Far Away in the Deep Space: Nearest-Neighbor-Based Dense Out-of-Distribution Detection

Silvio Galessso
University of Freiburg
galessos@cs.uni-freiburg.de

Max Argus
University of Freiburg

Thomas Brox
University of Freiburg

Abstract

The key to out-of-distribution detection is density estimation of the in-distribution data or of its feature representations. While good parametric solutions to this problem exist for well curated classification data, these are less suitable for complex domains, such as semantic segmentation. In this paper, we show that a k-Nearest-Neighbors approach can achieve surprisingly good results with small reference datasets and runtimes, and be robust with respect to hyperparameters, such as the number of neighbors and the choice of the support set size. Moreover, we show that it combines well with anomaly scores from standard parametric approaches, and we find that transformer features are particularly well suited to detect novel objects in combination with k-Nearest-Neighbors. Ultimately, the approach is simple and non-invasive, i.e., it does not affect the primary segmentation performance, avoids training on examples of anomalies, and achieves state-of-the-art results on the common benchmarks with +23% and +16% AP improvements on RoadAnomaly and StreetHazards respectively.

1. Introduction

Deep learning models can achieve remarkably good performance on a large number of tasks. However, when these models are evaluated on data different from the training distribution, their performance usually deteriorates substantially. Even worse, the models often do not even realize that they ran out of distribution and make wrong predictions with high confidence. For the safe deployment of machine learning systems in the open world, where there is no control over the distribution of the input data, the ability of a model to detect out-of-distribution (OoD) samples becomes important. Since such a system should be able to identify all unforeseen deviations from the training data, it cannot learn the distribution of the novel samples, but must base its decision on a model of the inlier distribution. This makes novelty detection a particularly challenging task, with an increasing number of research contributions in the last few years.

In this work, we focus on segmentation models for driving scenes, where we aim for detection and segmentation of unseen patterns in the image. This requires spatially resolved outputs, rather than an accumulated decision for the whole image. Moreover, driving scenes comprise very diverse patterns, which makes modelling this complex inlier distribution far from straightforward. The problem has been approached by several works already [14, 19, 29, 31], and a number of accepted evaluation benchmarks exist [4, 6, 17], on which we also assess our method.

Inspired by the success of non-parametric nearest neighbor methods in the scope of industrial anomaly detection [8, 24, 25], we investigate if and how this method can be applied in the much more diverse autonomous driving scenarios. The situation in these two domains is very different. In driving scenarios, we have a good set of annotated in-domain data and a working segmentation model for the in-distribution data, whereas in industrial anomaly de-
tection, pixel-level annotations are scarce even for the in-distribution data. On the other hand, objects in industrial inspection are often observed under very similar pose and lighting conditions with the object of interest in the center and controlled background, whereas the variation in driving scenes is much larger. Due to these differences, a nearest-neighbor approach has neither been shown nor publicly considered for driving scenarios.

We demonstrate that this has been an important omission as such approaches can also substantially improve performance in this domain. However, due to the larger diversity and the more semantic nature of the task the choice of the feature encoder, which feature granularity to use, and how to perform efficient sub-sampling on many million patterns is very important. We find a good and stable configuration, that outperforms all previous OoD segmentation methods on the StreetHazards, RoadAnomaly, and SegmentMelfYouCan benchmarks, often by a large margin and without relying on out-of-domain data during training. Moreover, while we use the features from the underlying semantic segmentation model, the output of that model is not affected. This enables an efficient integration of the novelty segmentation approach into existing semantic segmentation systems.

2. Related Work

2.1. Out-of-Distribution (OoD) Detection

In computer vision, detecting anomalous patterns is a task with several applications. One is visual inspection in industrial manufacturing, with the goal of identifying production defects [1, 39]. In this scenario, many examples of healthy items are easily available, whereas the distribution of the possible faults is unknown. The aim is to identify the presence and possibly the location of the defect. The application to a single object class at a time, together with the limited variation in appearance of said objects, make this sub-domain of OoD detection a somewhat specialized niche in terms of challenges and methods.

A more research oriented evaluation setting for out-of-distribution detection is that of image recognition [18, 34, 36], where the normal data distribution is subdivided in multiple semantic categories, which can in turn have nuanced appearances and subtypes. The goal in this case is to reliably identify images that do not belong to the semantic classes of the training set.

