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

Silvio Galesso
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. This is particularly challenging for dense anomaly detection in domains where the in-distribution data has a complex underlying structure. Nearest-Neighbors approaches have been shown to work well in object-centric data domains, such as industrial inspection and image classification. In this paper, we show that nearest-neighbor approaches also yield state-of-the-art results on dense novelty detection in complex driving scenes when working with an appropriate feature representation. In particular, we find that transformer-based architectures produce representations that yield much better similarity metrics for the task. We identify the multi-head structure of these models as one of the reasons, and demonstrate a way to transfer some of the improvements to CNNs. Ultimately, the approach is simple and non-invasive, i.e., it does not affect the primary segmentation performance, refrains from training on examples of anomalies, and achieves state-of-the-art results on RoadAnomaly, StreetHazards, and SegmentMeIfYouCan-Anomaly.

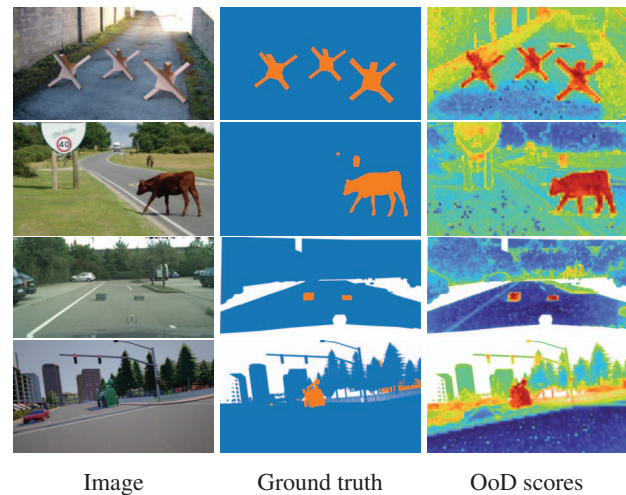| Image	Ground truth	OoD scores

Figure 1: Our approach uses a combination of conventional parametric anomaly detection and k-nearest-neighbors. The resulting anomaly scores can be used to identify semantic anomalies and obstacles (orange in the ground truth) in road scenes and achieve state-of-the-art anomaly detection performance on common benchmarks such as RoadAnomaly, StreetHazards, and SegmentMeIfYouCan-Anomaly.

1. Introduction

Deep learning models can achieve remarkably good performance on a large number of tasks. However, when these models are evaluated on data outside of the training distribution, their performance usually deteriorates substantially [23]. Even worse, the models often do not realize that they are out of distribution and make wrong predictions with high confidence [20]. 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 crucial. 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 deci-

sion 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 years.

In this work, we focus on segmentation datasets for autonomous driving, and we aim to detect and localize objects of unknown categories in the image. This requires spatially resolved outputs, rather than an accumulated decision for the whole image. Moreover, driving scenes comprise diverse patterns and multiple objects (see Figure 1), which makes modelling this complex inlier distribution challenging. The problem has been approached by several works already [29, 50, 55, 18], and a number of accepted evaluation benchmarks exist [22, 4, 7].

Inspired by the success of non-parametric nearest-neighbor methods in the scope of industrial anomaly de-

tection [11, 44, 45] and image recognition [49], we investigate if and how these approaches can be modified to detect anomalous objects in driving scenes. The settings are very different. In industrial inspection, objects are observed under similar conditions, which limits the scope of the in-distribution data, which is advantageous for nearest-neighbors. In OoD image recognition, while the data is more diverse than in industrial inspection, it is still object-centric, and decisions can be based on the global embedding vector. Novelty detection in driving scenes must deal with diverse in-distribution data and decisions must be made locally. Segmentation embeddings are local, and therefore spatially correlated, have high intra-class diversity, and are known to be prone to false positives at class boundaries [29]. Consequently, non-parametric approaches appear unfit to model the in-distribution data effectively, and indeed to the best of our knowledge no successful nearest-neighbor approach has so far been proposed in this setting.

Motivated by recent work showing the outstanding properties of attention-based representations [5, 37, 60], we apply k-nearest-neighbor based OoD detection on the representations produced by transformers models for semantic segmentation. As a key insight of this work, we find that transformer representations are a game changer for dense anomaly detection in driving scenes, outperforming their convolutional counterparts and state-of-the-art approaches. In particular, we find that the multiple heads of the transformer architecture play an important role. We show further evidence of the benefit of multiple heads by extending the idea to CNNs, improving their performance on OoD detection. We see a connection between this finding and the theoretical argument that nearest-neighbor approaches should fail in high-dimensional features spaces ("curse of dimensionality"). Nearest-neighbor approaches have been shown to work well in other settings [45, 49], but in complex road scenes they need the right representations to succeed.

