Self-Supervised Likelihood Estimation with Energy Guidance for Anomaly Segmentation in Urban Scenes

Yuanpeng Tu[1]*, Yuxi Li[2]*, Boshen Zhang[2]*, Liang Liu[2],
Jiangning Zhang[2], Yabiao Wang[2] †, Cai Rong Zhao[1]†
[1] Dept. of Electronic and Information Engineering, Tongji Univeristy, Shanghai
[2] YouTu Lab, Tencent, Shanghai
{2030809, zhaocairong}@tongji.edu.cn
{yukiyxli, boshenzhang, leoneliu, vtzhang, caseywang}@tencent.com

Abstract

Robust autonomous driving requires agents to accurately identify unexpected areas (anomalies) in urban scenes. To this end, some critical issues remain open: how to design advisable metric to measure anomalies, and how to properly generate training samples of anomaly data? Classical effort in anomaly detection usually resorts to pixel-wise uncertainty or sample synthesis, which ignores the contextual information and sometimes requires auxiliary data with fine-grained annotations. On the contrary, in this paper, we exploit the strong context-dependent nature of the segmentation task and design an energy-guided self-supervised framework for anomaly segmentation, which optimizes an anomaly head by maximizing the likelihood of self-generated anomaly pixels. For this purpose, we design two estimators to model anomaly likelihood, one is a task-agnostic binary estimator and the other depicts the likelihood as residual of task-oriented joint energy. Based on the proposed estimators, we devise an adaptive self-supervised training framework, which exploits the contextual reliance and estimated likelihood to refine mask annotations in anomaly areas. We conduct extensive experiments on challenging Fishyscapes and Road Anomaly benchmarks, demonstrating that without any auxiliary data or synthetic models, our method can still achieve comparable performance to supervised competitors. Code is available at https://github.com/yuanpengtu/SLEEG.

Introduction

Recent studies in semantic segmentation have achieved significant advances on close-set benchmarks of urban scenarios (Cordts et al. 2016). However, when it comes to deployment in the wild, it is necessary to enable segmentation models with the ability of anomaly detection.

Essentially, the key of segmentation with anomalies lies in two aspects: *Firstly,* the anomaly score should be designed to differentiate anomaly and normal pixels. *Second,* extra anomaly data is critical to identify which pixel belongs to anomaly areas. To address these issues, a fresh wave of approaches are proposed. To measure the likelihood of anomalies, some methods take insight from uncertainty estimation and devise a series of proxy tasks (Hendrycks and Gimpel 2017; Mukhoti and Gal 2018; Malinin and Gales 2018; Grcić, Bevandić, and Šegvić 2022; Tian et al. 2021). However, they usually design coupled objectives and require retraining of models, which might degrade their performance (Bogdoll, Nitsche, and Zöllner 2022). On the other hand, to generate training samples with anomaly pixels, outlier exposure is widely adopted (Bevandić et al. 2019; Grcić, Bevandić, and Šegvić 2022; Chan, Rottmann, and Gottschalk 2021) by training with auxiliary data, while these methods increase the cost due to additional labeling requirements, the adopted auxiliary data is also not guaranteed to be consistent with realistic scenes. There are also approaches using an extra reconstruction model and taking the reconstruction error as anomaly parts (Xia et al. 2020; Di Biase et al. 2021a; Vojir et al. 2021), which affects the efficiency and their accuracy highly relies on reconstruction quality.

In summary, most previous efforts are designed to capture anomaly samples in classification. Nevertheless, semantic segmentation differs since the its results inherently rely on spatial context. As shown in Fig. 1, given a pretrained model (Chen et al. 2018), the same patch yield different semantic uncertainty (measured as entropy of categorical distribution) when it is placed under a different context, even though the patch is filled with normal pixels. The empirical observation inspires that **we can automatically synthesize anomalies from normal pixels via a self-supervised copy-and-paste manner guided by spatial context.** Such self-supervised paradigm can (1) avoid the cost of explicitly annotating anomalies and (2) ensure the quality of generated anomalies by referring to their context. Therefore, we propose a new framework termed as **S**elf-supervised **L**ikelihood **E**stimation with **E**nergy **G**uidance (**SLEEG**), which extends off-the-shelf segmentation models to anomaly detectors with the guidance of energy model (LeCun et al. 2006) while avoiding the overhead of labeling anomaly data. The SLEEG framework is designed in a self-teaching paradigm, to properly depict the anomaly area, we propose two anomaly estimators based on the joint distribution of content and anomalies. The first is formulated as a simple task-agnostic classifier to differentiate anomaly and normal pixels. The other is a task-oriented estimator and can be regarded as residual

[*] These authors contributed equally.
[†] Corresponding author.

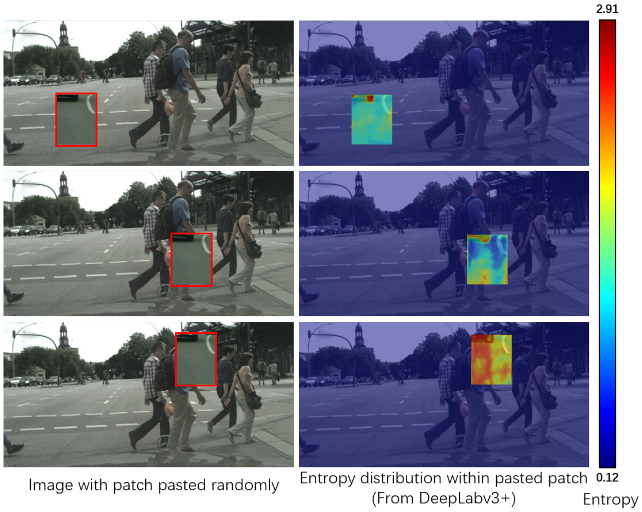
2.91
0.12
Image with patch pasted randomly
Entropy distribution within pasted patch
(From DeepLabv3+)
Entropy

Figure 1: Illustration of contextual reliance in anomaly segmentation tasks. The left column shows an image pasted with a random patch at different position. The right column illustrates the corresponding entropy distribution from segmentation results of DeepLab segmentation model (Chen et al. 2018). Different pasted positions result in various uncertainty distribution within the patch.

estimation of classic joint-energy model (JEM) (Grathwohl et al. 2020), with proper design of loss function, this estimator can be optimized through a dynamic energy-guided margin. Next, based on these estimators, we design an adaptive refinement mechanism to provide dynamic anomaly samples under different contexts, which is guided by anomaly likelihood and contextual information of each pixel.