The out-of-distribution image recognition problem can be further extended to semantic segmentation [4, 6, 17], which is the task we address in this work, and where we wish to identify the individual pixels that correspond to unknown entities. As in image recognition, the normal data distribution consists of several semantic categories. An additional challenge of this setting is its dense (per-pixel) nature: each location in the image is at the same time part of a semantically coherent object, and in relation with the surrounding environment. This ensures high diversity of the data samples, as well as a high degree of correlation between them.

OoD Detection with Deep Feature Representations

The outputs produced by a learned discriminative model, such as a neural network optimized for classification or segmentation, can also be used for uncertainty estimation and anomaly detection “out-of-the-box”. Anomaly scores like the predictive entropy or the the model’s top-logits constitute ubiquitous baselines, and represent viable approaches especially for dense anomaly detection, where computational scalability is a concern [12, 17, 19].

In the context of image classification and industrial inspection, however, it is more common to leverage the deep feature representations produced by the networks. In the former case, successful approaches are based on fitting parametric distributions to the normal features, and using the resulting (negative) likelihood of the test samples as an anomaly score [20, 37]. For industrial inspection, parametric density estimation has also been used [26], but recent literature has seen the use of non-parametric methods based on explicit feature matching and k nearest neighbors [8, 25].

Nearest neighbors are a valid choice to represent densities which cannot be effectively modeled parametrically, however it can easily incur in unsustainable computational costs due to the need to store sufficient normal features. PatchCore [25] circumvents this issue by computing a core-set of features best suited for the task, thereby obtaining an advantageous cost-performance trade-off.

2.2. Dense Out-of-Distribution Detection

Despite having fundamentally the same goal, industrial anomaly detection and dense out-of-distribution detection for autonomous driving are essentially different problems, due to their distinct data domains. This is confirmed by the fact that there is virtually no overlap between the respective state-of-the-art approaches. Most techniques for dense OoD detection use the output of a pre-trained semantic segmentation model, from exploiting the logit values [12, 17, 19], to the prediction entropy [7], to interpreting it as an energy/density/distance function [5, 14, 29].

A recent trend is to make use of an external source of out-of-distribution data, typically from a semantically disjoint dataset, to influence or regularize the optimization of the model [7, 14, 29]. Perhaps unsurprisingly, approaches that use of third-party anomalies tend to dominate the benchmarks.

One further class of methods use the proxy task of image reconstruction, conditioned on the predicted segmentation mask [21, 31]. These rely on the assumption that the con-
ditional generator will not be able to correctly reproduce out-of-domain objects.

To the best of our knowledge, we are the first to successfully apply non-parametric, nearest-neighbor based anomaly detection to the semantic segmentation setting, showing that an approach seemingly confined to industrial inspection can in fact excel in this domain.

3. Deep Neighbor Proximity for OoD Detection

At the core of our approach for out-of-distribution detection are deep k-Nearest-Neighbors. As illustrated in Figure 1, our method relies on the computation of distances between feature representations produced by the encoder of a semantic segmentation network. At test time we collect the distances between feature representations produced by the encoder of the in-distribution dataset – i.e. the training set for the segmentation network.

Consider a matrix of in-distribution reference features as $R \in \mathbb{R}^{N \times C}$, where $N$ is the number of reference features and $C$ is the dimensionality of each feature vector. The computation of $R$ will be described in Sections 3.2 and 3.3. For a test image, we extract the feature representation $T \in \mathbb{R}^{H \times W \times C}$, and “flatten” it in the spatial dimensions $H, W$. We first compute the matrix $D \in \mathbb{R}^{H \times W \times N}$ of distances between each possible combination of samples $t$ and $r$ in the feature sets:

$$d_{i,j} = \text{dist}(t_j, r_i) \quad \forall \ i \in \{1..H\cdot W\}, \ j \in \{1..N\}$$

(1)

where $\text{dist}$ is the euclidean distance. Then, for each test feature $i$ we compute the OoD score as the average of the distances to the closest $k$ neighbors:

$$s_j^i = \frac{1}{k} \cdot \min_{D_i \subseteq D_i} \sum_{d \in D_i} d_i$$

(2)

and successively reshape them into the original feature shape, to obtain: $S^N \in \mathbb{R}^{H \times W}$.

