Dense Open-set Recognition with Synthetic Outliers Generated by Real NVP

Matej Grcić, Petra Bevandić and Siniša Segvić
Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia

Keywords: Open-set Recognition, Semantic Segmentation, Real NVP.

Abstract: Today's deep models are often unable to detect inputs which do not belong to the training distribution. This gives rise to confident incorrect predictions which could lead to devastating consequences in many important application fields such as healthcare and autonomous driving. Interestingly, both discriminative and generative models appear to be equally affected. Consequently, this vulnerability represents an important research challenge. We consider an outlier detection approach based on discriminative training with jointly learned synthetic outliers. We obtain the synthetic outliers by sampling an RNVP model which is jointly trained to generate datapoints at the border of the training distribution. We show that this approach can be adapted for simultaneous semantic segmentation and dense outlier detection. We present image classification experiments on CIFAR-10, as well as semantic segmentation experiments on three existing datasets (StreetHazards, WD-Pascal, Fishyscapes Lost & Found), and one contributed dataset. Our models perform competitively with respect to the state of the art despite producing predictions with only one forward pass.

1 INTRODUCTION

Early computer vision workflows involved handcrafted feature engineering and shallow discriminative models. However, despite hard work and notable scientific contribution, the resulting features (Lowe, 2004; Sánchez et al., 2013; Rosten and Drummond, 2006) were insufficient in many computer vision tasks. Emergence of deep convolutional models (Krizhevsky et al., 2012) enabled implicit feature engineering. Consequently, computer vision research is now focused on designing deep models which are trained in end-to-end fashion (Simonyan and Zisserman, 2015; He et al., 2016; Huang et al., 2017).

Today, deep models deliver state-of-the-art performance in most computer vision tasks. However, whenever there is a domain shift between the train and the test distributions, we witness overconfidence of incorrect predictions (Lakshminarayanan et al., 2017; Guo et al., 2017). This issue poses a direct threat to the safety of AI-based systems for healthcare (Xia et al., 2020), autonomous driving (Zendel et al., 2018) and other critical application fields. Therefore, open-set recognition (Bendale and Boult, 2015) becomes an increasingly important research objective. The desired models should be able to detect foreign samples while also fulfilling their primary discriminative task. This problem has been also addressed in several related settings such as anomaly detection (Andrews et al., 2016), uncertainty estimation (Malinin and Gales, 2018), and out-of-distribution (OOD) detection (Hendrycks and Gimpel, 2017). We focus on open-set recognition and note that the proposed approach can be generalized to other settings.

Formally, OOD detection can be viewed as a form of binary classification where the model has to predict whether a given sample belongs to the training dataset. In practice, this can be achieved by designing the model to predict an OOD score given the input. We assume that the score function $s(\mathbf{x}) : \mathbf{X} \rightarrow \mathbb{R}$ assigns OOD score to any given sample. A convenient baseline approach (Hendrycks and Gimpel, 2017) expresses the score function in terms of maximum softmax probability: $s_{MSP}(\mathbf{x}) = 1 - \max_c P_\theta(c|\mathbf{x})$. The score function can also be expressed with softmax entropy: $s_H(\mathbf{x}) = -\sum_i P_\theta(c_i|\mathbf{x}) \log P_\theta(c_i|\mathbf{x})$ (Hendrycks et al., 2019b). We evaluate our dense OOD detection approach both with $s_{MSP}(\mathbf{x})$ and $s_H(\mathbf{x})$.

Typically, we wish to achieve open-set recognition with a single forward pass in order to promote cross-task synergy and allow real-time inference. Unfortunately, the two tasks may negatively affect each other since this a form of the multi-objective optimization.

This paper presents an open-set recognition ap-

Grcić, M., Bevandić, P. and Segvić, S.
Dense Open-set Recognition with Synthetic Outliers Generated by Real NVP.
DOI: 10.5220/0010260701330143
In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2021) - Volume 4: VISAPP*, pages 133-143
ISBN: 978-989-758-488-6

proach based on jointly learned synthetic examples generated at the border of the training distribution (Lee et al., 2018). Instead of adversarial generation (Goodfellow et al., 2014), we prefer to base our approach on invertible normalized flow (Dinh et al., 2017) due to better distribution coverage (Lucas et al., 2019) and opportunity to generate images of arbitrary resolution. The resulting method outperforms GANs in equivalent image-wide experiments. We argue that the proposed approach is especially suitable for adaptation to dense prediction due to capability of normalizing flows to generate images of arbitrary size. We evaluate our models on several datasets for dense open-set recognition and demonstrate competitive performance with respect to the state of the art.

2 RELATED WORK

Previous work covers various facets of OOD detection. As usual in computer vision research, datasets and benchmarks are critical research tools since they help us identify effective approaches. Generative models are important for our work since they allow us to generate synthetic training samples for discriminative OOD detection.

2.1 Out-Of-Distribution Detection

The most popular OOD detection approaches include likelihood evaluation (Nalisnick et al., 2019), assesing the prediction confidence (DeVries and Taylor, 2018), exploiting an additional negative dataset (Hendrycks et al., 2019b), and modification of the loss function (Lee et al., 2018). In theory, OOD detection can be elegantly implemented by evaluating the likelihood of a given sample with a suitable generative model. A generative model capable of exact likelihood evaluation should assign OOD samples a lower likelihood. However, (Nalisnick et al., 2019) and (Serrà et al., 2020) show that generative models with exact likelihood evaluation tend to assign simple outliers a higher likelihood than to complex inliers. Exhaustive analysis shows that flow-based generative models are not suitable for this task due to their architecture (Kirichenko et al., 2020). Still, (Grathwohl et al., 2020) manage to detect outliers using the gradient of sample likelihood, while (Ren et al., 2019) detect outliers by evaluating likelihood ratios of two generative models. We do not use these approaches since our primary task is open-set semantic segmentation which requires specifically designed discriminative models for dense prediction.