We implemented SLEEG with different models and evaluate on benchmarks of Fishyscapes (Blum et al. 2019) and Road Anomaly (Lis et al. 2019a). Experimental results show that SLEEG can bring consistent improvement over different baselines by training with only normal data from Cityscapes (Cordts et al. 2016). Further, compared with other state-of-the-art methods, SLEEG achieves competitive results **without training on labeled anomaly data or updating parameters of segmentation networks.** In summary, the contributions can be listed as:

- We propose SLEEG, a self-supervised framework for anomaly segmentation in a copy-and-paste manner. In the framework, we design two decoupled likelihood estimators from an energy model, a task-agnostic estimator for discriminative learning and a task-oriented estimator for residual learning of joint-energy.
- Based on the proposed energy-guided estimators, we propose a dynamic mask refinement mechanism by applying likelihood-guided pixel separation to help extract more informative anomaly areas for training.
- Without training on labeled auxiliary data or updating segmentation parameters, our SLEEG achieves competitive performance on both Fishyscapes (Blum et al. 2019) and Road Anomaly (Lis et al. 2019a) benchmarks.

Related Work

Anomaly Segmentation

Some of previous effort extends technique from anomaly detection (e.g. uncertainty estimation and outlier exposure) in classification into segmentation tasks to help identify anomaly pixels, recently there also appears new paradigm which exploits reconstruction error to highlight anomalies.

Uncertainty Estimation. Similar to image-level anomaly detection approaches, early uncertainty based methods (Hendrycks and Gimpel 2016; Lee et al. 2017) focused on measuring with maximum softmax probability since the model tends to output uniform prediction for unseen semantics. However, they are prone to misclassify pixels of tail classes as anomalies, since the same threshold is set for all pixels regardless of the class-wise discrepancy. To address this issue, Jung *et al.* (Jung et al. 2021) proposed standardized max logit (SML), which normalized the distribution of max logit from seen classes. Recent methods try to enhance ability of distinguishing hard samples from anomalous ones by re-training the classifiers with anomaly objectives. However, they generally suffer from accuracy decrease on seen categories (Bogdoll, Nitsche, and Zöllner 2022).

Outlier Exposure. Recent methods (Bevandić et al. 2018; Chan, Rottmann, and Gottschalk 2021; Di Biase et al. 2021b; Bevandić et al. 2019; Hendrycks, Mazeika, and Dietterich 2018) are intuitive, which utilize labeled samples from non-overlapped classes of an external dataset as anomalies to help models differentiate unexpected pixels against normal ones. Hendrycks et al. (Hendrycks, Mazeika, and Dietterich 2018) forced the model to predict uniform distribution of anomaly detection. (Chan, Rottmann, and Gottschalk 2021) and (Di Biase et al. 2021b) leveraged instance masks from COCO (Lin et al. 2014) and void category from Cityscapes (Cordts et al. 2016) to make models generalize to unexpected objects. However, they usually require re-training of the model and suffer from potential degradation in accuracy of in-distribution recognition. Besides, they require fine-grained annotation of anomalies (e.g. instance mask or bounding boxes), increasing the cost of labeling. Finally, utilizing specific datasets as outliers may lead the anomaly detectors biased toward specific domains, leading to degraded accuracy in real-world scenes.

Image Reconstruction. Methods based on image reconstruction (Lis et al. 2019b; Xia et al. 2020) usually employ generative adversarial networks (GANs) (Creswell et al. 2018) to fit the distribution of normal pixels, re-synthesize images conditioned on predicted segmentation results and localize the discrepancy between original images and reconstructed ones as anomalous objects. Nevertheless, these approaches usually heavily rely on the accurate segmentation maps and performance of reconstruction model, while it is still difficult for the segmentation models to distinguish hard in-distribution pixels and anomalous ones. On the other hand, their performance can be affected by the artifacts generated by GANs as well. Finally, they also suffer from time-consuming serialized training and inference processes of the reconstruction networks, making them hard to be applied in real-time scenarios (Bogdoll, Nitsche, and Zöllner 2022).

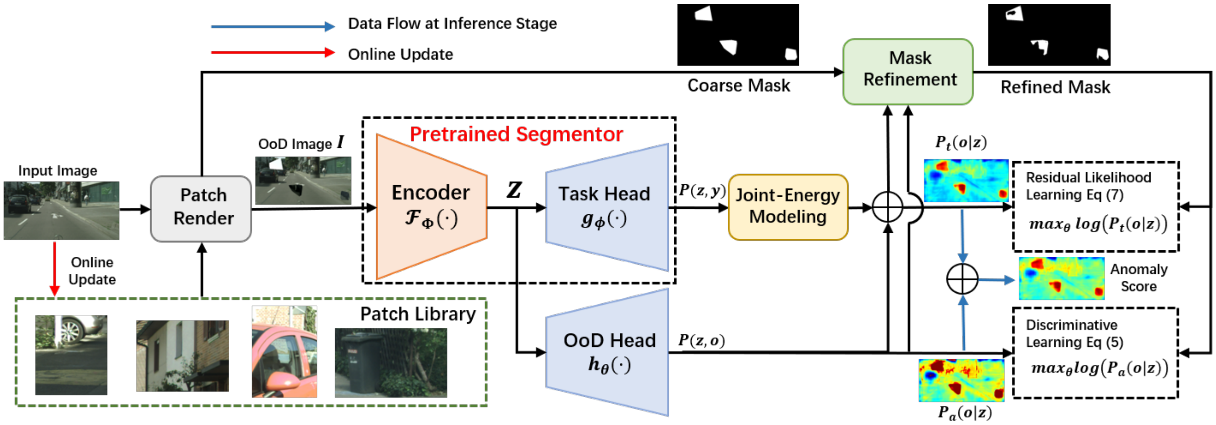
Data Flow at Inference Stage
Online Update
Coarse Mask
Mask Refinement
Refined Mask
Input Image
Patch Render
OoD Image I
Pretrained Segmentor
Encoder $\mathcal{F}_\Phi(\cdot)$
Z
Task Head $g_\phi(\cdot)$
$P(z,y)$
Joint-Energy Modeling
$P_t(o|z)$
Residual Likelihood Learning Eq (7)
$max_\theta\, log(P_t(o|z))$
Anomaly Score
Discriminative Learning Eq (5)
$max_\theta log(P_a(o|z))$
$P_a(o|z)$
OoD Head $h_\theta(\cdot)$
$P(z,o)$
Online Update
Patch Library

Figure 2: Illustration of proposed SLEEG framework, an OoD head is extended and trained in a self-supervised manner to enable a pretrained segmentation model with anomaly detection ability.