3.1. Out-of-Distribution Scores

The distances obtained with the procedure described above can be directly used as anomaly scores, however they can also be combined with those obtained from the model predictions. For this we choose the $\text{max logit}$ approach, which uses the negative values of the top-class logits produced by the segmentation model, and has been shown to outperform comparable alternatives [12, 17].

In order to combine $S^N$ with the max-logit scores $S^L$, we first bring them to the same resolution, by upsampling $S^N$ to the original image size. Subsequently, we simply scale both to the same range using their respective extrema estimated on the training set:

$$S^N = S^N / \max S^N$$

(3)

$$S^L = (S^L - \min S^L_{\text{train}}) / (\max S^L_{\text{train}} - \min S^L_{\text{train}})$$

(4)

and finally compute the combined scores:

$$S^C = S^N + S^L$$

(5)

In the following text, we will use the abbreviations ML (Max-Logit) for methods based only on $S^L$, whereas DNP (Deep Neighbor Proximity) and cDNP (combined Deep Neighbor Proximity) refers to those based on $S^N$ and $S^C$ respectively.

3.2. Model Architectures and Feature Extraction

To assess the versatility of our method, we apply it to four different feature extraction architectures: ResNet [16], ConvNeXt [22], MiT [33] and ViT [11].

ResNet and ConvNeXt are both CNNs consisting of a cascade of $4$ computational stages. We can extract convolutional features at the end of each stage. Earlier stage features have higher resolution but less semantic content. Both are designed such that the 3rd stage contains more internal layers than the others stages.

MiT is also a hierarchical 4-stage architecture, but it uses alternating multi-head self-attention blocks and convolutional layers. Here we can test the representations from the output of each stage, as above, but also the internal features of the self-attention mechanism: queries, keys, and values.

ViT is a “pure” transformer, as it is entirely composed of self-attention blocks that output a constant number of patch features, corresponding to a constant resolution. For this architecture we test the features taken from the output of different transformer blocks, as well as the queries, keys, and values produced by the attention mechanism.

In order to learn the feature embeddings we embed them as backbones in compatible encoder-decoder segmentation models, and follow their respective standard training procedures. For ResNet and ConvNeXt we use UPerNet [32], for MiT we use SegFormer [33], and for ViT we use Segmenter [28] and SETR [38]. These choices follow established practices in semantic segmentation literature, so that they would be viable options in practical applications.

Feature Selection: For each encoder architecture there are many choices of features to extract for computing neighbor distances, from different depth levels to different functional layers. We evaluate the ones described above for each model type, so as to identify the most suitable for the task, and present the results in Section 4.4.

3.3. Reference Feature Subsampling

In order to have a tractable amount of reference features we sub-sample the representations obtained from the
training set. For this stage we evaluate three options: random sub-sampling, greedy coreset reduction (GCS), and per-class greedy coreset reduction (PC-GCS). The second method has been used by PatchCore [25], and consists in a selection procedure that aims to preserve the best coverage of the training features in the representation space, according to a nearest-neighbor distance criterion.

PC-GCS is a proposed variant of GCS applied separately to each category present in the segmentation dataset. PC-GCS makes sense in this setting because industrial inspection images, on which PatchCore is originally applied, are single-class and less diverse in appearance than the segmentation data we use, and therefore an application to coherent sub-components, such as semantic categories, is closer to its original intended scenario. PC-GCS also preserves the balance between classes of the original dataset.

4. Experiments

In this section we present results that inform about the choice of features (4.4), subsampling strategy (4.5), and number of neighbors (4.6). In Section 4.8, we show how the nearest-neighbor-based approaches we introduced in Section 3.1 perform in comparison with the max-logit baseline, and finally how they compare with current state-of-the-art approaches in Section 4.9. However, we first begin with an introduction of the experimental setup.

4.1. Datasets and Benchmarks

Different benchmarks for the evaluation of dense OoD performance exist, and while they all revolve on semantic segmentation data for autonomous driving, they are quite different in nature.

StreetHazards [17] is a synthetic dataset and benchmark, featuring 12 in-distribution categories in the annotated training/validation sets, and 250 diverse OoD objects, annotated as 1 category in the test set. Its size (1500 test images), variety of OoD objects and locations makes it a valid benchmark for research.

RoadAnomaly [21] is a benchmark made of a set of images downloaded from the web, which features objects, such as animals or vehicles, from categories alien to the typical driving ontology e.g. Cityscapes [9] or BDD100k [35].