OOD detection can also be expressed throughout a modification of the model architecture. This has been attempted by adding a confidence prediction head which is supposed to emit low confidence prediction for OOD samples (DeVries and Taylor, 2018). However, learned confidence is able to account only for difficulties which have been seen in the training data (aleatoric uncertainty), while remaining mostly unable to manage OOD inputs (epistemic uncertainty).

A preprocessing approach known as ODIN (Liang et al., 2018) does not require any change in model's training procedure. They add a small perturbation to the input which leads to better separation between in- and out-of-distribution samples. However, this approach results in only slight improvements over the max-softmax baseline (Bevandic et al., 2019). Additionally, it requires multiple passes trough the model, which considerably increases the inference latency.

Discriminative OOD detection can be improved by utilizing a diverse negative dataset (Hendrycks et al., 2019b; Bevandic et al., 2019). Additionally, outliers can be detected by observing the distributional uncertainty of prior networks (Malinin and Gales, 2018). We differ from these approaches, since we do not require training on negative data.

The negative dataset can be replaced by an adversarial network which generates artificial outliers (Lee et al., 2018; Nitsch et al., 2020). This setup requires that the discriminative classifier output uniform distribution in synthetic samples. Consequently, the generator network receives a signal from classifier's loss which moves generated samples to the border of the training distribution. Our approach is closely related with this method and hence we present an empirical comparison in the experiments. The main differences are that i) our method uses RNVP instead of GAN which allows us to generate outliers of different sizes, and ii) we present an adaptation for dense open-set recognition.

2.2 Dense Open-set Recognition

OOD detection can be combined with pixel-level classification to achieve dense open-set recognition. Testing open-set performance of dense discriminative models proved to be a difficult task. Available benchmarks fail to fully cover all open-world situations but still pose a challenge to present models. However, evaluating the model on multiple benchmarks can give us a better notion on open-set recognition performance. The WildDash benchmark (Zendel et al., 2018) evaluates capability of the model to deal with challenging real-world situations and outright negative images. We do not evaluate directly on this dataset since it does not include test

images with mixed content (both inliers and outliers). The Fishyscapes Lost & Found benchmark (Blum et al., 2019) evaluates the ability to detect small OOD objects on the road surface. The StreetHazards dataset (Hendrycks et al., 2019a) includes synthetic images created by the Unreal Engine, while the BDD-Anomaly dataset (Hendrycks et al., 2019a) is created from the Berkley Deep Drive dataset (BDD) (Yu et al., 2018) by selecting *motorcycle* and *train* classes as anomalies. The WD-Pascal dataset (Bevandic et al., 2019) contains WildDash images with pasted animals from the Pascal VOC (Everingham et al., 2010) dataset.

There are several prior approaches which adress dense open-set recognition. Epistemic uncertainty attempts to discern uncertainty due to insufficient knowledge from uncertainty due to insufficient supervision (Kendall and Gal, 2017). However, their approach assumes that MC dropout corresponds to Bayesian model sampling which may not be satisfied in practice. Additionally, Bayesian model sampling is unable to account for distributional shift (Malinin and Gales, 2018). OOD detection can also be framed as comparison between the original image and its sythesized version which is conditionally generated from the predictions (Xia et al., 2020). However image-to-image translation is a hard problem; OOD input is not a necessary condition for getting a poor reconstruction. Employing confidence of multiple DNNs trained in one-vs-all setting (Franchi et al., 2020) currently achieves the best OOD detection results on StreetHazards dataset. All these approaches require multiple forward passes through complex models and are therefore much slower than our approach.

Recognition of OOD input can also be improved by training on several positive and negative datasets. However that requires extraordinary efforts for making different datasets mutually compatible (Lambert et al., 2020). A more affordable approach assumes one specialized inlier dataset and one general purpose noisy negative dataset (Bevandic et al., 2019). Different than both these approaches, we do not use negative training data. Instead we jointly train a generative model for synthetic negative samples which we use for training of the discriminative model. Consequently, our approach does not incur any bias due to particular choice of the negative dataset.

2.3 Generative Modeling

Many generative models approximate the data distribution p_D using the model distribution p_θ defined by:

$$p_\theta(\mathbf{x}) = \frac{p'_\theta(\mathbf{x})}{Z}, \quad Z = \int_{\mathbf{x}} p'_\theta(\mathbf{x}) dx. \tag{1}$$

In the previous equation, p'_θ denotes the unnormalized distribution modeled with a deep model parameterized with θ, while Z is a normalization constant.

Restricted Boltzmann machines (RBM) (Salakhutdinov et al., 2007) learn the data distribution p_D by utilizing a two-phase training procedure. The positive phase increases the value of $p'_\theta(\mathbf{x})$ for every datapoint $\mathbf{x}_i$, while the negative phase updates the normalization constant Z in order to keep $p_\theta(\mathbf{x})$ a valid distribution.

Invertible normalizing flows (Dinh et al., 2015) are trained by likelihood maximization which essentially increases the value of $p_\theta(\mathbf{x})$ for every datapoint $\mathbf{x}_i$. This corresponds to the positive phase of RBM's training. Different than RBM's, normalizing flows require that the nonlinear transformation is bijective. In particular, normalizing flows consider an invertible transformation $f_\theta : \mathbf{X} \rightarrow \mathbf{Z}$ which maps the desired complex data distribution to a simpler latent distribution. Consequently, the change of variable formula: $p_\theta(\mathbf{x}) = p_z(f_\theta(\mathbf{x}))|\det\frac{\partial f_\theta(\mathbf{x})}{\partial \mathbf{x}}|$ can be applied. Since we transform normalized latent distribution with f_θ, which is bijective by design, the change of variables gives us a guarantee that the $p_\theta(\mathbf{x})$ will remain normalized. Hence, the negative phase for the flow-based models is unnecessary. Real NVP (RNVP) (Dinh et al., 2017) is a flow-based generative model which captures the dataset distribution by learning to bijectively transform it with a powerful set of affine transformations to a simpler latent distribution such as a unit Gaussian.