To summarize, our contributions are the following:

- We show that k-nearest-neighbors (kNNs) with deep supervised segmentation features can achieve state-of-the-art performance on all common benchmarks (StreetHazards, RoadAnomaly, SegmentMeIfYouCan, FS Lost&Found), at a favorable computational trade-off and *without* training on out-of-distribution data.

- In contrast to [49], we investigate and find major differences in the performance of several feature encoders. We find attention-based models to be substantially and consistently superior to CNNs.

- We investigate the effects of the "curse of dimensionality" on the performance of kNNs and identify the multi-head design of transformers as a key to their success, demonstrating its beneficial effects on CNNs too.

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 [2, 64]. 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 location of the defect [11, 44, 45].

A more academic evaluation setting for OoD detection is image classification [25, 61, 58], where the normal data distribution is divided into classes, which can in turn have nuanced appearances and subtypes. The goal is to reliably identify images that do not belong to the semantic classes of a training set, such as CIFAR10 [31] or ImageNet [13]. For this purpose several scoring functions have been proposed, mostly based on discriminative parametric models [32, 62, 27, 48, 26, 53, 41], optionally aided by the use of third-party outliers (outlier exposure) [24, 35, 38, 30].

While anomalous image recognition is an essential research problem, safety-critical real world applications such as autonomous driving deal with complex multi-object scenes and require accurate localization of unrecognized objects [22, 4, 7]. This is the setting we address in this work, where we aim to identify the individual pixels that correspond to unknown entities. As in image recognition, the data includes several semantic categories; however, each sample does not correspond to a whole object but to part of an object in relation with its context. Novelty detection in semantic segmentation has been studied less, most methods rely on scoring functions operating on the output of pre-trained segmentation models [22, 29, 15, 6, 33]. Recent approaches [19, 42] make clever use of the mask transformer architecture [9] for segmentation, which decouples mask and class predictions, for improved OoD detection.

A trend in recent works for dense OoD detection is to use of outlier exposure [8, 50, 18], mentioned above.

2.2. OoD Detection with k-Nearest-Neighbors

Unknown patterns in the inference data can be detected by retrieval and comparison with available in-distribution samples. The particular nature and structure of the in-distribution data determines the difficulty and scalability of the task, i.e. how many in-distribution samples are needed to effectively represent the data distribution and distinguish it from anomalous entities, and how easy it is to compute a suitable feature representation.

Retrieval based approaches are very successful at detecting defects in products [11, 44, 45, 64], precisely because they can rely on abundant images of healthy samples with little variation in appearance [2]. These methods use various types of learned deep features in combination with kNNs.

A kNN-based method has recently been proven successful in the image recognition setting, too [49]. Here the training data is more diverse, sometimes featuring 1k categories, but methods can rely on class labels and object-centric, single-instance images, which help the feature extraction process [43]. The evaluation setting in this line of research often uses different datasets for in- and out-of-distribution samples [27, 49].

In dense out-of-distribution detection, each pixel is a sample and a potential anomaly. High correlation with surrounding context further complicates the task, since anomalous pixels are correlated to nearby in-distribution pixels [22, 29]. The outliers are not from another dataset, but part of the scene, coexisting and interacting with the in-distribution objects. The training data ontologies have fewer categories [12, 22], but higher intra-class variation. Results for an embedded kNN approach similar to ours are reported in [4], but without the right representation choice its performance is inferior to most baselines.

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 the local representation maps of the test sample and a library of *reference features* obtained from the in-distribution dataset – i.e. the training set for the segmentation network.

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

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

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

$$s_j^{\mathrm{N}} = \frac{1}{k} \cdot \min_{\substack{D_i' \subset D_i \\ |D_i'| = k}} \sum_{d \in D_i'} d, \quad (2)$$

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

Although different distance functions were tried, L_2 was found to perform best. See Appendix for details.

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 (i.e. *parametric* scores).

In order to combine $\mathbf{S}^{\mathrm{N}}$ with the parametric scores $\mathbf{S}^{\mathrm{P}}$, we first bring them to the same resolution, by upsampling $\mathbf{S}^{\mathrm{N}}$ to the original image size. Subsequently, we simply scale both to the same range using their respective extrema estimated on the training set:

$$\overline{\mathbf{S}^{\mathrm{N}}} = \mathbf{S}^{\mathrm{N}} / \max \mathbf{S}_{\mathrm{train}}^{\mathrm{N}} \quad (3)$$

$$\overline{\mathbf{S}^{\mathrm{P}}} = (\mathbf{S}^{\mathrm{P}} - \min \mathbf{S}_{\mathrm{train}}^{\mathrm{P}}) / (\max \mathbf{S}_{\mathrm{train}}^{\mathrm{P}} - \min \mathbf{S}_{\mathrm{train}}^{\mathrm{P}}) \quad (4)$$

and finally compute the combined scores:

$$\mathbf{S}^{\mathrm{C}} = \overline{\mathbf{S}^{\mathrm{N}}} + \overline{\mathbf{S}^{\mathrm{P}}}. \quad (5)$$

In the following text, we will use the abbreviations DNP (Deep Neighbor Proximity) and cDNP (combined Deep Neighbor Proximity) to refer to the approaches and results based on $\mathbf{S}^{\mathrm{N}}$ and $\mathbf{S}^{\mathrm{C}}$ respectively.