SegmentMeIfYouCan - Anomaly Track [6] is an extension of RoadAnomaly, containing mostly images for which the OoD ground truth is undisclosed.

Except for StreetHazards, which comes with a training set, we use the other benchmarks with Cityscapes trained models. For our ablation experiments we use RoadAnomaly and StreetHazards.

4.2. Metrics

The most important metric for the task at hand is the Average Precision (AP), which is a holistic metric, averaged over several threshold values. Secondly, we report the False Positive Rate at 95% True Positive Rate (FPR95), which measures the performance at a high detection threshold - relevant for safety critical applications.

4.3. Training the Feature Extractors

We follow the standard recommended training procedure for each of our semantic segmentation models, optimizing for the cross entropy objective using exclusively the respective training dataset. We rely on the common mmsegmentation [23] framework, and adhere to the default training and optimization settings reported in its configuration files. We train all models for 80k iterations with polynomial learning rate schedule, using SGD with learning rate $1e^{-2}$ for UPerNet-ResNet and Segmenter-ViT, and AdamW with learning rate $1e^{-5}$ for the other backbones.

Please note that we select the network snapshots based on segmentation performance – i.e. after full convergence – even though this might negatively impact OoD detection results. In fact, OoD detection performance is observed to vary greatly over the epochs, and tends to reach its peak before full convergence of the segmentation loss, before declining again due to overfitting/overconfidence [15].

All encoders are initialized with the respective publicly available ImageNet [10] pre-trained parameters. For fair comparison, on our ViT models we use the DeiT [30] weights, instead of the original ones trained on a larger undisclosed dataset.

![Figure 2. DNP performance (AP on RoadAnomaly) using features from the last 4 layers/stages of ViT-B. Results for Queries, Keys, Values, and features from the end of each stage. The last 2 layers produce the most useful features, and the self-attention features perform equally well and always better than the final representations.]

4.4. Choosing the Best Features

In this section we explore the efficacy of our approach with different feature representations, as anticipated in Section 3.2. For both CNN architectures, we observe the best suited representations are those extracted at the third stage (out of four). These features likely strike the right balance between resolution and semantic abstraction. See Appendix for more complete results.
More interesting cases are those of the transformer encoders, MiT and ViT, which include self-attention layers, from which query, key and value representations can be extracted. A comparison between the last 4 layers of ViT-B is shown in Figure 2, including the aforementioned self-attention features, as well as standard end-of-the-block representations. Results indicate a clear superiority of the self-attention features, which perform approximately equally well. A similar analysis on MiT confirms the superiority of queries, keys, and values, which achieve ~59% better AP than standard end-of-the-block features.

In the following sections, we will use the keys as default features for transformer backbones, in accordance with other works [27], but we note that all three perform equally, and leave a more informed choice for future research.

4.5. Reference Feature Subsampling Methods

Here we discuss the results on the choice of the subsampling method: random, GCS, PC-CGS, which are summarized in the plots in Figure 3 for the architectures of Table 1. We evaluate for different values of \( N \), from 5\( k \) to 100\( k \), for each setup we run 3 test runs with different seeds affecting the sub-sampling randomness. We can first of all observe that ranking between the methods is not consistent across architectures.

For ResNet, GCS dominates on average, but with high standard deviation. For ConvNeXt, random is the best with fewer reference features but is outperformed by PC-GCS otherwise. For MiT, PC-CGS lags behind with few reference features, but catches up eventually. For ViT the three methods perform similarly, with GCS having a higher variance again. While these results do not indicate a clear winner, they also tell that the simplest sub-sampling approach – random – is competitive.

4.6. Impact of the Number of Neighbors

In this section we present the results of an ablation on the number of neighbors (\( k \)) considered in the computation of the \( S^N \) scores, i.e., how many elements the average in Equation 2 contains. Figure 4 shows the effect of \( k \) on our approach for the RoadAnomaly dataset and with ConvNeXt and Segmenter features. The best AP is obtained with \( k=1 \) and there is a monotonic decrease as \( k \) increases. The drop in performance, however, is slow and with low standard deviation, which is useful for reliably picking \( k \).

Although not being the optimal value in this study, we choose \( k=5 \) for all other experiments to have a more robust expected distance computation (Equation 2).