3 THE PROPOSED METHOD

We approach the task of open set recognition by exploiting synthetic negative samples produced by a RNVP model which is jointly trained with the open-set classifier. We aim to improve OOD detection without harming recognition accuracy. We first apply the proposed setup in the image-wide setup and later adapt for dense OOD detection.

3.1 Joint Training of Discriminative Classifier and Real NVP

We assume that OOD performance of a classifier can be improved by introducing an additional loss term that encourages it to emit the uniform distribution in synthetic outliers (Lee et al., 2018). This term can be expressed as Kullback-Leibler (KL) divergence between the model prediction in OOD samples and the uniform distribution. The resulting compound classi-

fier loss is:

$$L_{\text{cls}}(\theta_C) = -\mathbb{E}_{\hat{x},\hat{y}\sim P_{in}}[\log P_{\theta_C}(y=\hat{y}|\hat{x})] + \lambda \cdot \mathbb{E}_{x\sim P_{out}}[\text{KL}(U||P_{\theta_C}(y|x))] \,. \quad (2)$$

Note that the distribution P_{out} corresponds to synthetic outliers generated by RNVP. Hence, the classifier loss encourages RNVP to generate data which is classified with high entropy. However, we also apply the standard negative log-likelihood loss to the RNVP model:

$$L_{\text{RNVP}}(\theta_R) = -\mathbb{E}_{\mathbf{x}\sim P_{in}}[\log p_{\mathbf{z}}(\mathbf{f}_{\theta_R}(\mathbf{x})) + \log \left| \det \left(\frac{\partial \mathbf{f}_{\theta_R}(\mathbf{x})}{\partial \mathbf{x}} \right) \right|] \,. \quad (3)$$

This loss encourages RNVP to generate the data which resembles the training dataset. The two losses (L_{cls} and L_{RNVP}) are opposed, but together they cause the RNVP to generate data which resemble inliers but get classified to uniform distribution. Consequently, we colloquially state that our RNVP model generates samples at the border of the training distribution (Lee et al., 2018).

Figure 1 illustrates the proposed training procedure. Real images are processed by the generative model f_{RNVP} and by the discriminative model f_{cls}. The generative model outputs the likelihood which is subjected to generative NLL loss. The discriminative model outputs posterior probability which is subjected to cross-entropy with respect to the labels. At the same time, a random latent vector $\mathbf{z}$ is drawn from a Gaussian distribution. A synthetic outlier is obtained by sampling RNVP (f^{-1}_{RNVP}), and processed by f_{cls}. The resulting posterior is subjected to KL loss with respect to the uniform distribution. We summarize the joint training procedure in Algorithm 1.

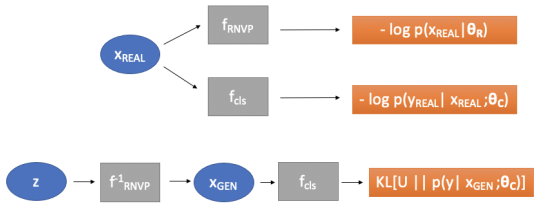


Figure 1: Schematic overview of the proposed training procedure using losses (2) and (3).

3.2 Dense Open-set Recognition

This section extends the described joint training of synthetic outliers towards dense open-set recognition. Different than in the image-wide case where an image is either an inlier or an outlier, dense OOD detection has to deal with images of mixed content. A real-world image may contain OOD objects or exclusively consist of inlier content. Thereby, outliers typically occlude the background, which results in well-defined

Algorithm 1: Joint training of an open-set classifier and an RNVP model for generation of synthetic outliers.

Require: $\lambda > 0$
Define RNVP: $\mathbf{z} = \mathbf{f}_{\theta_R}(\mathbf{x}), \mathbf{x} = \mathbf{f}^{-1}_{\theta_R}(\mathbf{z})$
Define Classifier: $P_{\theta_C}(y|\mathbf{x})$
Define Optimizers: $O_{\text{R}}(\theta_R), O_{\text{C}}(\theta_R)$
repeat
  $\mathbf{x}$, y = obtain_minibatch()
  $\mathbf{z}$ = sample $N(0,1)$
  $L_{\text{cls}} = -\log P_{\theta_C}(y|\mathbf{x}) + \lambda \text{KL}(U||P_{\theta_C}(y|\mathbf{f}^{-1}_{\theta_R}(\mathbf{z})))$
  $\theta_{\text{C}} += O_{\text{C}}.update(\nabla_{\theta_C} L_{\text{cls}})$
  $L_{\text{RNVP}} = -\log(p_z(\mathbf{f}_{\theta_R}(\mathbf{x}))) - \log\left(\left|det(\frac{\delta \mathbf{f}_{\theta_R}(\mathbf{x})}{\delta \mathbf{x}})\right|\right)$
  $\theta_{\text{R}} += O_{\text{R}}.update(\nabla_{\theta_R} L_{\text{RNVP}} + \nabla_{\theta_C} L_{\text{cls}})$
until *convergence*

borders between the outlier and the inlier content. We embed these observations into our dense open-set recognition approach as follows. We jointly train RNVP with a semantic segmentation model as we described in 3.1. Different than in the image-wide setup, we paste the synthetic outliers provided by RNVP at random locations within our regular training images. We train the dense prediction model to predict correct semantic segmentation in inlier pixels, and the uniform distribution within the patches generated by RNVP. The discriminative loss is defined by:

$$L_{\text{seg}}(\theta) = -\sum_i \sum_j [\![\mathbf{s}_{ij} = 0]\!] \log P_\theta(\mathbf{y}_{ij}|\mathbf{x}) + \lambda \sum_i \sum_j [\![\mathbf{s}_{ij} = 1]\!] \text{KL}(U \,||\, P_\theta(\mathbf{y}_{ij}|\mathbf{x})) \,. \quad (4)$$

In the above equation, $\mathbf{y}$ represents the ground truth labels, while $\mathbf{s}$ represents the OOD mask where zeros correspond to unchanged pixels of the training image (inliers) and ones denote the pasted RNVP output (outliers).