In terms of parametric scores we considered all the best options in recent literature for dense OoD detection. While more results are available in the Appendix, here we report those obtained with *LogSumExp* operator, which performed slightly better. LogSumExp can be interpreted as an energy functional built on a discriminative model's logits [16, 50].

3.2. Model Architectures and Feature Extraction

To assess the versatility of our method, we apply it to four different feature extraction architectures: ResNet [21], ConvNeXt [36], MiT [57] and ViT [14].

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 practice we use the networks above as encoders in semantic segmentation models, and thereby learn the feature

embeddings as part of the standard supervised segmentation training procedure. For ResNet and ConvNeXt we use UPerNet [56], for MiT we use SegFormer [57], and for ViT we use Segmenter [47] and SETR [63], following established practices in semantic segmentation literature.

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 4.3.

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). GCS has been used by PatchCore [45], and consists in a greedy selection procedure aiming to find the subset of $\mathbb{R}$ with the closest solution to the nearest neighbor problem. In [45], GCS is found to greatly outperform random sampling.

PC-GCS is our 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. 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

Our evaluation investigates the influence of: the choice of features (4.3), subsampling strategy (4.7), and number of neighbors (4.9). Section 4.4, shows how our proposed nearest-neighbor-based approaches (Section 3.1) perform in comparison with the parametric baseline, and finally how they compare with current state-of-the-art approaches in Section 4.5. Unless otherwise stated, we use $N = 100k$ reference features, and $k = 3$ neighbors.

4.1. Evaluation Benchmarks and Metrics

Several 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 [22] is a synthetic dataset and benchmark, featuring 12 in-distribution categories in the annotated training/validation sets, and 250 diverse OoD objects, annotated as one category in the test set. Its size (1500 test images), variety of OoD objects and locations

makes it an important benchmark for research. **Road-Anomaly** [34] is a benchmark made of images downloaded from the web, which features objects, such as animals or vehicles, with categories alien to the typical driving ontology e.g. Cityscapes [12] or BDD100k [59]. **SegmentMeIfYou-Can - Anomaly** [7] is an extension of RoadAnomaly, containing mostly images for which the OoD ground truth is undisclosed. **Fishyscapes Lost&Found** [4] is a benchmark for the detection of road obstacles (e.g. lost cargo, small objects), originally designed as an extension of Cityscapes.

Models trained on Cityscapes are used for all benchmarks, except for StreetHazards, which comes with its own training set. Our ablation experiments use RoadAnomaly and StreetHazards. We use the same model checkpoints for all Cityscapes-based benchmarks.

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 (**FPR**$_{95}$), which measures the performance at a high detection threshold - relevant for safety critical applications.

4.2. Training the Feature Extractors

We follow standard training procedures for each of our semantic segmentation models, optimizing for the cross entropy objective using exclusively the respective training dataset. We use the `mmsegmentation` [40] framework and adhere to the default optimization settings when available (see Appendix). For comparability we select 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 overconfidence [20].

All encoders are initialized with the respective publicly available ImageNet [13] pre-trained parameters. For fair comparison, on our ViT models we use the DeiT [51] weights, instead of the original ones trained on a larger undisclosed dataset.

4.3. Choosing the Best Reference Features

The efficacy of our approach with different feature representations is evaluated here, 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.

For the transformer encoders, MiT and ViT, which include self-attention layers, from which *query*, *key* and *value* representations can be extracted, as well as "end-of-the-block" representations [52]. Our results in the Appendix indicate a clear superiority of the self-attention features,

Model	Method	RoadAnomaly		StreetHazards	
		AP↑	FPR$_{95}$↓	AP↑	FPR$_{95}$↓
UPerNet ResNet50	LogSumExp	28.58	62.98	23.87	19.31
	DNP	27.81	54.72	15.30	25.88
	cDNP	33.55	42.14	25.09	14.30
UPerNet ConvNeXt-T	LogSumExp	40.04	59.43	15.79	25.19
	DNP	40.89	40.02	20.30	18.66
	cDNP	46.74	41.72	26.55	13.94
SegFormer MiT-B2	LogSumExp	69.62	27.74	16.11	25.33
	DNP	72.59	18.38	37.33	20.44
	cDNP	77.92	**15.97**	37.44	16.75
Segmenter ViT-S	LogSumExp	56.39	34.54	18.16	28.94
	DNP	79.47	19.75	**45.69**	18.37
	cDNP	**79.78**	18.18	43.89	**15.69**