4.7. Computational Costs

A major point of concern with k-nearest-neighbors is the computational cost due to the calculation of the distances between all feature pairs. This depends on two major factors: the number of reference features \( N \), and the network architecture, which determines the test feature resolution \( H \times W \) and channel size \( C \). We estimate the runtime increase of our approach on an NVIDIA RTX 2080Ti, taking into account the asynchronous nature of computations on GPU, which is the device we use for computing distances.

In Figure 5 we show the average runtime increase of cDNP over the baseline max-logit, for 1280×720 resolu-
In terms of pure runtime increase, ResNet50 is the most expensive architecture with ~120ms for \( N=100k \) – due to its comparatively high feature size and resolution, while MiT, which has the lowest feature resolution, is the least expensive with ~29ms for \( N=100k \). Looking at the AP gains, however, MiT and DeiT offer the best trade-offs, with the latter delivering by far the best improvements even at \( N=10k \), at the small cost of ~13ms.

The only network that has no AP gain with less than 100k reference features is ConvNeXt-T, but has improvements in terms of FPR\(_{95}\) nonetheless.

### 4.8. Comparing Max-Logit, DNP, and cDNP

In this paragraph we compare the three out-of-distribution scores defined in Section 3.1, i.e. the baseline max-logit (ML) and the nearest-neighbor based approaches: DNP and cDNP. We compare results for four feature extractors, on RoadAnomaly and StreetHazards: numerical results are shown in Table 1, while summary plots are provided in Figure 6.

Consistently for all benchmarks, architectures, and metrics, cDNP performs better than its two component scores ML and DNP. However, the extent of its superiority changes for different architectures and is much greater for the attention-based encoders, particularly for ViT. This can also easily be seen in Figure 6.

In fact, while ML is occasionally better than DNP for ResNet and ConvNeXt, for the last two models the performance of DNP – the pure kNN method – is already superior to that of ML, and performs almost as well as cDNP in the case of ViT. The performance gains of cDNP are particularly noticeable in terms of FPR\(_{95}\), where they are substantial for all architectures.

### 4.9. Comparison with the State of the Art

Here we compare our approach with the state-of-the-art in dense out-of-distribution detection. Results for RoadAnomaly and StreetHazards can be seen Table 2, where we include cDNP applied to a few different architectures: ConvNeXt-S (a slightly larger version of the best CNN from Section 4.8), Segmenter-B, and SETR (based on ViT-L). We added the last model to show that our approach can work with different encoder sizes and segmentation heads, and we used an official snapshot, to test the method with high accuracy off-the-shelf parameters.

Included in the table are methods that use out-of-domain data at training time (OE), most notably PEBAL [29] and
DenseHybrid [14]. While these approaches work well on the benchmarks at hand, they break the interpretation of out-of-distribution data as completely unknown.

Despite not taking advantage of negative training data, our approach is by far the best performing on RoadAnomaly, on both AP and FPR$_{95}$. Segmenter and SETR perform comparably well, followed by ConvNeXt and PE-BAL. On StreetHazards Segmenter has the best AP, followed by ConvNeXt – which has the lowest FPR$_{95}$ – and DenseHybrid.

In Table 3 we report results for the SegmentMeIfYou-Can (SMIYC) Anomaly benchmark, whose ground truth is undisclosed, including our results with Segmenter-ViT-B. Our approach is the best in terms of AP followed by Entropy-Maximization and is second best in terms of FPR$_{95}$ after DenseHybrid, but outperforming Entropy-Maximization. Our method outperforms all others not relying on outlier exposure according to both metrics.

![Figure 6. Comparison between max-logit (ML), DNP, and cDNP, with all architectures, on RoadAnomaly and StreetHazards. The metrics shown are AP (higher is better) and FPR$_{95}$ (lower is better). In all four plots cDNP outperforms the ML and DNP sub-components. For the transformer encoders, DNP alone obtains remarkably good results. Note that the architectures on the x axis are not formally continuous.](image)

Table 3. State of the art results on the SMIYC-Anomaly benchmark (test set, official leaderboard). OE denotes the use of negative/anomalous training data. Best results are bold, second best are underlined.