Figure 2 shows the proposed training setup for dense open-set recognition. We generate a synthetic outlier by sampling RNVP, and paste it at a random position in the training image. Images with mixed content are given to our discriminative model for dense prediction, which optimizes the discriminative loss (4). Consequently, the gradients of (4) are back-propagated to the RNVP. Additionally, we maximize the likelihood of the image patch replaced by the generated outlier. This is necessary in order to keep the samples generated by RNVP at the border of the training distribution (Lee et al., 2018). Due to the convenient architecture of RNVP we are able to generate synthetic samples that vary in spatial dimensions.

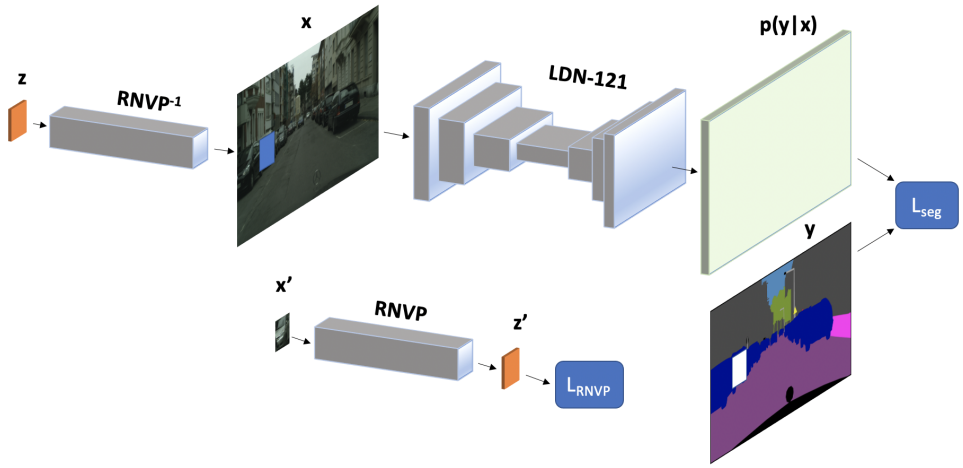


Figure 2: Schematic overview of the proposed training procedure for dense open-set recognition. The loss L_{seg} corresponds to (4), while the loss L_{RNVP} corresponds to (3).

Hence, we can train our model with synthetic outliers of different sizes. This is not straightforward to accomplish with other types of generative models and we consider that as a distinct advantage of our approach. The proposed procedure is summarized by Algorithm 2.

Semantic segmentation is known as a memory intensive task. Hence, we optimize memory consumption by using gradient checkpointing (Chen et al., 2016; Krešo et al., 2020) which trades computation time for lower memory consumption. We apply the checkpointing procedure both on our dense classifier and the RNVP.

Algorithm 2: Dense open-set recognition classifier training.

Require: $\lambda > 0$
Define RNVP: $\mathbf{z} = \mathbf{f}_{\theta_R}(\mathbf{x}), \mathbf{x} = \mathbf{f}_{\theta_R}^{-1}(\mathbf{z})$
Define Classifier: $P_{\theta_C}(\mathbf{y}|\mathbf{x})$
Define Optimizers: $O_{\mathrm{R}}(\theta_R), O_{\mathrm{C}}(\theta_R)$
repeat
  $\mathbf{x}, \mathbf{y} = \text{obtain\_minibatch}()$
  $\mathbf{z} = \text{sample} N(0, I)$
  $\mathbf{x}_{ood} = \mathbf{f}_{\theta_R}^{-1}(\mathbf{z})$
  $\mathbf{x}, \mathbf{x'}, \mathbf{y}, \mathbf{s} = \text{process\_batch}(\mathbf{x}, \mathbf{x}_{ood}, \mathbf{y})$
  $L_{\text{seg}}(\theta_C) =$
    $-\sum_i \sum_j [\![\mathbf{s}_{i,j} = 0]\!] \log P_{\theta_C}(\mathbf{y}_{i,j}|\mathbf{x})$
    $+\lambda \sum_i \sum_j [\![\mathbf{s}_{i,j} = 1]\!] \mathrm{KL}(U \,||\, P_{\theta_C}(\mathbf{y}_{i,j}|\mathbf{x}))$
  $\theta_{\mathrm{C}} += O_{\mathrm{C}}.update(\nabla_{\theta_C} L_{\text{seg}})$
  $L_{\text{RNVP}} = -\log(p_z(\mathbf{f}_{\theta_R}(\mathbf{x}')))$
    $-\log \left| det(\frac{\delta \mathbf{f}_{\theta_R}(\mathbf{x}')}{\delta \mathbf{x}'}) \right|$
  $\theta_{\mathrm{R}} += O_{\mathrm{R}}.update(\nabla_{\theta_R} L_{\text{RNVP}} + \nabla_{\theta_C} L_{\text{seg}})$
until *convergence*

3.3 Effects of Temperature Scaling onto OOD Detection

The loss (4) encourages high entropy of the softmax output in outlier pixels. This improves the outlier detection performance with respect to the standard cross-entropy loss. However, OOD accuracy can be further improved by applying temperature scaling during the inference phase (Guo et al., 2017; Liang et al., 2018). Dividing pre-softmax logits with a constant $T > 1$ moves the softmax output of every sample closer (but not equally closer) to the uniform distribution. We empirically show that such practice yields more appropriate values of the scoring function s and enables recognition of some previously undetected outliers (Liang et al., 2018).

We observe that dense classifiers tend to assign a low max-softmax score in pixels at semantic borders. Consequently, any of these pixels end up wrongly classified as outliers. This happens because the border pixels typically have two dominant logits (belonging to the two neighboring classes), while the other logits have significantly smaller values. On the other hand, undetected outlier pixels do not follow such pattern. It is easy to show that temperature scaling influences more the max-softmax score in pixels with homogeneous non-maximum logits than in pixels with two dominant logits. This practice improves separation of OOD score between border and outlier pixels, as well as the general OOD detection performance.