Energy Based Modeling

There is also a series of methods applying energy function (LeCun et al. 2006) to depict probabilities of anomaly data. These approaches apply a free-energy function (LeCun et al. 2006) as anomaly indicators and focus on minimizing energy for normal instances while maximizing energy for outlier samples. And the energy value is taken as measurement to predict anomaly probability of samples. Previous energy-based models generally employ Markov Chain Monte Carlo to estimate energy score whereas high-quality samples cannot be generated in this manner. To address this issue, (Tian et al. 2021) takes the insight from absenting learning, and utilizes the joint-energy (Grathwohl et al. 2020) with smooth terms to help switch between normal classification task and anomaly detection, while still requiring outlier exposure strategy with fine-grained annotation.

Methodology

In this section we introduce the detail of SLEEG framework, which is depicted as Fig. 2. Given an image $\mathcal{I} \in \mathbb{R}^{3\times H\times W}$ with H, W indicating its spatial resolution, its spatial coordinate set is defined as Ω, we associate the pixel x_ω at each coordinate $\omega \in \Omega$ with a triplet variable $(z_\omega, y_\omega, o_\omega)$, where $z_\omega \in \mathbb{R}^D$ represents the encoded feature of dimension D, $y_\omega \in \{0, 1, \cdots, K-1\}$ denotes predicted categorical labels over K close-set semantic classes, and $o_\omega \in \{0, 1\}$ is a binary indicator to denote whether x_ω belongs to anomaly. Our approach follows the classical Encoder-Decoder meta-architecture in semantic segmentation, where an encoder $\mathcal{F}_\Phi(\cdot) : \mathbb{R}^{3\times H\times W} \rightarrow \mathbb{R}^{D\times H\times W}$ extracts deep features and a segmentation decoder $g_\phi(\cdot) : \mathbb{R}^D \rightarrow \mathbb{R}^K$ is responsible to predict categorical distribution, both Φ, ϕ can be pretrained parameters from off-the-shelf segmentation models. Similarly, we further extend a learnable anomaly decoder $h_\theta(\cdot) : \mathbb{R}^D \rightarrow \mathbb{R}^2$ (termed as OoD head in Fig. 2) from original segmentation model, the output of which is utilized to derive two different anomaly estimators. The learnable parameter θ is trained in a self-supervised pipeline by maximizing the estimator scores in pseudo anomaly areas.

Anomaly Estimators for Likelihood Maximization

In anomaly detection, our goal is to accurately model the anomaly likelihood given input data $p(o_\omega|z_\omega)$. To this end, we resort to the Bayes Rule to derive the conditional probability, however, since the feature encoding z_ω is jointly modeled with both semantic y_ω and anomaly o_ω, the likelihood probability can be derived with different marginalization

$$p(o_\omega|z_\omega) = \frac{p(z_\omega, o_\omega)}{\sum_{o_\omega=0}^{1} p(z_\omega, o_\omega)} = \frac{p(z_\omega, o_\omega)}{\sum_{y_\omega=1}^{K} p(z_\omega, y_\omega)} \tag{1}$$

This converts the estimation from conditional distribution $p(o_\omega|z_\omega)$ to the joint distribution of $p(z_\omega, o_\omega)$ and $p(z_\omega, y_\omega)$. To analytically estimate the joint distribution, we recap the energy-based model (LeCun et al. 2006; Grathwohl et al. 2020) by reinterpreting the decoder $g_\phi(\cdot)$ as an energy function, which estimates the joint distribution $p(z_\omega, y_\omega)$

$$p(z_\omega, y_\omega; \phi) = \frac{1}{\mathcal{T}(\phi)} \exp(g_\phi(z_\omega))[y_\omega] \tag{2}$$

where $[y_\omega]$ is the y_ω-th index of output categorical vector, and $\mathcal{T}(\phi) = \int_{z_\omega} \sum_{y_\omega} \exp(g_\phi(z_\omega))[y_\omega] dz_\omega$ is an unknown normalization factor. Similarly, the learnable anomaly decoder $h_\theta(\cdot)$ can be regarded as an energy function to estimate the distribution of $p(z_\omega, o_\omega)$

$$p(z_\omega, o_\omega; \theta) = \frac{1}{\Gamma(\theta)} \exp(h_\theta(z_\omega))[o_\omega] \tag{3}$$

where $\Gamma(\theta)$ is another constant factor similar to $\mathcal{T}(\phi)$. By inserting Eq. (3) into Eq. (1), we can derive analytical representation of $p(o_\omega|z_\omega)$ which only focuses on anomalies regardless of segmentation tasks. On the other hand, when inserting both Eq .(3) and Eq .(2) into Eq. (1), we essentially take semantic distribution $p(z_\omega, y_\omega)$ into account for estimation. Hence we obtain two different anomaly estimators.

Task Agnostic Estimator (TAE). By taking the anomaly segmentation as a pixel-wise binary classification problem, we can easily derive the likelihood from Eq. (3) and Eq. (1) with the normalization factor $\Gamma(\theta)$ eliminated

$$p_a(o_\omega|z_\omega; \theta) = \frac{\exp(h_\theta(z_\omega))[o_\omega]}{\sum_{o_\omega \in \{0,1\}} \exp(h_\theta(z_\omega))[o_\omega]} \tag{4}$$

Since Eq. (4) is a normalized probability function, we can optimize this estimator via simple cross-entropy loss

$$\mathcal{L}_a(\theta) = -\, E_{x_\omega \in \mathcal{S}_{ood}} [\log p_a(o_\omega = 1|z_\omega;\theta)] \\ -\, E_{x_\omega \in \mathcal{S}_{id}} [\log p_a(o_\omega = 0|z_\omega;\theta)] \qquad (5)$$

where $\mathcal{S}_{ood}$ denote the set of anomaly pixels and $\mathcal{S}_{id}$ is the set of normal pixels.