Table 1: Comparison between the the parametric (Log-SumExp) baseline, DNP, and their combination cDNP, for four encoder/decoder architectures, on RoadAnomaly and StreetHazards. Best results per model are underlined, best overall are bold. DNP/cDNP always outperform LogSum-Exp, cDNP is the best approach overall.

which perform approximately equally and outperform the block features by ∼20% and ∼55% for ViT and MiT respectively. Based on performance, we choose the last layer representations for both transformer-based encoders.

In the following sections, we will use the *keys* as default features for transformer backbones, in accordance with other works [46], but we observed no clear superiority of one over the other.

4.4. Comparing Parametric Scores, DNP, and cDNP

The three out-of-distribution scores defined in Section 3.1, i.e. the model's own parametric scores (LogSumExp) and the nearest-neighbor based approaches: DNP and cDNP are compared in Table 1 for the four feature extractors on RoadAnomaly and StreetHazards. Consistently for all benchmarks, architectures, and metrics, cDNP performs better than its two component scores LogSumExp and DNP. However, the extent of its superiority changes for different architectures and is much greater for the attention-based encoders, particularly for ViT.

In fact, while LogSumExp is better than DNP for ResNet, for the other models the performance of DNP – the pure kNN method – is superior to that of LogSumExp, 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.

4.5. Comparison to the State of the Art

A comparison with the current state-of-the-art on RoadAnomaly, StreetHazards, SegmentMeIfYouCan (SMIYC) Anomaly, and Fishyscapes Lost&Found is shown in the Tables 2(a-c).

While our comparison is focused on methods that do not use outlier exposure for training (OE) – as this technique breaks the interpretation of out-of-distribution data as completely unknown – for completeness we include results for PEBAL [50] and DenseHybrid [18], two recent approaches based on OE. Other notable results are those of M2F-EAM [19], and the concurrent RbA [42]: both cleverly exploit the Mask2Former segmentation model.

Along cDNP-Segmenter-ViT-B and cDNP-ConvNeXt-S, we include results for cDNP-SETR (based on ViT-L) We added the last model to show that the 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.

On RoadAnomaly, StreetHazards, and SMIYC-Anomaly – whose ground truth is undisclosed, cDNP performs best, followed by RbA and M2F-EAM.

In Table 2(c) we report results for Fishyscapes Lost&Found-test. This benchmark focuses on road obstacles which are smaller and with less semantic variation. In this case we also include our method's results using a model trained with outlier exposure, following the protocol described in [8]. Here cDNP outperforms the other methods in terms of AP, but has a higher FPR$_{95}$.

The results on Fishyscapes Lost&Found identify a potential limitation of our method in its current formulation, i.e. comparatively worse performance in detecting very small anomalies. This is due to the lower resolution of the transformer patches and kNN scores. The Appendix has results for SMIYC-Obstacle, another benchmark with small obstacles.

4.6. Qualitative Results

Qualitative examples of our method on RoadAnomaly and StreetHazards, using Segmenter-ViT-B are shown in Figure 2. In both cases, the anomaly score-maps show the superiority of the kNN-based scores compared to the parametric ones: cDNP has both fewer false negatives (it marks most anomalous locations) and false positives (especially at class boundaries and for distant objects).

More examples from both datasets, including results from both ViT-B and UPerNet-ConvNeXt-S are shown in Figure 3. Here the benefits of the combined scores over parametric ones and DNP can be seen, especially for ConvNeXt. While ViT generally performs better than ConvNeXt, the example in the third row is an exception.

A qualitative comparison with the RbA [42] approach is shown in Figure 4. More qualitative results can be found in the Appendix.