Table 1: OOD detection performance of the VGG13 model (Simonyan and Zisserman, 2015) trained on the CIFAR10 dataset. The RNVP-based approach outperforms both the max-softmax baseline (Hendrycks and Gimpel, 2017) and the GAN-based approach (Lee et al., 2018) across multiple OOD datasets and metrics. Our RNVP-based approach achieves 85.98% accuracy on the SVHN test set, while the GAN-based approach achieves 80.27%.

	TNR at TPR 95%	AUROC	OOD det. acc.
OOD dataset	Baseline / GAN outliers / RNVP outliers (ours)		
SVHN	14.0/12.7/**14.8**	46.2/46.2/**83.0**	66.9/65.9/**78.9**
LSUN (resize)	14.0/26.8/**46.5**	40.8/61.9/**87.5**	63.2/73.2/**79.3**
TinyImageNet (resize)	14.0/28.1/**33.7**	39.8/66.2/**79.7**	62.9/73.2/**73.3**

4 A NOVEL DENSE OOD DETECTION DATASET

We propose a novel OOD detection dataset which we obtained by relabeling the Mapillary Vistas (Neuhold et al., 2017) dataset. The original Vistas dataset consists of 18 000 training images and 2 000 validation images with 66 classes. We propose to use human classes as outliers due to their dispersion across scenes and visual diversity from other objects. We create a novel dense OOD dataset by excluding all images with class *person* and the three rider classes to the test subset. Consequently, our dataset has 8 003 train images and 830 validation images. The test set contains 11 167 images ($8\,003 + 830 + 11\,167 = 20\,000$). We refer to our dataset as Vistas-NP (no persons)[1].

The obtained dataset is similar to BDD-Anomaly (Hendrycks et al., 2019a) which selects the classes *motorcycle* and *train* classes as visual anomalies. However, the class *motorcycle* is often visually alike to class *bike*, while the class *train* is often visually similar to class *bus*. Therefore, the error of recognizing an OOD pixel on a motorcycle as a *bike* receives an equal penalty as the error of recognizing that pixel as a *person*. We believe that choosing persons as outliers is a more sensible choice since the whole category is removed from the dataset. Another advantage of Vistas dataset is better variety. All images from BDD dataset originate from the USA. Contrary, the Mapillary Vistas dataset contains a more extensive set of world-wide driving scenes.

Table 2 shows a comparison of the Vistas-NP test vs. FS Lost & Found (Blum et al., 2019) and BDD-Anomaly (Hendrycks et al., 2019a) test. FS Lost & Found contains 100 publicly available images while BDD-Anomaly test set includes 361 images. Our test subset has significantly more images with diverse instances of anomaly classes. Consequently, Vistas-NP is able to provide a more comprehensive insight into OOD detection performance.

[1]https://github.com/matejgrcic/Vistas-NP

Table 2: Comparison of Vistas-NP (ours) with respect to FS Lost & Found (Blum et al., 2019) and BDD-Anomaly (Hendrycks et al., 2019a). Note that FS Lost & Found recommends training on Cityscapes.

	Vistas-NP (ours)	FS L&F	BDD-A
	Label shares (%)		
Inlier	94.2	81.2	82.3
Outlier	0.6	0.2	0.8
Ignore	5.2	18.6	16.9
	Number of images		
Train	8 003	5 000	6 688
Test	11 167	100	361

5 EXPERIMENTS

We explore open-set recognition performance of the proposed RNVP-based approach. We first address the image-wide setup on CIFAR-10, where we compare the outlier detection performance of our RNVP-based approach with the original GAN-based approach (Lee et al., 2018) and the max-softmax baseline (Hendrycks and Gimpel, 2017). Subsequently, we evaluate an adaptation of our approach for dense prediction as proposed in 3.2. We demonstrate effectiveness on public open-set recognition datasets (Lost & Found, WD-Pascal, and StreetHazards) as well as on the proposed novel dataset Vistas-NP.

5.1 Image-wide OOD Detection

We evaluate our image-wide open-set recognition approach in experiments with the VGG-13 backbone (Simonyan and Zisserman, 2015) on CIFAR-10 (Krizhevsky et al., 2009). We evaluate OOD detection performance using multiple threshold-free metrics (Hendrycks and Gimpel, 2017) on outliers from LSUN (Yu et al., 2015) and Tiny-ImageNet. We use the maximum softmax probability (MSP) as a baseline.

We compare our RNVP-based method with the original formulation of this method, which is based

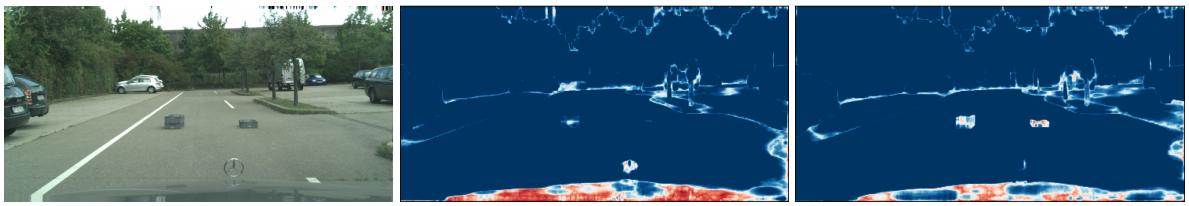
on adversarial generative training (Lee et al., 2018). We demonstrate advantage of RNVP-based setup in experiments with the same setup as proposed in (Lee et al., 2018). Table 1 shows the resulting OOD detection performance. Our RNVP-based approach outperforms other approaches across all metrics without losing the classification accuracy. We train the classifier for 100 epochs with batch size 64. In contrast to (Lee et al., 2018), we set the loss-modulation hyperparameter $\lambda = 1$ and do not validate it for a particular OOD dataset. The employed RNVP model consists of 3 residual blocks with 32 feature maps in every coupling layer. Downsampling is performed twice. RNVP's parameters are optimized with Adam optimizer. The training lasts for approximately 10 hours on a single NVIDIA Titan Xp GPU. Max memory allocation peaks at 1.3 GB with gradient checkpointing (Chen et al., 2016) of RNVP and 2.3 GB without checkpointing. Note that (Lee et al., 2018) also reports results for a different setup where outlier samples are sampled from external negative datasets. Those results are not relevant in the scope of this work.