Task Oriented Residual Estimator (TORE). When taking joint probability $p(z_\omega, y_\omega)$ from Eq. (2) for marginalization in Eq. (1), the estimated likelihood $p_t(o_\omega|z_\omega;\theta)$ is coupled with constants $\Gamma(\theta)$ and $\mathcal{T}(\phi)$, which is intractable, hence we transform the likelihood into logarithmic form

$$\log p_t(o_\omega|z_\omega;\theta) = h_\theta(z_\omega)[o_\omega] + \text{JEM}(z_\omega) + C(\phi,\theta) \\ \text{JEM}(z_\omega) = -\log \sum_{y_\omega} \exp(g_\phi(z_\omega))[y_\omega] \qquad (6)$$

where $C(\phi,\theta) = \log(\mathcal{T}(\phi)/\Gamma(\theta))$ is a constant w.r.t z_ω. Note that the second term in Eq. (6) is exactly the negative joint-energy model (JEM) (Grathwohl et al. 2020), which can be regarded as "coarse" estimation of uncertainty, therefore Eq. (6) essentially takes decoder h_θ to estimate the residual of JEM score, hence to "refine" the anomaly area. To optimize the estimator while handling intractable factor $C(\phi,\theta)$, we exploit a margin loss to compare the likelihood of anomaly and normal pixels and eliminate additive constant $C(\phi,\theta)$

$$\mathcal{L}_o(\theta) = \{E_{x_\omega \in \mathcal{S}_{id}}[\log p_t(o_\omega = 1|z_\omega;\theta)] - \\ E_{x_{\omega'} \in \mathcal{S}_{ood}}[\log p_t(o_{\omega'} = 1|z_{\omega'};\theta)] + \gamma\}_+ \qquad (7)$$

where γ is a hyperparameter to control the margin, $\{\cdot\}_+$ indicates truncating the value to 0 if it is negative.

Difference from other energy-based methods. Note that the loss term in Eq. (7) can be reformulated as[1]

$$\mathcal{L}_o(\theta) = \{E_{x_\omega \in \mathcal{S}_{id}}[h_\theta(z_\omega)] - E_{x_{\omega'} \in \mathcal{S}_{ood}}[h_\theta(z_{\omega'})] + \widehat{\gamma}\}_+ \\ \widehat{\gamma} = \gamma + E_{x_\omega \in \mathcal{S}_{id}}[\text{JEM}(z_\omega)] - E_{x_{\omega'} \in \mathcal{S}_{ood}}[\text{JEM}(z_{\omega'})] \quad (8)$$

Therefore, our TORE is different from other methods that directly update classifiers to optimize JEM (Grcić, Bevandić, and Šegvić 2022; Tian et al. 2021), instead it takes estimated joint-energy to dynamically control $\widehat{\gamma}$ of anomaly confidence between anomaly and normal pixels. This dynamic margin $\widehat{\gamma}$ is also verified more helpful than a static margin γ in the ablation study. Finally, the predicted anomaly score can be expressed as the combination of both estimators

$$\mathcal{A}(x_\omega) = \log p_a(o_\omega|z_\omega;\theta^*) + \lambda \log p_t(o_\omega|z_\omega;\theta^*) \qquad (9)$$

where λ is a balance factor and θ^* denotes the parameters of OoD head after optimization. In Eq. (9) we omit the constant term $C(\phi,\theta)$ since it does not affect the relative order of different pixels.

Self-supervised Training with Refined Patch

In this section we describe our self-training pipeline, which is illustrated in Fig. 2. Different from classical outlier exposure (Bevandić et al. 2021; Di Biase et al. 2021a), we aim at generating training samples with both anomaly pixels $\mathcal{S}_{ood}$ and normal area $\mathcal{S}_{id}$ in a self-supervised manner without manual mask annotation.

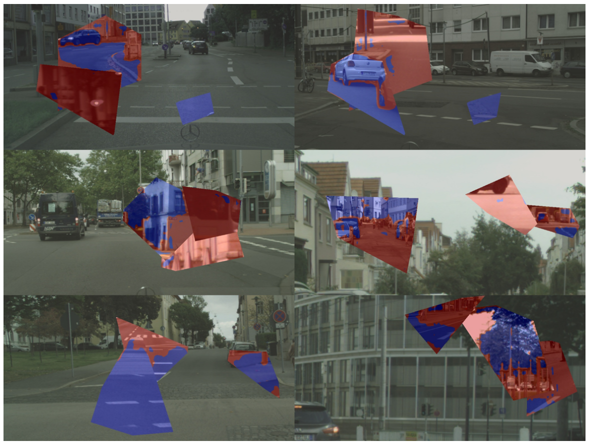

Figure 3: Visualization of generated patches with random shapes on training set of Cityscapes. Area with redred mask denotes anomaly pixels after mask refinement, the blueblue area represents the ignored pixels from pasted patches.

Anomaly Patch Rendering. Intuitively, due to the contextual reliance of the segmentation network, a random patch can be an anomaly part once placed under somewhat unnatural pattern, even if the patch is cropped from a normal image. With this consideration, we design a dynamic copy-and-paste strategy to generate pseudo anomaly samples (as shown in Fig. 2). In detail, for each input image, we first randomly crop N rectangle patches from other images as candidates. To ensure the geometric diversity of anomalies, we extract the Harris corner points (Harris, Stephens et al. 1988) within each candidate and crop the minimum polygon containing all these points as anomaly patch, finally, all cropped polygons are randomly pasted on input image.