4.7. Reference Feature Subsampling Methods

Here we discuss the results on the choice of the subsampling method: random, GCS, PC-CGS, which are sum-

Method	OE	RoadAnomaly AP	RoadAnomaly FPR$_{95}$	StreetHazards AP	StreetHazards FPR$_{95}$
DML[6]		37.0	37.0	14.7	17.3
MOoSe[15]		43.6	32.1	15.2	17.6
PEBAL[50]	✓	62.4	28.3	-	-
DenseHybrid[18]	✓	-	-	30.2	13.0
M2F-EAM[19]		66.7	13.4	-	-
RbA[42]		78.5	11.8	-	-
cDNP-Segmenter-B		85.6	**9.8**	46.2	14.9
cDNP-SETR-L		**85.9**	13.8	-	-

(a)

Method	OE	SMIYC-Anomaly AP	SMIYC-Anomaly FPR$_{95}$
Resynth. [34]		52.3	25.9
PEBAL[50]	✓	49.1	40.8
NFlowJS[17]		56.9	34.7
ObsNet[3]		75.4	26.7
M2F-EAM[19]		76.3	93.9
DenseHybrid[18]	✓	78.0	9.8
RbA[42]		86.1	15.9
cDNP-Segmenter-B		**88.9**	**11.4**

(b)

Method	OE	FS-Lost&Found AP	FS-Lost&Found FPR$_{95}$
M2F-EAM [19]		9.4	41.5
NFlowJS		39.4	9.0
DenseHybrid[18]	✓	43.9	6.2
PEBAL [50]	✓	44.2	7.6
FlowEneDet		50.2	**5.2**
GMMSeg [33]		55.6	6.6
cDNP-Segmenter-B		**62.2**	8.9
cDNP-Segmenter-B	✓	69.8	7.5

(c)

Table 2: Comparison with the state-of-the art on RoadAnomaly (a), StreetHazards (a), SMIYC-Anomaly (b), and Fishyscapes-Lost&Found-test (c). OE denotes the use of outlier exposure, according to each specific approach. Best results *without* OE are shown in bold. Our approach performs best overall, except for Lost&Found, where it has higher FPR$_{95}$.
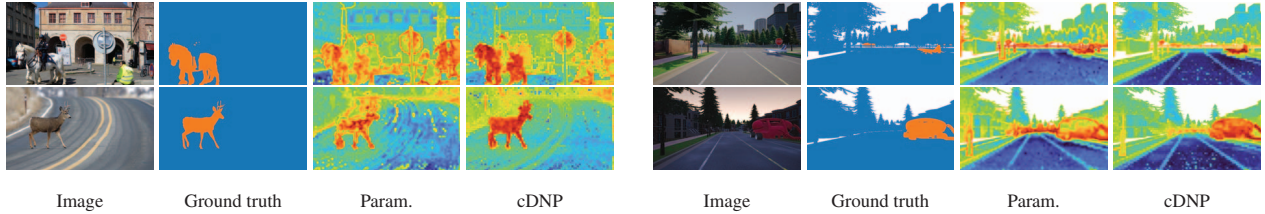

Figure 2: Qualitative results for Segmenter-ViT-B on RoadAnomaly (left) and StreetHazards (right). OoD objects are indicated in orange in the ground truth. It can be observed how cDNP scores are better markers for anomalous entities than the parametric ones. The latter also wrongly mark object boundaries and distant objects as anomalous, more often than cDNP.
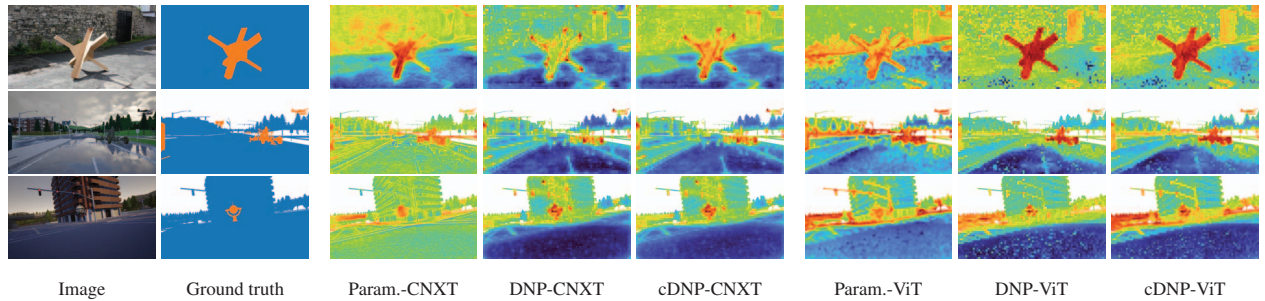

Figure 3: Qualitative results with UperNet-ConvNeXt-S (CNXT) and Segmenter-ViT-B. The top two examples are from RoadAnomaly, followed by two StreetHazards ones. The score maps show how the combination of parametric scores and DNP ones is an improvement over both, mostly through the removal/filtering of false positives. In the first and second row, ViT is clearly outperforming ConvNeXt, whereas in the last row it is the other way around.

marized in Figure 5 for ConvNeXt-T and ViT-S on Road-Anomaly. We evaluate for different values of N: 1k, 10k, 100k, and 1 million for random only, due to the prohibitive pre-processing costs of the coreset approaches. Each setup is evaluated with three random seeds.