5.2 Dense Open-set Recognition

We consider a dense open-set recognition approach which jointly trains a generative model of synthetic outliers as described in 3.2. We use a Ladder-style DenseNet-121 (Krešo et al., 2020) (LDN-121) on all datasets. LDN-121 is chosen due to its memory efficiency and prediction accuracy. Note that the proposed procedure is independent from the particular dense prediction model. We always train on random 512×512 crops which we sample from images resized to 512 pixels (shorter edge). We preserve the original label resolution for all datasets except Vistas-NP. Labels of the Vistas-NP dataset are resized to 512 pixels. The output of LDN-121 is bilinearly upsampled to the matching label resolution. During the training we use the batch size 6 (this would not be possible without checkpointing) and set λ to $1 * 10^{-3}$. The temperature scaling procedure uses $T = 2$ for softmax entropy and $T = 10$ for max-softmax OOD score. The value of λ is chosen in a way that it does not affect model's semantic segmentation performance on one held-out training image, while the parameter T is choosen so it gives the best OOD results on the held-out image. Parameters of LDN-121 are optimized using Adam optimizer with learning rate $1 * 10^{-4}$ for backbone parameters and $4 * 10^{-4}$ for upsampling path. For LDN-121 we decay learning rate throughout epochs using cosine annealing procedure to minimal value of $1 * 10^{-7}$. Additionally, LDN's backbone is initialized with ImageNet weights. Architecture of RNVP consists of 3 residual blocks with 32 feature maps in every coupling layer. Downsampling is performed three times. Parameters of RNVP are optimized using Adam with default hyperparameters. The generated outliers have spatial dimensions uniformly selected from the set $\{64, 72, 80, 88, 96, 104, 112, 120, 128, 136, 144\}$. Consequently, each outlier takes 1.5-8% of the image area.

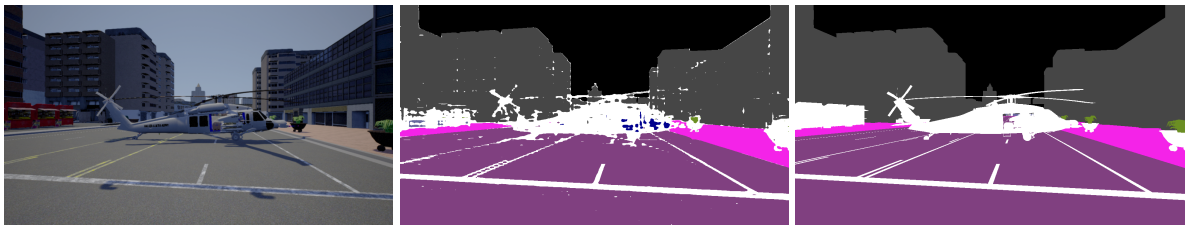
We demonstrate general applicability of the proposed method by training it on Cityscapes (Cordts et al., 2016) and testing dense outlier detection performance on Fishyscapes Lost and Found (Blum et al., 2019). We train all models for 54k iterations. We test the contribution of our jointly trained model by substituting synthetic outliers with Gaussian noise. We refer to this baseline as LDN + noise. Table 3 shows significant improvement of the proposed approach with respect to both baselines. The first two columns show the tested model and the achieved mIoU accuracy on the Cityscapes validation subset. The last two columns show OOD detection performance, where AP stands for average precision while F95 stands for TPR at FPR 95%. We show the performance of our models when outliers are detected using the max-softmax probability (MSP) and the entropy of softmax output (H).

Figure 3 shows qualitative results on FS Lost & Found. Ideally, OOD pixels should be painted in red which signals low-confidence predictions. The baseline fails to detect two boxes at the road as anomalies,



(a) (b) (c)

Figure 3: Model performance on FS Lost & Found dataset (Blum et al., 2019). Figure (a) shows the original image. Figure (b) shows the output of baseline model, while figure (c) shows the output of the model trained in the proposed setup. Our approach significantly improves the dense OOD detection performance.



(a) (b) (c)

Figure 4: Qualitative results on a StreetHazard test image in which the outlier object corresponds to a helicopter. We show the input image (a), dense open-set prediction by the proposed method (b), and the ground-truth labels (c). Outlier objects are marked in white. A pixel is marked as outlier if the corresponding max-softmax OOD score is higher than 80%.

Table 3: Dense open-set recognition on Fishyscapes Lost & Found (Blum et al., 2019) with LDN-121 (Krešo et al., 2020). Models are trained on Cityscapes dataset (Cordts et al., 2016). SO denotes synthetic outliers. AP stands for average precision, while F95 represents TPR at FPR 95%.

	Citysc.	FS L&F	
Model	mIoU	AP	F95
LDN, MSP (baseline)	**72.1**	3.9	30.8
LDN + noise	71.3	4.9	27.0
LDN+SO, T=1, MSP	71.5	6.8	26.8
LDN+SO, T=1, H	71.5	12.5	26.0
LDN+SO, T=10, MSP	71.5	**16.5**	**23.3**

while the proposed model performs much better.

Our training lasts for 94k iterations. We also test the LDN-121 using the proposed Vistas-NP dataset. As before, our OOD detection baseline is a discriminatively trained closed-set model activated with max-softmax. Table 4 shows the resulting dense open-set recognition performance. The first two columns correspond to the mIoU and the AP performance on the test split. The last column corespons to AP performance on WD-Pascal[2] (Bevandic et al., 2019). Note that WD-Pascal contains people marked as inliers. However, there are only few images with people so they do not affect average precision score significantly. The bottom section of the table illustrates advantages of softmax entropy and temperature scaling.