Adaptive Mask Refinement. As the patches are randomly pasted, they can still contain objects which fits the context well. Taking these pixels as anomaly will reversely hinder the detection results. Therefore, to fully leverage the contextual information, we take the estimated anomaly likelihood $\mathcal{A}(x_\omega)$ as guidance to measure the inconsistency between context and pasted patches and further refine the pasted polygons. Formally, we define the coordinates set of the i-th pasted polygon area as Ω_i^p, and aim at finding a threshold η^* to separate out pixels more likely to be anomalies

$$\mathcal{S}_{ood} = \{x_\omega | \omega \in \cup_{i=1}^N \Omega_i^p \wedge \mathcal{A}(x_\omega) \geq \eta^*\} \qquad (10)$$

To properly refine the pasted patch, we design the threshold η^* in a dynamic manner such that pixels within each separated group share similar anomaly likelihood. Therefore we search the threshold by minimizing the variance of anomaly scores within each pixel group in pasted area

$$\eta^* = \arg\min_\eta Var_{x_\omega \in \mathcal{S}_{ood}}[\mathcal{A}(x_\omega)] + Var_{x_\omega \notin \mathcal{S}_{ood}}[\mathcal{A}(x_\omega)] \\ \textbf{s.t.} \quad \omega \in \cup_{i=1}^N \Omega_i^p \quad \min_\omega \mathcal{A}(x_\omega) \leq \eta \leq \max_\omega \mathcal{A}(x_\omega) \quad (11)$$

[1] For simplicity, we ignore the index symbol $[o_\omega = 1]$ here.

Method	OoD Data	Re-training	FS LAF		FS Static	
			$\text{FPR}_{95} \downarrow$	AP ↑	$\text{FPR}_{95} \downarrow$	AP ↑
Synboost (Di Biase et al. 2021a)	✓	✗	15.79	43.22	18.75	72.59
Density - Logistic Regression (Blum et al. 2021)	✓	✓	24.36	4.65	13.39	57.16
Bayesian Deeplab (Mukhoti and Gal 2018)	✗	✓	38.46	9.81	15.50	48.70
OoD Training - Void Class (Blum et al. 2021)	✓	✓	22.11	10.29	19.40	45.00
Discriminative Outlier Detection Head (Bevandić et al. 2021)	✓	✓	19.02	31.31	**0.29**	**96.76**
Dirichlet Deeplab (Malinin and Gales 2018)	✓	✓	47.43	34.28	84.60	31.30
DenseHybrid† (Grcić, Bevandić, and Šegvić 2022)	✓	✓	6.18	43.90	5.51	72.27
PEBAL† (Di Biase et al. 2021b)	✓	✓	7.58	44.17	1.73	92.38
CoroCL† (Liu et al. 2023)	✓	✓	**2.27**	**53.99**	0.52	95.96
MSP (Hendrycks and Gimpel 2017)	✗	✗	44.85	1.77	39.83	12.88
Entropy (Hendrycks and Gimpel 2017)	✗	✗	44.83	2.93	39.75	15.41
Density - Single-layer NLL (Blum et al. 2021)	✗	✗	32.90	3.01	21.29	40.86
kNN Embedding - density (Blum et al. 2021)	✗	✗	30.02	3.55	20.25	44.03
Density - Minimum NLL (Blum et al. 2021)	✗	✗	47.15	4.25	17.43	62.14
Image Resynthesis (Lis et al. 2019a)	✗	✗	48.05	5.70	27.13	29.60
SML (Jung et al. 2021)	✗	✗	21.52	31.05	19.64	53.11
GMMSeg (Liang et al. 2022)	✗	✗	6.61	55.63	15.96	**76.02**
FED-U (Gudovskiy, Okuno, and Nakata 2023)	✗	✗	11.38	20.45	21.58	67.80
SLEEG (ours)	✗	✗	6.69	**59.66**	**10.49**	68.93

Table 1: Comparison with previous methods on the FS test set. "OoD Data" indicates training with additional labeled anomaly data. "Re-training" means updating parameters of segmentation network. † indicates training with WideResNet38. Bold values and underlined values represent the best and second best results (Comparison is conducted within each group).

The problem in Eq. (11) is solved in a grid scanning manner, which is equivalent to finding the maximum gap between mean values of groups (Otsu 1979). After the refinement, the pasted areas not attributed to $\mathcal{S}_{ood}$ is labeled as ignored and will not be used for loss computation. On the other hand, we take all pixels outside the pasted area as normal set $\mathcal{S}_{id}$. Fig. 3 shows some examples after our refinement, it can be clearly observed that some part fitting the context well is automatically filtered out and other areas that are more contradictory to context are left for training. With the set separation above, we can reversely apply Eq. (5) and Eq. (7) to train the OoD head in a self-supervised paradigm.

Experiments

Datasets

We evaluate SLEEG in several widely used anomaly datasets: FishyScapes (FS) Lost & Found (Blum et al. 2021), FishyScapes (FS) Static (Blum et al. 2021), Road Anomaly (Lis et al. 2019a), Segment-Me-If-You-Can(SMIYC)(Chan et al. 2021) and StreetHazards (Hendrycks et al. 2019b).

FS Lost & Found contains high-resolution samples for autonomous driving scenes. FS Lost & Found is built based on the Lost & Found (Pinggera et al. 2016) and adopts the same collection setup as Cityscapes (Cordts et al. 2016), which is also a segmentation benchmark for urban scenes. Specifically, 37 types of unexpected real road obstacles and 13 different street scenarios are included. A validation set of 100 samples is publicly available, while the hidden test set with 275 images is unknown. All the methods need to submit the code to the website[2] to evaluate on this test set.

FS Static is artificially built upon the validation set of Cityscapes, where unexpected objects are collected from the Pascal VOC dataset (Everingham et al. 2010) and positioned randomly on the images. Specifically, only objects not belonging to the pre-defined classes of Cityscapes are used. This dataset consists of a public validation set with 30 images and a hidden test set with 1,000 images.

Road Anomaly includes real-world images collected online, where anomalous obstacles encounter on or locate near the road. All the images are re-scaled to a size of 1,280 × 720 and pixel-wise annotations of unexpected objects are provided. Since there exists larger domain gap between Road anomaly and Cityscapes, generalization ability of models is essential to the performance on this dataset, following previous works (Tian et al. 2021; Liang et al. 2022), we evaluate SLEEG on the test set consisting of 60 images.

Other Benchmarks. Besides the datasets above, we further evaluate on other benchmarks including SMIYC (Chan et al. 2021) and Streethazards (Hendrycks et al. 2019b), which contain anomalies collected either from real-world or virtual game engine. The corresponding results can be found in supplementary material.

