.9	16.4	202.3
LDN-121 + SML [28]	✗	75.4	13.3	202.6
LDN-121 + DenseHybrid (ours)	✗	**63.7**	**15.7**	**202.4**

5.5 Impact of anomaly detector design

Table 6 compares the proposed DenseHybrid approach with its generative and discriminative components – $\hat{p}(\mathbf{x})$ and $P(d_{in}|\mathbf{x})$. The hybrid anomaly score based on the ratio of these two distributions outperforms each of the two components. The results are averaged over the last three epochs.

Table 6. Validation of DenseHybrid components on Fishyscapes validation set

Anomaly detector	FS LostAndFound		FS Static	
	AP	FPR_{95}	AP	FPR_{95}
Discriminative $(1 - P(d_{in}\|\mathbf{x}))$	42.9 ± 4.2	42.1 ± 7.0	47.8 ± 5.0	41.6 ± 8.3
Generative $\hat{p}(\mathbf{x})$	60.5 ± 2.6	7.4 ± 0.8	54.2 ± 2.1	6.2 ± 0.7
Hybrid $(1 - P(d_{in}\|\mathbf{x}))/\hat{p}(\mathbf{x})$	**63.8** ± 2.9	**6.1** ± 0.7	**60.0** ± 2.0	**4.9** ± 0.6

5.6 Implementation details

We adapt the standard segmentation networks [30,55] to enable co-operation with our hybrid anomaly detector. We append an additional branch g_γ which is in our case BN-ReLU-Conv1x1. The additional branch computes the discriminative anomaly output. We obtain generative anomaly output by computing sum of exponentiated logits. We build our recognition models based on dense classifiers. We fine-tune all our models on mixed content images with pasted negative instances from ADE20k. In the case of SMIYC we fine-tune LDN-121 [30] for 10 epochs on images from Cityscapes [12], Vistas [42] and Wilddash2 [52]. In the case of Fishyscapes we use DeepLabV3+ with WideResNet38 [55]. We fine-tune the model for 10 epochs on Cityscapes. We train LDN-121 on StreetHazards for 120 epochs in closed world and then fine-tune the recognition model on mixed-content images. Other details are available in the supplement.

6 Conclusion

Discriminative and generative approaches to dense anomaly detection assume different failure modes. We propose to achieve a synergy of these two approaches by fusing the data posterior and the data likelihood derived from the standard discriminative model. The proposed hybrid setup relies on unnormalized distributions. Hence, we try to eschew evaluation of the intractable normalization constant both during training and inference. The proposed DenseHybrid architecture yields state-of-the-art performance on the standard anomaly segmentation benchmarks as well as competitive dense recognition performance in the open world. The latter is measured with the novel open-mIoU score which takes into account classification in both inliers and anomalous pixels. Future work should focus on reducing the revealed performance gap between closed-world and open-world recognition in order to improve the progress toward safe autonomous driving systems.

Acknowledgements

This work has been supported by Croatian Science Foundation grant IP-2020-02-5851 ADEPT, as well as by European Regional Development Fund grants KK.01.1.1.01.0009 DATACROSS and KK.01.2.1.02.0119 A-Unit. We thank Marin Oršić for insightful discussions during early stages of this work.