The ranking between the methods changes with the architecture and the number of reference samples, but all methods are similarly competitive given enough features. The performance does not saturate with 1M reference samples, however the inference costs in that regime make the approach less attractive. Throughout the following experiments we use PC-GCS, based on its results and faster pre-

processing times than GCS.

4.8. Computational Costs

A major point of concern with k-nearest-neighbors is the computational cost due to the distance computations and search on large feature sets. This depends on two major factors: the number of reference features N, and the network architecture, which determines the test feature resolution $H \cdot W$ and channel size C. We estimate the runtime for each k-nearest-neighbors distance computation using the `IndexFlat` exact search index provided by the `faiss` [28] library, on an NVIDIA RTX 2080Ti card.

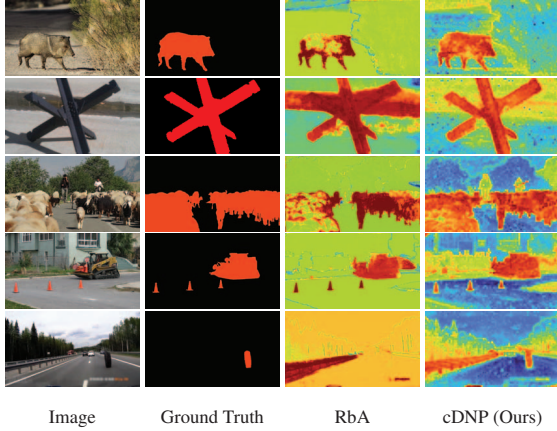Image   Ground Truth   RbA   cDNP (Ours)

Figure 4: Qualitative comparison between cDNP and RbA. The scoremaps from the latter have sharper contours, due to the mask based inference: this can lead to more accurate detection of smaller objects (fourth example), but also to stark false negatives/positives (first three examples). Both methods fail, albeit differently, on the last example.
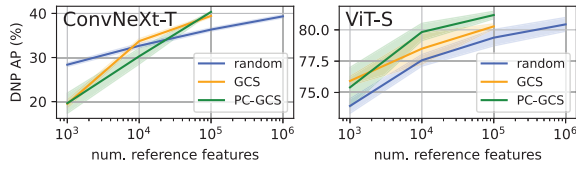

Figure 5: Results of different subsampling methods for ConvNeXt and ViT on RoadAnomaly. On the x-axes are the number of samples, on the y-axes the performance of DNP. All sampling methods perform well given enough reference features, with a slight superiority of PC-GCS.
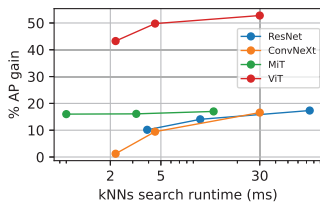

Figure 6: Percentage performance gains in terms of AP (cDNP vs. parametric) over nearest-neighbors search time. For each of the four architectures, three reference set sizes are considered: $N \in \{10^3, 10^4, 10^5\}$.

The trade-off between the performance gain brought by our approach (cDNP) over the best performing parametric scores (LSE) and runtime is shown in Figure 6. The x-axis shows the average kNNs search runtime, for images of 1280×720px. We compare all architectures with reference feature set sizes 1k, 10k, 100k on RoadAnomaly.

ResNet50 is the most expensive architecture, due to its comparatively high feature size and resolution, while MiT, which has the lowest feature resolution, is the least expensive. MiT yields the same relative AP gain as the CNNs, but its absolute performance is much better, as per Table 1. ConvNeXt and ViT have the same cost, but the latter offers the highest AP gain in return.

4.9. Impact of the Number of Neighbors

The number of selected neighbors (k) is an important hyperparameter for any kNN-based approach, and typically depends on the size of the reference set [49] and on the source of the features. An overview of the optimal k values for our ConvNeXt-T and ViT-S models is presented in Figure 7, which shows that for $k > 3$ the performance of both models decreases.

The choice of $k=3$ gives a near-optimal value and ensures a more robust distance expectation (Equation 2).
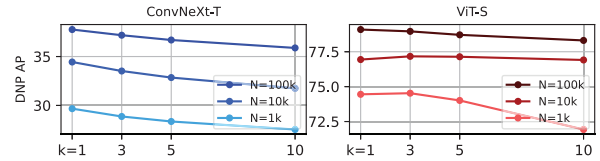

Figure 7: Effect of the number of neighbors (k) on DNP scores for ConvNeXt-T and ViT-S, on RoadAnomaly. The optimal k value depends on the number of reference features N for ViT, but the difference is marginal.