Implementation Details

For fair comparisons, we follow the similar settings of (Jung et al. 2021; Di Biase et al. 2021b) to utilize the segmentation model of DeepLab series (Chen et al. 2018) with ResNet101 (He et al. 2016) backbone, which is pre-trained on Cityscapes and keep fixed without further re-training. The lightweight OoD head consists of three stacked Conv-BN-ReLU blocks and is trained for $40,000$ iterations with batchsize of 8, the balance factor is set as $\lambda = 0.5$ for

[2] https://fishyscapes.com/results

Method	OoD Data	FS LAF Validation		FS Static Validation		Road Anomaly Test	
		$\mathrm{FPR}_{95}\downarrow$	AP ↑	$\mathrm{FPR}_{95}\downarrow$	AP ↑	$\mathrm{FPR}_{95}\downarrow$	AP ↑
Synboost (Di Biase et al. 2021a)	✓	34.47	40.99	47.71	48.44	59.72	41.83
DenseHybrid† (Grcić, Bevandić, and Šegvić 2022)	✓	6.10	63.80	4.90	60.00	-	-
PEBAL† (Tian et al. 2021)	✓	4.76	58.81	**1.52**	**99.61**	**44.58**	**45.10**
CoroCL† (Liu et al. 2023)	✓	**0.85**	**92.46**	2.52	70.61	-	-
MSP (Hendrycks and Gimpel 2017)	✗	45.63	6.02	34.10	14.24	68.44	20.59
MaxLogit (Hendrycks et al. 2019a)	✗	38.13	18.77	28.50	27.99	64.85	24.44
SynthCP (Xia et al. 2020)	✗	45.95	6.54	34.02	23.22	64.69	24.87
SML (Jung et al. 2021)	✗	14.53	36.55	16.75	48.67	49.74	25.82
GMMSeg (Liang et al. 2022)	✗	13.11	43.47	-	-	47.90	34.42
FED-U (Gudovskiy, Okuno, and Nakata 2023)	✗	11.35	37.05	20.15	46.32	-	-
SLEEG (ours)	✗	**10.90**	**70.90**	**3.85**	**77.23**	**42.60**	**38.10**

Table 2: Comparison on FS validation sets and Road Anomaly. "OoD Data" indicates the method adopts additional labeled data as anomalies for training. † indicates training with WideResNet38 backbone. Bold values represent the best results (Comparison is conducted within each group) .

test. And we apply a warmup strategy for mask refinement. First we take JEM as anomaly score for $6,000$ iterations and then replace it with estimated anomaly likelihood. Following common setting (Jung et al. 2021; Blum et al. 2021; Xia et al. 2020), the average precision (AP) and false positive rate (FPR_{95}) at true positive rate of 95% are adopted as metrics to perform comprehensive evaluation. Among these metrics, FPR_{95} and AP are more crucial since there exists severe imbalance between anomaly and normal pixels.

Experiments Results

FS Leaderboard. Tab. 1 provides results on test sets of FS benchmark. Following (Blum et al. 2019), previous methods are categorized based on whether they require re-training the segmentation network or extra OoD data. Compared with previous methods without re-training or extra anomaly data, SLEEG works effectively to outperform most competitors by large margins. SLEEG can even surpass some methods with re-training and extra anomaly data (Tian et al. 2021; Di Biase et al. 2021a) by large margins and achieve a new SOTA performance of AP on Lost & Found track, which includes anomalies in more realistic scenes. On the artificially-generated FS Static, SLEEG is on par with the SOTA method DenseHybrid, which requires extra data and adopts stronger Wide-ResNet as its feature extractor. Although SLEEG is inferior to some methods relying on extra data in FS Static, this is due to the external data (Lin et al. 2014) adopted falls into similar distribution of anomalies in FS Static (Everingham et al. 2010), thus they suffer a large performance gap between artificial Static track and realistic LAF, while SLEEG keeps superiority on both tracks.

FS Validation Sets. Tab. 2 shows comparisons on the validation sets of FS Lost & Found and Static. In terms of AP, our anomaly detector achieves the best results for all metrics on FS Lost & Found and competitive performance on FS Static. Specifically, when compared with GMMSeg (Liang et al. 2022), SLEEG yields significant improvements of 26.4% on AP and reduce FPR_{95} to 10.9%. The results demonstrate the effectiveness and robustness of our SLEEG on detecting real-world anomaly instances. Similar to results in Tab. 1, methods with auxiliary anomaly data (*e.g.* PEBAL) or synthetic model (*e.g.* SynthCP) usually suffer from significant performance gap between artificial anomaly in FS Static and anomalies in realistic scenes from LAF, since they tend to overfit label anomalies. In contrast, SLEEG can yield relative consistent performance gain.

Patch Policy	Refinement	FS LAF Validation		Road Anomaly Test	
		$\mathrm{FPR}_{95}\downarrow$	AP ↑	$\mathrm{FPR}_{95}\downarrow$	AP ↑
Void as OoD	-	19.3	47.3	69.1	21.3
Square Patch	✗	25.4	45.7	67.9	18.2
	✓	**13.8**	**62.5**	**47.8**	**34.8**
Convex Patch	✗	13.3	61.9	44.8	35.9
	✓	**10.9**	**98.3**	**42.6**	**38.1**

Table 3: Investigation of patch policies on FS and Road Anomaly validation set. "Void as OoD" denotes training with pixels of void class in Cityscapes as anomaly samples.

Road Anomaly Test Sets. We further compare SLEEG with recent advanced anomaly segmentation methods on Road Anomaly in Tab. 2, it is observed that SLEEG outperforms most competitors by a large margin when no labeled anomaly data is available. Since there exists larger inherent domain shift between Road anomaly and Cityscapes than Fishyscapes, previous methods (*e.g.* SML (Jung et al. 2021)) that perform well on Fishyscapes are prone to suffer from poor accuracy on Road anomaly. However, our SLEEG yields significant improvements on both two datasets, demonstrating the robustness of SLEEG in tackling open-world scenes with diverse styles.

In-depth Discussion

Effectiveness of Different Estimators. We explore the contribution of TAE and TORE on FS Lost & Found and Road Anomaly test sets. We take the pure JEM (Grathwohl et al. 2020) measurement as baseline detector. It can be observed

Estimator		FS LAF Validation		Road Anomaly Test	
TAE	TORE	$\text{FPR}_{95}\downarrow$	AP$\uparrow$	$\text{FPR}_{95}\downarrow$	AP$\uparrow$
✗	✗	27.0	31.0	79.9	18.9
✓	✗	25.7	57.5	47.6	33.5
✗	✓	18.4	47.9	51.2	31.6
✓	✓	**10.9**	**70.9**	**42.6**	**38.1**

Table 4: Ablation results for different likelihood estimators on FS validation set and Road Anomaly validation set.