| Method          | OE | AP $\uparrow$ | FPR$_{95}$ $\downarrow$ |
|-----------------|----|---------------|--------------------------|
| PEBAL [29]      | ✓  | 49.14         | 40.82                    |
| NFlowJS [13]    |    | 56.92         | 34.71                    |
| ObsNet [2]      |    | 75.44         | 26.69                    |
| DenseHybrid [14]| ✓  | 77.96         | 9.81                     |
| Max-Entropy [7] | ✓  | 85.47         | 15.00                    |
| cDNP-Segmenter-B (Ours) |   | **88.90**     | **11.42**                |

5. Conclusion

In this work we presented combined Deep Neighbor Proximity (cDNP), an approach for dense out-of-distribution detection based on k nearest neighbors, which is simple, has low computational cost and achieves state-of-the-art performance on common driving-focused anomaly detection benchmarks. Our method can be easily combined with standard parametric ones for a performance boost, but also delivers an exceptional standalone performance when paired with attention-based models, most notably with ViT.

We did an extensive comparative study to find the right settings for our approach and to verify that it is practical and robust to hyperparameter changes. Our findings go hand-in-hand with recently discovered properties of transformers’ representations, and open the way to their use in a real-world, safety-critical setting.

Our work uncovers and exploits said properties of transformer features, but stops short of explaining their root cause. Other than that, the main current limitations of our method are the added computational requirements, and the resolution of the kNN based scores, which is lower than that of the original image and depends on the encoder architecture. For the former issue, possible solutions include better feature sub-sampling or aggregation strategies, as well as optimized nearest neighbor computation. The latter could be solved with a guided upsampling strategy.

Moreover, while we presented the approach in its simplicity, a number of increments and improvements are possible, such as the integration of other types of anomaly scores (energy based, standardized max logits), or the inclusion of outlier data in the training pipeline, which has been an important component of prior state-of-the-art methods.

4.10. Qualitative Results

Figures 7 and 8 show qualitative examples of our method on RoadAnomaly and StreetHazards respectively, using Segmenter-ViT-B. In both cases, the superiority of the k-nearest-neighbor based scores compared to max-logit can be seen. Moreover, the fruitful combination effect between ML and DNP can be seen in the background in-distribution elements, such as buildings and roads, for which the cDNP scores are lower than those of DNP.

In Figure 9 we show more examples from both datasets, including results from both ViT-B and UPerNet-ConvNeXt-S. While the attention based model generally performs better than the CNN, the example in the third row is an exception, as the scores obtained from ViT contain a large number of false positives.
Figure 7. Qualitative results for Segmenter-B on RoadAnomaly. It can be observed how DNP scores are better markers for anomalous entities than ML, and that their combination cDNP has lower OoD scores for background classes such as roads and buildings.

Figure 8. Qualitative results for Segmenter-B on StreetHazards. In the ground truth mask OoD objects are indicated in orange. Here DNP and cDNP perform similarly, both better than ML. In the first and third row, DNP and cDNP have fewer false positives, especially for objects far from the camera. In the second row, ML fails to detect large parts of the unknown object, which is correctly identified by our approach: in this case DNP performs slightly better than cDNP.

Figure 9. Qualitative results with ConvNeXt-S and Segmenter-ViT-B. The top two examples are from RoadAnomaly, followed by two StreetHazards ones. The score maps show how the combination of ML and DNP is an improvement over both, mostly through the removal/filtering of false positives. In the first and third row, ViT is clearly outperforming ConvNeXt, whereas in the last row it is the other way around: cDNP-ConvNeXt successfully isolates the unknown entity, while ViT cDNP produces many false positives.
References

[1] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad — a comprehensive real-world dataset for unsupervised anomaly detection. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 9584–9592, 2019. 2

[2] Victor Besnier, Andrei Bursuc, David Picard, and Alexandre Briot. Triggering failures: Out-of-distribution detection by learning from local adversarial attacks in semantic segmentation. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pages 15681–15690, 2021. 7

[3] Giancarlo Di Biase, Hermann Blum, Roland Y. Siegwart, and César Cadena. Pixel-wise anomaly detection in complex driving scenes. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 16913–16922, 2021. 6

[4] Hermann Blum, Paul-Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. The fishyscapes benchmark: Measuring blind spots in semantic segmentation. arXiv preprint arXiv:1904.03215, 2019. 1, 2

[5] Jun Cen, Peng Yun, Junhao Cai, Michael Yu Wang, and Ming Liu. Deep metric learning for open world semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 15333–15342, 2021. 2, 6