5. Feature Dimensionality and Partitioning

A known limitation of non-parametric pattern recognition approaches, such as kNNs, is that they scale poorly to high dimensional problems. The term "curse of dimensionality" refers to the observation that, as the data dimensionality increases, the data space becomes more sparsely populated and the distance to the nearest sample approaches the distance to the farthest one [54, 1]. This section investigates the effect of this phenomenon in our setting and evaluates the role it plays in the large performance gap between different feature extractors observed in the previous experiments.

Section 5.2 explores the multi-head design as one of the reasons for the success of transformer features, and finds that a similar partitioning strategy can be used to boost OoD detection performance in CNNs.

Lower-order distance functions have also been tested to mitigate the curse of dimensionality, although with mixed success [1, 39]. See results in Appendix.

5.1. Curse of Dimensionality

The results in Section 4 leave no doubt that the representation size does not prevent the use of nearest neighbors.
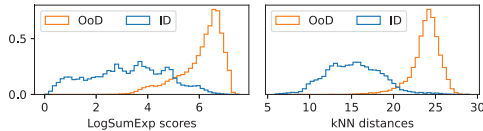
Figure 8: Density histograms of parametric scores and kNN distances, which do not collapse under the curse of dimensionality and show better ID/OoD separation.
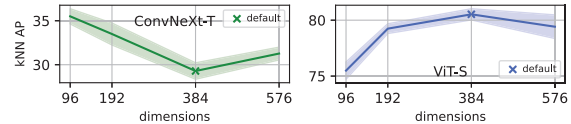

Figure 9: Effect of feature size on OoD detection performance using kNNs. The default size is the optimal for ViT, but smaller features perform better for ConvNeXt.

This is confirmed in Figure 8, where we compare the distributions of kNN distance and parametric scores. Although the kNN distances do not collapse on a single value, it is still possible that their high dimensionality impairs performance. This would partially explain the performance difference between feature extractors reported in Table 1.

To investigate this, we trained versions of UperNet-ConvNeXt-T and Segmenter-ViT-S varying the size of the features used for kNNs. Interestingly, according to the results reported in Figure 9, ConvNeXt features benefit greatly from a reduced dimensionality, whereas ViT representations work best at the default size. This is despite the fact that the default feature size is the same in both cases.

5.2. Feature Groups and kNNs

One major difference between ConvNeXt and ViT is that the latter computes its representations in a multi-head fashion, with each head being able to attend to different portions of the input and being responsible for a specific group of feature dimensions. This results in a functional partitioning of the representations.

We hypothesize that this design choice affects representation learning and impacts how the features behave with kNNs. We propose an ablation study in which the overall feature dimensionality is fixed to its default value (384), but the number of independent feature groups changes.

We achieve feature partitioning in ConvNeXt by using depthwise-separable convolutions [10], these divide the input features into equally sized groups, and process them independently into separate outputs before concatenating.

The results of the ablation study are shown in Figure 10, and show that the two networks behave similarly. For both architectures a clear optimal number of groups exists: for ViT-S it happens to coincide with the default number of heads, whereas for ConvNeXt it's three. Fewer groups result in a drastic performance reduction, whereas the performance with more groups decreases more slowly, likely as a result of reduced representational power per group.

6. 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
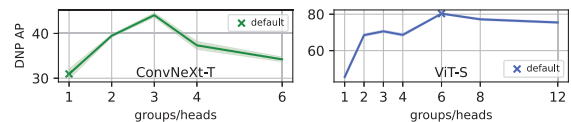

Figure 10: Effect of the number of groups (ConvNeXt) and heads (ViT) on OoD detection performance using kNNs. The behavior is similar for the two architectures, although the optimal number of feature groups is different.

is simple, cost-effective, and achieves state-of-the-art performance on common driving-focused anomaly detection benchmarks. The method is easily combined with standard parametric scores for a performance boost, but also delivers exceptional standalone performance when paired with attention-based models. We conducted a thorough comparative study to verify that the approach performs well on various datasets and is robust to parameter changes.

The large boost in performance we discovered with transformer-based representations is in line with the good feature clustering properties of transformers [5, 37, 60]. We believe that the self-attention mechanism, based on similarity between feature tokens, is a major reason for these advantageous feature properties in similarity-centric approaches, although this is hard to prove. We found clear evidence that the multiple heads have a positive influence on the similarity metric, and we were even able to transfer this advantage to CNNs via corresponding group structures. Our results indicate that this effect is tied to the "curse of dimensionality", which – although not catastrophically – harms the performance of k-nearest-neighbors. A deeper understanding of these effects holds the promise of principled progress in the field.

A limitation of the current approach is its lower resolution, which depends on the encoder architecture and can harm performance in the presence of very small anomalies. Upsampling strategies or alternative architectures can be explored for whenever small-sized objects matter.