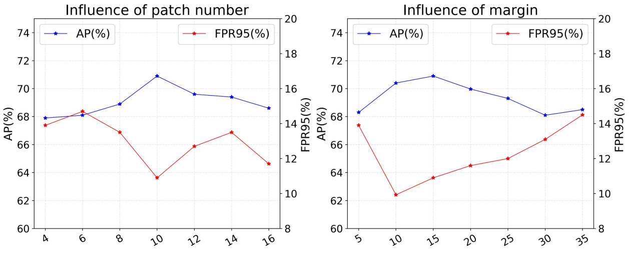

Figure 4: Investigation on the influence on AP and false positive rate with varied patch number N (left) and margin value γ (right) on FS Lost & Found validation set.

from Tab. 4 that TAE brings significant improvements to the baseline by training with self-generated OoD samples. Further, by dynamically adjusting the margin of anomaly scores, TORE also brings a large performance boost of around 13% in AP and significant reduction in FPR_{95} as well. Finally, SLEEG is capable of achieving the best performance by seamlessly integrating two estimators.

Specifically, as indicated by Eq. (8), we also compare the automatic energy-guided margin $\widehat{\gamma}$ with a manually tuned static value γ for the training of TORE, which is equivalent to eliminating the JEM terms in Eq. (8). For both setting, we set $\gamma = 15$, the results can be illustrated in Tab. 6. It is clear that the margin with a dynamically controlled component as Eq. (8) can consistently outperforms the static margin, indicating the residual form of anomaly likelihood can adaptively capture effective training samples.

Investigation on Patch Policy. Tab. 3 shows the comparison results for different patch policies on FS Lost & Found validation set and road anomaly validation set. We design a baseline which regards pixels labeled with "void" from Cityscapes as auxiliary anomaly data (denoted as "Void as OoD"). From Tab. 3, training with simple square patches performs slightly inferior to baseline, since this is prone to implicitly learn useless shape-related prior. With more various shapes, convex patches brings significant improvements. Besides, by performing our refinement strategy, SLEEG can also achieve significant improvement even with simple square patch. Finally, imposing further refinement on the convex patch achieves the best performance.

Parameter Sensitivity. We further investigate the influence of patch number and margin value on the FS Lost & Found validation set. As shown in Fig. 4, SLEEG performs best when margin $\gamma = 15$ and patch number $N = 10$. The results demonstrate that SLEEG achieves similar performance across all settings, implying SLEEG is not very sensitive to both patch number and margin value. Finally, we set patch number and margin to 10 and 15 respectively.

Model	Method	FS LAF Val		FS Static Val	
		$\text{FPR}_{95}\downarrow$	AP$\uparrow$	$\text{FPR}_{95}\downarrow$	AP$\uparrow$
OCR Net	SML	18.28	39.96	15.07	47.90
	JEM	22.90	23.27	16.80	34.03
	Void Class	16.65	46.62	17.74	29.30
	SLEEG	**8.89**	**72.51**	**7.6**	**73.01**
ISA Net	SML	18.67	28.76	14.86	32.15
	JEM	35.57	22.65	16.22	45.22
	Void Class	29.17	43.53	17.18	24.61
	SLEEG	**12.28**	**65.79**	**3.33**	**80.03**
FCN	SML	39.80	17.59	14.53	28.92
	JEM	39.36	21.83	13.47	30.51
	Void Class	24.29	43.44	14.79	22.72
	SLEEG	**21.01**	**63.11**	**5.31**	**61.29**

Table 5: Comparison between SLEEG and other methods on FS validation sets with different segmentation models. "Void Class" denotes training models with pixels that fall into void class as anomaly samples. The mIoU for OCRNet (Yuan and Wang 2020), ISANet (Huang et al. 2019) and FCN (Shelhamer 2017) are 80.66, 80.51, 77.72 respectively.

Margin Type	FS LAF Validation		Road Anomaly Test	
	$\text{FPR}_{95}\downarrow$	AP$\uparrow$	$\text{FPR}_{95}\downarrow$	AP$\uparrow$
Static γ	12.3 1	65.2	56.2	34.4
Dynamic $\widehat{\gamma}$	**10.9**	**70.9**	**42.6**	**38.1**

Table 6: Ablation results of comparing static/dynamic margin (Eq. (8)) on FS and Road Anomaly validation set.

Extension to More Segmentation Models. Finally, as shown in Tab. 5, we also validate the generalization ability of SLEEG by training with different advanced segmentation frameworks, including OCRNet (Yuan and Wang 2020), ISANet (Huang et al. 2019) and FCN (Shelhamer 2017), where SLEEG consistently surpasses SML and JEM across all the frameworks with more than 20% AP improvement on average. Additionally, it is observed that SLEEG performs better on FS Lost& Found validation set when adopting networks with higher mIoU scores, implying that SLEEG can work effectively with various frameworks.

Conclusion

In this paper, we propose SLEEG, a simple and flexible anomaly segmentation model without re-training or labeled anomaly data, which exploits a task-agnostic binary estimator, and a task-oriented energy residual estimator for anomaly likelihood estimation, and incorporate them with an adaptive copy-and-paste mask policy for self-supervised learning. Extensive experimental results verify the effectiveness of our method and competitive performance is achieved on both FS Lost & Found validation and test sets by SLEEG.

Acknowledgements

This work was supported by National Natural Science Fund of China (62076184, 61976158, 61976160, 62076182, 62276190), in part by Fundamental Research Funds for the Central Universities and State Key Laboratory of Integrated Services Networks (Xidian University), in part by Shanghai Innovation Action Project of Science and Technology (20511100700) and Shanghai Natural Science Foundation (22ZR1466700).

References

Bevandić, P.; Krešo, I.; Oršić, M.; and Šegvić, S. 2018. Discriminative out-of-distribution detection for semantic segmentation. *arXiv preprint arXiv:1808.07703*.

Bevandić, P.; Krešo, I.; Oršić, M.; and Šegvić, S. 2019. Simultaneous semantic segmentation and outlier detection in presence of domain shift. In *German conference on pattern recognition*, 33–47. Springer.

Bevandić, P.; Krešo, I.; Oršić, M.; and Šegvić, S. 2021. Dense outlier detection and open-set recognition based on training with noisy negative images. *arXiv preprint arXiv:2101.09193*.