Finally, we evaluate the proposed method on the StreetHazards (Hendrycks et al., 2019a) dataset. The dataset contains 12 training classes. As proposed in (Hendrycks et al., 2019a), we calculate average precision for every image and report the mean value. The model is trained for 43k iterations. Table 5 presents the obtained results and compares it with the previous work. We achieve the best AUROC score, equal the best AP score, and outperform all previous approaches with respect to segmentation accuracy by a wide margin. Note that the best method (Franchi et al., 2020) uses ensemble learning.

Figure 4 shows results of LDN-121 trained in the proposed procedure. We mark pixels as outliers if the max-softmax score is higher than 80%. Parameter T is set to 10.

[2]https://github.com/pb-brainiac/semseg_od

Table 4: Dense open-set recognition with LDN-121 (Krešo et al., 2020) trained on the Vistas-NP dataset. Our model improves dense OOD detection on both Vistas-NP test set and WD-Pascal (Bevandic et al., 2019) without impairing the segmentation accuracy.

	Vistas-NP		WD-Psc.
Model	mIoU	AP	AP
LDN, MSP (baseline)	61.5	8.6	7.0
LDN+SO, T=1, MSP	61.6	9.3	14.1
LDN+SO, T=1, H	61.6	13.7	17.8
LDN+SO, T=10, MSP	61.6	16.2	20.5
LDN+SO, T=2, H	**61.6**	**16.9**	**21.5**

The proposed procedure consumes different amounts of GPU memory depending on the spatial dimensions of generated outliers. We asses the memory consumption using NVIDIA Titan Xp and batch size 4. We measure memory allocation of 9.61 GB for maximal outlier size, while the minimal outlier size consumes 7.88 GB of memory. When we apply the gradient checkpointing, memory allocation peaks at 5.55 GB, while the minimal allocation equals to 4.92 GB. Information about the GPU memory allocation is obtained using $torch.cuda.max\_memory\_allocated()$.

6 CONCLUSION

We have presented a novel dense open-set recognition approach based on discriminative training with jointly trained synthetic outliers. The synthetic outliers are obtained by sampling a generative model based on normalized flow that is trained alongside a dense discriminative model in order to produce samples at the border of the training distribution. We paste the generated samples into densely annotated training images, and learn dense open-set recognition models which perform simultaneous semantic segmentation and dense outlier detection. Experiments

Table 5: Results on StreetHazards (Hendrycks et al., 2019a) dataset. Our method equals the best AP score, achieves the best AUROC score and the third best FPR95. Additionally, we achieve the best mIoU accuracy.

Model	mIoU ↑	AP ↑	FPR95 ↓	AUROC ↑
LDN, MSP (Hendrycks and Gimpel, 2017)	56.2	7.3	30.8	89.0
Dropout (Gal and Ghahramani, 2016)(Xia et al., 2020)	/	7.5	79.4	69.9
AE (Baur et al., 2018)(Xia et al., 2020)	/	2.2	91.7	66.1
MSP + CRF (Hendrycks et al., 2019a)	/	6.5	29.9	88.1
SynthCP, t=1 (Xia et al., 2020)	/	8.1	46.0	81.9
SynthCP, t=0.999 (Xia et al., 2020)	/	9.3	28.4	88.5
Ensemble OVA (Franchi et al., 2020)	54.0	**12.7**	**21.9**	91.6
OVNNI (Franchi et al., 2020)	54.6	12.6	22.2	91.2
LDN + SO, T=1, MSP	**59.7**	8.6	26.1	90.2
LDN + SO, T=10, MSP	**59.7**	12.1	29.1	90.8
LDN + SO, T=1, H	**59.7**	11.3	25.7	91.1
LDN + SO, T=2, H	**59.7**	**12.7**	25.2	**91.7**

on CIFAR-10 show that synthetic outliers generated by RNVP lead to better open-set performance then their GAN counterparts. We present dense open-set recognition experiments on a novel dataset which we call Vistas-NP, as well as on three public datasets which were proposed in the prior work. The proposed approach is competitive with respect to the state of the art on the StreetHazards dataset. Additionally, we outperform baselines on two other dense OOD detection datasets. Suitable avenues for future work include increasing the capacity of the generative model, combining the proposed approach with noisy outliers from some large general-purpose dataset, and devising more involved approaches for simultaneous discriminative and generative modeling.

ACKNOWLEDGMENTS

We thank Ivan Grubišić, Marin Oršić and Jakob Verbeek on their useful comments. This work has been supported by the European Regional Development Fund under the project "A-UNIT - Research and development of an advanced unit for autonomous control of mobile vehicles in logistics" (KK.01.2.1.02.0119).

REFERENCES

Andrews, J. T. A., Tanay, T., Morton, E., and Griffin, L. D. (2016). Transfer representation-learning for anomaly detection. In *ICML 2016*.

Baur, C., Wiestler, B., Albarqouni, S., and Navab, N. (2018). Deep autoencoding models for unsupervised anomaly segmentation in brain MR images. In Crimi, A., Bakas, S., Kuijf, H. J., Keyvan, F., Reyes, M., and van Walsum, T., editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I*, volume 11383 of *Lecture Notes in Computer Science*, pages 161–169. Springer.

Bendale, A. and Boult, T. E. (2015). Towards open world recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1893–1902. IEEE Computer Society.

Bevandic, P., Kreso, I., Orsic, M., and Segvic, S. (2019). Simultaneous semantic segmentation and outlier detection in presence of domain shift. In Fink, G. A., Frintrop, S., and Jiang, X., editors, *Pattern Recognition - 41st DAGM German Conference, DAGM GCPR 2019, Dortmund, Germany, September 10-13, 2019, Proceedings*, volume 11824 of *Lecture Notes in Computer Science*, pages 33–47. Springer.