References

[1] Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional space. In Jan Van den Bussche and Victor Vianu, editors, *Database Theory — ICDT 2001*, pages 420–434, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg. 7

[2] 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

[3] 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. 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, 3, 4

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 2, 8

[6] 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

[7] 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

[8] 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, 5

[9] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022. 2

[10] Francois Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 8

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

[12] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 4

[13] 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. 2, 4

[14] 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

[15] 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, 6

[16] Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations*, 2020. 3

[17] Matej Grcic, Petra Bevandi'c, and Sinivsa vSegvi'c. Dense anomaly detection by robust learning on synthetic negative data. *ArXiv*, abs/2112.12833, 2021. 6

[18] Matej Grcic, Petra Bevandi'c, and Sinivsa vSegvi'c. Densehybrid: Hybrid anomaly detection for dense open-set recognition. *ArXiv*, abs/2207.02606, 2022. 1, 2, 5, 6

[19] Matej Grcić, Josip Šarić, and Siniša Šegvić. On advantages of mask-level recognition for outlier-aware segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2936–2946, June 2023. 2, 5, 6

[20] 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. 1, 4

[21] 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

[22] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza 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

[23] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. 1

[24] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019. 2

[25] 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. 2

[26] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. 2

[27] Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2, 3

[28] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 6

[29] Sanghun Jung, Jungsoo Lee, Daehoon Gwak, Sungha Choi, and Jaegul Choo. Standardized max logits: A simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15425–15434, October 2021. 1, 2, 3

[30] Julian Katz-Samuels, Julia B Nakhleh, Robert Nowak, and Yixuan Li. Training OOD detectors in their natural habitats. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 10848–10865. PMLR, 17–23 Jul 2022. 2

[31] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2

[32] 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. 2

[33] Chen Liang, Wenguan Wang, Jiaxu Miao, and Yi Yang. GMMSeg: Gaussian mixture based generative semantic segmentation models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 2, 6

[34] 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. 4, 6

[35] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21464–21475. Curran Associates, Inc., 2020. 2

[36] 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. 3

[37] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8364–8375, 2022. 2, 8

[38] Yifei Ming, Ying Fan, and Yixuan Li. POEM: Out-of-distribution detection with posterior sampling. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 15650–15665. PMLR, 17–23 Jul 2022. 2

[39] Evgeny M. Mirkes, Jeza Allohibi, and Alexander Gorban. Fractional norms and quasinorms do not help to overcome the curse of dimensionality. *Entropy*, 22(10), 2020. 7

[40] MMSegmentation Contributors. OpenMMLab Semantic Segmentation Toolbox and Benchmark, 7 2020. 4

[41] Peyman Morteza and Yixuan Li. Provable guarantees for understanding out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7831–7840, 2022. 2

[42] Nazir Nayal, Mısra Yavuz, João F. Henriques, and Fatma Güney. Rba: Segmenting unknown regions rejected by all, 2023. 2, 5, 6

[43] Senthil Purushwalkam and Abhinav Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3407–3418. Curran Associates, Inc., 2020. 3

[44] 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. 2

[45] 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. 2, 4

[46] 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. 5

[47] 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. 4

[48] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. 2

[49] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. *ICML*, 2022. 2, 3, 7

[50] 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. 1, 2, 3, 5, 6

[51] 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. 4

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszko-reit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Il-lia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vish-wanathan, and R. Garnett, editors, *Advances in Neural Infor-mation Processing Systems*, volume 30. Curran Associates, Inc., 2017. 4

[53] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Pro-ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[54] Roger Weber, Hans-Jörg Schek, and Stephen Blott. A quan-titative analysis and performance study for similarity-search methods in high-dimensional spaces. In *Very Large Data Bases Conference*, 1998. 7

[55] Yingda Xia, Yi Zhang, Fengze Liu, Wei Shen, and Alan Lod-don Yuille. Synthesize then compare: Detecting fail-ures and anomalies for semantic segmentation. *ArXiv*, abs/2003.08440, 2020. 1

[56] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understand-ing. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 4

[57] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transform-ers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 3, 4

[58] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmark-ing generalized out-of-distribution detection. *arXiv preprint arXiv:2210.07242*, 2022. 2

[59] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tool-ing. *arXiv preprint arXiv:1805.04687*, 2(5):6, 2018. 4

[60] Andrii Zadaianchuk, Matthaeus Kleindessner, Yi Zhu, Francesco Locatello, and Thomas Brox. Unsupervised se-mantic segmentation with self-supervised object-centric rep-resentations. *arXiv preprint arXiv:2207.05027*, 2022. 2, 8

[61] 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. 2

[62] 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. 2

[63] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmen-tation from a sequence-to-sequence perspective with trans-formers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021. 4

[64] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *ECCV 2022*, 2022. 2