Blum, H.; Sarlin, P.-E.; Nieto, J.; Siegwart, R.; and Cadena, C. 2019. Fishyscapes: A benchmark for safe semantic segmentation in autonomous driving. In *Proc. of the IEEE/CVF International Conference on Computer Vision Workshops*, 0–0.

Blum, H.; Sarlin, P.-E.; Nieto, J.; Siegwart, R.; and Cadena, C. 2021. The fishyscapes benchmark: Measuring blind spots in semantic segmentation. *International Journal of Computer Vision*, 129(11): 3119–3135.

Bogdoll, D.; Nitsche, M.; and Zöllner, J. M. 2022. Anomaly Detection in Autonomous Driving: A Survey. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4488–4499.

Chan, R.; Lis, K.; Uhlemeyer, S.; Blum, H.; Honari, S.; Siegwart, R.; Fua, P.; Salzmann, M.; and Rottmann, M. 2021. SegmentMeIfYouCan: A Benchmark for Anomaly Segmentation.

Chan, R.; Rottmann, M.; and Gottschalk, H. 2021. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 5128–5137.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proc. European Conference on Computer Vision (ECCV)*, 801–818.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3213–3223.

Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; and Bharath, A. A. 2018. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1): 53–65.

Di Biase, G.; Blum, H.; Siegwart, R.; and Cadena, C. 2021a. Pixel-wise Anomaly Detection in Complex Driving Scenes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 16918–16927.

Di Biase, G.; Blum, H.; Siegwart, R.; and Cadena, C. 2021b. Pixel-wise anomaly detection in complex driving scenes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 16918–16927.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338.

Grathwohl, W.; Wang, K.-C.; Jacobsen, J.-H.; Duvenaud, D.; Norouzi, M.; and Swersky, K. 2020. Your classifier is secretly an energy based model and you should treat it like one. In *Proc. International Conference on Learning Representations (ICLR)*.

Grcić, M.; Bevandić, P.; and Šegvić, S. 2022. DenseHybrid: Hybrid Anomaly Detection for Dense Open-set Recognition. *arXiv preprint arXiv:2207.02606*.

Gudovskiy, D.; Okuno, T.; and Nakata, Y. 2023. Concurrent Misclassification and Out-of-Distribution Detection for Semantic Segmentation via Energy-Based Normalizing Flow. *arXiv preprint arXiv:2305.09610*.

Harris, C.; Stephens, M.; et al. 1988. A combined corner and edge detector. In *Alvey vision conference*, volume 15, 10–5244. Manchester, UK.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hendrycks, D.; Basart, S.; Mazeika, M.; Mostajabi, M.; Steinhardt, J.; and Song, D. 2019a. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*.

Hendrycks, D.; Basart, S.; Mazeika, M.; Mostajabi, M.; Steinhardt, J.; and Song, D. 2019b. Scaling Out-of-Distribution Detection for Real-World Settings.

Hendrycks, D.; and Gimpel, K. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Hendrycks, D.; and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *Proc. International Conference on Learning Representations (ICLR)*.

Hendrycks, D.; Mazeika, M.; and Dietterich, T. 2018. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*.

Huang, L.; Yuan, Y.; Guo, J.; Zhang, C.; Chen, X.; and Wang, J. 2019. Interlaced Sparse Self-Attention for Semantic Segmentation. *arXiv preprint arXiv:1907.12273*.

Jung, S.; Lee, J.; Gwak, D.; Choi, S.; and Choo, J. 2021. Standardized max logits: A simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 15425–15434.

LeCun, Y.; Chopra, S.; Hadsell, R.; Ranzato, M.; and Huang, F. 2006. A tutorial on energy-based learning. *Predicting structured data*, 1(0).

Lee, K.; Lee, H.; Lee, K.; and Shin, J. 2017. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*.

Liang, C.; Wang, W.; Miao, J.; and Yang, Y. 2022. GMMSeg: Gaussian Mixture based Generative Semantic Segmentation Models. *arXiv preprint arXiv:2210.02025*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proc. European Conference on Computer Vision (ECCV)*, 740–755. Springer.

Lis, K.; Nakka, K.; Fua, P.; and Salzmann, M. 2019a. Detecting the Unexpected via Image Resynthesis. In *Proc. IEEE International Conference on Computer Vision (ICCV)*.

Lis, K.; Nakka, K.; Fua, P.; and Salzmann, M. 2019b. Detecting the unexpected via image resynthesis. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2152–2161.

Liu, Y.; Ding, C.; Tian, Y.; Pang, G.; Belagiannis, V.; Reid, I.; and Carneiro, G. 2023. Residual Pattern Learning for Pixel-wise Out-of-Distribution Detection in Semantic Segmentation. arXiv:2211.14512.

Malinin, A.; and Gales, M. J. 2018. Predictive Uncertainty Estimation via Prior Networks. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*.

Mukhoti, J.; and Gal, Y. 2018. Evaluating bayesian deep learning methods for semantic segmentation. *arXiv preprint arXiv:1811.12709*.

Otsu, N. 1979. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1): 62–66.

Pinggera, P.; Ramos, S.; Gehrig, S.; Franke, U.; Rother, C.; and Mester, R. 2016. Lost and found: detecting small road hazards for self-driving vehicles. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1099–1106. IEEE.

Shelhamer. 2017. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39(4): 640–651.

Tian, Y.; Liu, Y.; Pang, G.; Liu, F.; Chen, Y.; and Carneiro, G. 2021. Pixel-wise Energy-biased Abstention Learning for Anomaly Segmentation on Complex Urban Driving Scenes. *arXiv preprint arXiv:2111.12264*.

Vojir, T.; Šipka, T.; Aljundi, R.; Chumerin, N.; Reino, D. O.; and Matas, J. 2021. Road anomaly detection by partial image reconstruction with segmentation coupling. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 15651–15660.

Xia, Y.; Zhang, Y.; Liu, F.; Shen, W.; and Yuille, A. L. 2020. Synthesize then compare: Detecting failures and anomalies for semantic segmentation. In *Proc. European Conference on Computer Vision (ECCV)*, 145–161. Springer.

Yuan, Y.; and Wang, J. 2020. Object-Contextual Representations for Semantic Segmentation.