Blum, H., Sarlin, P., Nieto, J. I., Siegwart, R., and Cadena, C. (2019). The fishyscapes benchmark: Measuring blind spots in semantic segmentation. *CoRR*, abs/1904.03215.

Chen, T., Xu, B., Zhang, C., and Guestrin, C. (2016). Training deep nets with sublinear memory cost. *CoRR*, abs/1604.06174.

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 *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 3213–3223. IEEE Computer Society.

DeVries, T. and Taylor, G. W. (2018). Learning confidence for out-of-distribution detection in neural networks. *CoRR*, abs/1802.04865.

Dinh, L., Krueger, D., and Bengio, Y. (2015). NICE: nonlinear independent components estimation. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. (2017). Density estimation using real NVP. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Everingham, M., Gool, L. V., Williams, C. K. I., Winn, J. M., and Zisserman, A. (2010). The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.*, 88(2):303–338.

Franchi, G., Bursuc, A., Aldea, E., Dubuisson, S., and Bloch, I. (2020). One versus all for deep neural network incertitude (OVNNI) quantification. *CoRR*, abs/2006.00954.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Balcan, M. and Weinberger, K. Q., editors, *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1050–1059. JMLR.org.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680.

Grathwohl, W., Wang, K., Jacobsen, J., Duvenaud, D., Norouzi, M., and Swersky, K. (2020). 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*.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Identity mappings in deep residual networks. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, volume 9908 of *Lecture Notes in Computer Science*, pages 630–645. Springer.

Hendrycks, D., Basart, S., Mazeika, M., Mostajabi, M., Steinhardt, J., and Song, D. (2019a). A benchmark for anomaly segmentation. *CoRR*, abs/1911.11132.

Hendrycks, D. and Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Hendrycks, D., Mazeika, M., and Dietterich, T. G. (2019b). Deep anomaly detection with outlier exposure. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2261–2269. IEEE Computer Society.

Kendall, A. and Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5574–5584.

Kirichenko, P., Izmailov, P., and Wilson, A. G. (2020). Why normalizing flows fail to detect out-of-distribution data. *CoRR*, abs/2006.08545.

Krešo, I., Krapac, J., and Šegvić, S. (2020). Efficient ladder-style densenets for semantic segmentation of large images. *IEEE Transactions on Intelligent Transportation Systems*.

Krizhevsky, A., Nair, V., and Hinton, G. (2009). Cifar-10 (canadian institute for advanced research).

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Bartlett, P. L., Pereira, F. C. N., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 1106–1114.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6402–6413.

Lambert, J., Zhuang, L., Sener, O., Hays, J., and Koltun, V. (2020). MSeg: A composite dataset for multi-domain semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR)*.

Lee, K., Lee, H., Lee, K., and Shin, J. (2018). Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Liang, S., Li, Y., and Srikant, R. (2018). Enhancing the reliability of out-of-distribution image detection in neural networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 60(2):91–110.

Lucas, T., Shmelkov, K., Alahari, K., Schmid, C., and Verbeek, J. (2019). Adaptive density estimation for generative models. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 11993–12003.

Malinin, A. and Gales, M. J. F. (2018). Predictive uncertainty estimation via prior networks. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 7047–7058.

Nalisnick, E. T., Matsukawa, A., Teh, Y. W., Görür, D., and Lakshminarayanan, B. (2019). Do deep generative models know what they don't know? In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Neuhold, G., Ollmann, T., Bulò, S. R., and Kontschieder, P. (2017). The mapillary vistas dataset for semantic understanding of street scenes. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5000–5009. IEEE Computer Society.

Nitsch, J., Itkina, M., Senanayake, R., Nieto, J., Schmidt, M., Siegwart, R., Kochenderfer, M. J., and Cadena, C. (2020). Out-of-distribution detection for automotive perception. *arXiv preprint arXiv:2011.01413*.

Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., DePristo, M. A., Dillon, J. V., and Lakshminarayanan, B. (2019). Likelihood ratios for out-of-distribution detection. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 14680–14691.

Rosten, E. and Drummond, T. (2006). Machine learning for high-speed corner detection. In Leonardis, A., Bischof, H., and Pinz, A., editors, *Computer Vision - ECCV 2006, 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part I*, volume 3951 of *Lecture Notes in Computer Science*, pages 430–443. Springer.

Salakhutdinov, R., Mnih, A., and Hinton, G. E. (2007). Restricted boltzmann machines for collaborative filtering. In Ghahramani, Z., editor, *Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, volume 227 of *ACM International Conference Proceeding Series*, pages 791–798. ACM.

Sánchez, J., Perronnin, F., Mensink, T., and Verbeek, J. J. (2013). Image classification with the fisher vector: Theory and practice. *Int. J. Comput. Vis.*, 105(3):222–245.

Serrà, J., Álvarez, D., Gómez, V., Slizovskaia, O., Núñez, J. F., and Luque, J. (2020). Input complexity and out-of-distribution detection with likelihood-based generative models. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Xia, Y., Zhang, Y., Liu, F., Shen, W., and Yuille, A. L. (2020). Synthesize then compare: Detecting failures and anomalies for semantic segmentation. *CoRR*, abs/2003.08440.

Yu, F., Xian, W., Chen, Y., Liu, F., Liao, M., Madhavan, V., and Darrell, T. (2018). BDD100K: A diverse driving video database with scalable annotation tooling. *CoRR*, abs/1805.04687.

Yu, F., Zhang, Y., Song, S., Seff, A., and Xiao, J. (2015). LSUN: construction of a large-scale image dataset using deep learning with humans in the loop. *CoRR*, abs/1506.03365.

Zendel, O., Honauer, K., Murschitz, M., Steininger, D., and Domínguez, G. F. (2018). Wilddash - creating hazard-aware benchmarks. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VI*, volume 11210 of *Lecture Notes in Computer Science*, pages 407–421. Springer.