[6] Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Pascal Fua, Mathieu Salzmann, and Matthias Rottmann. Segmentmeifyoucan: A benchmark for anomaly segmentation, 2021. 1, 2, 4

[7] Robin Chan, Matthias Rottmann, and Hanno Gottschalk. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pages 5108–5117, 2021. 2, 7

[8] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. arXiv preprint arXiv:2005.02357, 2020. 1, 2

[9] Marius Cords, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 4

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009. 4

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020. 3

[12] Silvio Galesso, Maria Alejandra Bravo, Mehdi Naouar, and Thomas Brox. Probing contextual diversity for dense out-of-distribution detection. arXiv preprint arXiv:2208.14195, 2022. 2, 3, 6

[13] Matej Grcic, Petra Bevandi’c, and Sinisva vSegvi’c. Dense anomaly detection by robust learning on synthetic negative data. ArXiv, abs/2112.12833, 2021. 7

[14] Matej Grcic, Petra Bevandi’c, and Sinisva vSegvi’c. Dense-hybrid: Hybrid anomaly detection for dense open-set recognition. ArXiv, abs/2207.02606, 2022. 1, 2, 6, 7

[15] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In International conference on machine learning, pages 1321–1330. PMLR, 2017. 4

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. 3

[17] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohamadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. arXiv preprint arXiv:1911.11132, 2019. 1, 2, 3, 4
[18] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15262–15271, 2021.

[19] Sanghun Jung, Jungsoo Lee, Daehoon Gwak, Sungha Choi, and Jaegul Choo. Standardized max logits: A simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 15425–15434, October 2021.

[20] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Advances in neural information processing systems, 31, 2018.

[21] Krzysztof Maciej Lis, Krishna Kanth Nakka, Pascal Fua, and Mathieu Salzmann. Detecting the unexpected via image resynthesis. International Conference On Computer Vision (ICCV), pages 2152–2161, 2019.

[22] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11976–11986, 2022.

[23] MMSegmentation Contributors. OpenMMLab Semantic Segmentation Toolbox and Benchmark, 7 2020.

[24] Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen. Panda: Adapting pretrained features for anomaly detection and segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2806–2814, 2021.

[25] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 14318–14328, June 2022.

[26] Marco Rudolph, Bastian Wandt, and Bodo Rosenhahn. Same but different: Semi-supervised defect detection with normalizing flows. In Proceedings of the IEEE/CVF winter conference on applications of computer vision, pages 1907–1916, 2021.

[27] Oriane Siméoni, Gilles Puy, Huy V. Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. November 2021.

[28] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 7262–7272, 2021.

[29] Yu Tian, Yuyuan Liu, Guansong Pang, Fengbei Liu, Yuanhong Chen, and G. Carneiro. Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes. ArXiv; abs/2111.12264, 2021.

[30] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In International Conference on Machine Learning, volume 139, pages 10347–10357, July 2021.

[31] Yingda Xia, Yi Zhang, Fengze Liu, Wei Shen, and Alan Lorraine Yuille. Synthesize then compare: Detecting failures and anomalies for semantic segmentation. ArXiv, abs/2003.08440, 2020.

[32] Tete Xiao, Yingcheng Liu, Bo lei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In Proceedings of the European conference on computer vision (ECCV), pages 418–434, 2018.

[33] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems, 34:12077–12090, 2021.

[34] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenzhan Peng, Haoqi Wang, Guanyao Chen, Bo Li, Yiyu Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. arXiv preprint arXiv:2210.07242, 2022.

[35] Fisher Yu, Wenyi Xian, Yangying Chen, Fangchen Liu, Mike Liao, Vashishth Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. arXiv preprint arXiv:1805.04687, 2(5):6, 2018.

[36] Hongjie Zhang, Ang Li, Jie Guo, and Yanwen Guo. Hybrid models for open set recognition. In European Conference on Computer Vision, pages 102–117, Springer, 2020.

[37] Hongjie Zhang, Ang Li, Jie Guo, and Yanwen Guo. Hybrid models for open set recognition. In European Conference on Computer Vision, pages 102–117, Springer, 2020.

[38] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiutian Zhu, Zekun Luo, Yahiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 6881–6890, 2021.

[39] Yang Zhou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pretraining for anomaly detection and segmentation. In ECCV 2022, 2022.