References

1. van Amersfoort, J., Smith, L., Jesson, A., Key, O., Gal, Y.: On feature collapse and deep kernel learning for single forward pass uncertainty. arXiv preprint arXiv:2102.11409 (2021)
2. Bevandic, P., Kreso, I., Orsic, M., Segvic, S.: Simultaneous semantic segmentation and outlier detection in presence of domain shift. In: 41st DAGM German Conference, DAGM GCPR (2019). https://doi.org/10.1007/978-3-030-33676-9_3
3. Bevandić, P., Krešo, I., Oršić, M., Šegvić, S.: Dense open-set recognition based on training with noisy negative images. Image and Vision Computing **124**, 104490 (2022)
4. Biase, G.D., Blum, H., Siegwart, R., Cadena, C.: Pixel-wise anomaly detection in complex driving scenes. In: Computer Vision and Pattern Recognition, CVPR (2021)
5. Blum, H., Sarlin, P.E., Nieto, J., Siegwart, R., Cadena, C.: The fishyscapes benchmark: Measuring blind spots in semantic segmentation. International Journal of Computer Vision **129**(11), 3119–3135 (2021)
6. Blum, H., Sarlin, P., Nieto, J.I., Siegwart, R., Cadena, C.: Fishyscapes: A benchmark for safe semantic segmentation in autonomous driving. In: 2019 IEEE/CVF International Conference on Computer Vision Workshops. pp. 2403–2412. IEEE (2019). https://doi.org/10.1109/ICCVW.2019.00294
7. Boult, T.E., Cruz, S., Dhamija, A.R., Günther, M., Henrydoss, J., Scheirer, W.J.: Learning and the unknown: Surveying steps toward open world recognition. In: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019. pp. 9801–9807. AAAI Press (2019)
8. Cen, J., Yun, P., Cai, J., Wang, M.Y., Liu, M.: Deep metric learning for open world semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15333–15342 (October 2021)
9. Chan, R., Lis, K., Uhlemeyer, S., Blum, H., Honari, S., Siegwart, R., Salzmann, M., Fua, P., Rottmann, M.: Segmentmeifyoucan: A benchmark for anomaly segmentation. CoRR **abs/2104.14812** (2021)
10. Chan, R., Rottmann, M., Gottschalk, H.: Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In: International Conference on Computer Vision, ICCV (2021)
11. Chao, P., Kao, C., Ruan, Y., Huang, C., Lin, Y.: Hardnet: A low memory traffic network. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV. pp. 3551–3560. IEEE (2019). https://doi.org/10.1109/ICCV.2019.00365
12. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2016)
13. DeVries, T., Taylor, G.W.: Learning confidence for out-of-distribution detection in neural networks. CoRR **abs/1802.04865** (2018)
14. Dhamija, A.R., Günther, M., Boult, T.E.: Reducing network agnostophobia. In: Annual Conference on Neural Information Processing Systems 2018, NeurIPS (2018)

15. Du, Y., Mordatch, I.: Implicit generation and modeling with energy based models. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada. pp. 3603–3613 (2019), https://proceedings.neurips.cc/paper/2019/hash/378a063b8fdb1db941e34f4bde584c7d-Abstract.html
16. Everingham, M., Eslami, S.M.A., Gool, L.V., Williams, C.K.I., Winn, J.M., Zisserman, A.: The pascal visual object classes challenge: A retrospective. Int. J. Comput. Vis. **111**(1), 98–136 (2015)
17. Farabet, C., Couprie, C., Najman, L., LeCun, Y.: Learning hierarchical features for scene labeling. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1915–1929 (2013)
18. Franchi, G., Bursuc, A., Aldea, E., Dubuisson, S., Bloch, I.: One versus all for deep neural network incertitude (OVNNI) quantification. CoRR **abs/2006.00954** (2020)
19. Franchi, G., Bursuc, A., Aldea, E., Dubuisson, S., Bloch, I.: TRADI: tracking deep neural network weight distributions. In: 16th European Conference on Computer Vision, ECCV (2020)
20. Grathwohl, W., Wang, K., Jacobsen, J., Duvenaud, D., Norouzi, M., Swersky, K.: Your classifier is secretly an energy based model and you should treat it like one. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020 (2020)
21. Grcić, M., Bevandić, P., Šegvić, S.: Dense open-set recognition with synthetic outliers generated by real NVP. In: 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP (2021)
22. Hawkins, D.M.: Identification of Outliers. Monographs on Applied Probability and Statistics, Springer (1980). https://doi.org/10.1007/978-94-015-3994-4, https://doi.org/10.1007/978-94-015-3994-4
23. Hendrycks, D., Basart, S., Mazeika, M., Mostajabi, M., Steinhardt, J., Song, D.: Scaling out-of-distribution detection for real-world settings. arXiv preprint arXiv:1911.11132 (2019)
24. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: 5th International Conference on Learning Representations, ICLR (2017)
25. Hendrycks, D., Mazeika, M., Dietterich, T.G.: Deep anomaly detection with outlier exposure. In: 7th International Conference on Learning Representations, ICLR (2019)
26. Hong, Y., Pan, H., Sun, W., Jia, Y.: Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes. CoRR **abs/2101.06085** (2021)
27. Jiang, D., Sun, S., Yu, Y.: Revisiting flow generative models for out-of-distribution detection. In: International Conference on Learning Representations (2022)
28. Jung, S., Lee, J., Gwak, D., Choi, S., Choo, J.: Standardized max logits: A simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation. In: International Conference on Computer Vision, ICCV (2021)
29. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: Neural Information Processing Systems (2017)
30. Kreso, I., Krapac, J., Segvic, S.: Efficient ladder-style densenets for semantic segmentation of large images. IEEE Trans. Intell. Transp. Syst. **22** (2021)

31. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems. pp. 6402–6413 (2017)
32. Lee, K., Lee, H., Lee, K., Shin, J.: Training confidence-calibrated classifiers for detecting out-of-distribution samples. In: 6th International Conference on Learning Representations, ICLR (2018)
33. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: Neural Information Processing Systems, NeurIPS (2018)
34. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. In: 6th International Conference on Learning Representations, ICLR (2018)
35. Lis, K., Honari, S., Fua, P., Salzmann, M.: Detecting road obstacles by erasing them. CoRR **abs/2012.13633** (2020)
36. Lis, K., Nakka, K.K., Fua, P., Salzmann, M.: Detecting the unexpected via image resynthesis. In: International Conference on Computer Vision, ICCV (2019)
37. Liu, W., Wang, X., Owens, J.D., Li, Y.: Energy-based out-of-distribution detection. In: NeurIPS (2020)
38. Lucas, T., Shmelkov, K., Alahari, K., Schmid, C., Verbeek, J.: Adaptive density estimation for generative models. In: Neural Information Processing Systems (2019)
39. Malinin, A., Gales, M.J.F.: Predictive uncertainty estimation via prior networks. In: Annual Conference on Neural Information Processing Systems (2018)
40. Mukhoti, J., Gal, Y.: Evaluating bayesian deep learning methods for semantic segmentation. CoRR **abs/1811.12709** (2018)
41. Nalisnick, E.T., Matsukawa, A., Teh, Y.W., Görür, D., Lakshminarayanan, B.: Do deep generative models know what they don't know? In: 7th International Conference on Learning Representations, ICLR (2019)
42. Neuhold, G., Ollmann, T., Bulò, S.R., Kontschieder, P.: The mapillary vistas dataset for semantic understanding of street scenes. In: IEEE International Conference on Computer Vision, ICCV (2017)
43. Orsic, M., Segvic, S.: Efficient semantic segmentation with pyramidal fusion. Pattern Recognit. **110**, 107611 (2021). https://doi.org/10.1016/j.patcog.2020.107611
44. Pinggera, P., Ramos, S., Gehrig, S., Franke, U., Rother, C., Mester, R.: Lost and found: detecting small road hazards for self-driving vehicles. In: International Conference on Intelligent Robots and Systems, IROS (2016)
45. Salakhutdinov, R., Hinton, G.: Deep boltzmann machines. In: Twelth International Conference on Artificial Intelligence and Statistics. PMLR (2009)
46. Scherreik, M.D., Rigling, B.D.: Open set recognition for automatic target classification with rejection. IEEE Trans. Aerosp. Electron. Syst. **52**(2), 632–642 (2016)
47. Serrà, J., Álvarez, D., Gómez, V., Slizovskaia, O., Núñez, J.F., Luque, J.: Input complexity and out-of-distribution detection with likelihood-based generative models. In: 8th International Conference on Learning Representations, ICLR (2020)
48. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Inf. Process. Manag. **45**(4), 427–437 (2009)
49. Sun, Y., Guo, C., Li, Y.: React: Out-of-distribution detection with rectified activations. In: NeurIPS (2021)
50. Vojir, T., Šipka, T., Aljundi, R., Chumerin, N., Reino, D.O., Matas, J.: Road anomaly detection by partial image reconstruction with segmentation coupling. In: International Conference on Computer Vision, ICCV (2021)

51. Xia, Y., Zhang, Y., Liu, F., Shen, W., Yuille, A.L.: Synthesize then compare: Detecting failures and anomalies for semantic segmentation. In: 16th European Conference on Computer Vision, ECCV (2020)
52. Zendel, O., Honauer, K., Murschitz, M., Steininger, D., Dominguez, G.F.: Wilddash - creating hazard-aware benchmarks. In: European Conference on Computer Vision (ECCV) (2018)
53. Zhang, L.H., Goldstein, M., Ranganath, R.: Understanding failures in out-of-distribution detection with deep generative models. In: 38th International Conference on Machine Learning, ICML (2021)
54. Zhao, Z., Cao, L., Lin, K.: Revealing distributional vulnerability of explicit discriminators by implicit generators. CoRR **abs/2108.09976** (2021)
55. Zhu, Y., Sapra, K., Reda, F.A., Shih, K.J., Newsam, S.D., Tao, A., Catanzaro, B.: Improving semantic segmentation via video propagation and label relaxation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 8856–8865. Computer Vision Foundation / IEEE (2019)