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

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

Max Argus
University of Freiburg

Thomas Brox
University of Freiburg

| Model | Features | Ft. size | Ft. res. | AP (DNP) |
|---|---|---|---|---|
| ResNet50 | Stage 1 | 256 | 180×320 | 16.9 |
| | Stage 2 | 512 | 90×160 | 20.8 |
| | Stage 3 | 1024 | 45×80 | 25.6 |
| | Stage 4 | 2048 | 22×40 | 21.2 |
| ConvNeXt-T | Stage 1 | 96 | 180×320 | 17.0 |
| | Stage 2 | 192 | 90×160 | 17.2 |
| | Stage 3 | 384 | 45×80 | 32.5 |
| | Stage 4 | 768 | 22×40 | 27.9 |
| MiT-B2 | Stage 4 - Q | 512 | 23×40 | 79.4 |
| | Stage 4 - K | 512 | 23×40 | 76.0 |
| | Stage 4 - V | 512 | 23×40 | 77.9 |
| | Stage 4 | 512 | 23×40 | 48.6 |
| ViT-B | Layer 12 - Q | 768 | 45×80 | 78.1 |
| | Layer 12 - K | 768 | 45×80 | 85.4 |
| | Layer 12 - V | 768 | 45×80 | 77.7 |
| | Layer 12 | 768 | 45×80 | 71.6 |

Table 1: Overview of the feature selection results, including feature sizes and resolutions, as well as the performance of DNP on each, in terms of AP on RoadAnomaly.

## 1. Ablation: Feature Selection for kNNs

Here we present the results of our approach on different types of features, as summarized in the main paper.

In Table 1 we report feature size (number of channels), resolution, and respective DNP performance for different feature options within the four encoders (ResNet, ConvNeXt, MiT, ViT). Most importantly, the results confirm the superiority of self-attention features (queries/keys/values) for MiT and ViT.

## 2. DNP with Self-Supervised Representations

Although in this work we mostly apply DNP to feature representation which have been trained for semantic segmentation in a supervised way, the approach could in principle be applied to other types of representations. This is because it relies on a set of in-distribution reference features which carry information about the training data.

| | RoadAnomaly AP | RoadAnomaly $FPR_{95}$ |
|---|---|---|
| DNP ViT-B iBOT | 55.28 | 19.72 |
| DNP ViT-B DINO | 67.83 | 18.99 |

Table 2: DNP performance on RoadAnomaly using representations from self-supervision approaches iBOT and DINO. While the supervised features (trained for semantic segmentation on Cityscapes) yield the best results, DNP with DINO features outperforms the other architectures.

To explore the capabilities of our method in this direction, we combined it with feature extractors trained via self-supervision, using the popular iBOT [7] and DINO [1] approaches, known to perform well on dense and local downstream tasks such as segmentation.

Table 2 shows the resulting DNP performances using parameters obtained from the respective iBOT and DINO official repositories, compared to the supervised features which we trained for semantic segmentation. Although the supervised representations obtain the best results, DINO features perform well with an AP of 67.83, outperforming the CNN architectures.

## 3. Additional Results

### 3.1. SegmentMeIfYouCan - Obstacle

Here we report our results on the Onstacle track of the SegmentMeIfYouCan (SMIYC) benchmark, which considers obstacles on the road surface.

Table 3 shows the official leaderboard results on the benchmark's test set (undisclosed ground truth). Excluding outlier exposure, there are two methods that currently

| Method | OE | AP | $FPR_{95}$ |
|---|---|---|---|
| DaCUP [6] | | 81.50 | 1.13 |
| NFlowJS | | 85.55 | 0.41 |
| Maximized Entropy | ✓ | 85.07 | 0.75 |
| DenseHybrid | ✓ | 87.08 | 0.24 |
| cDNP-Segmenter-ViT-B | | 72.70 | 1.40 |

Table 3: Results on the SMIYC-Obstacle test benchmark. OE marks the use of outlier exposure.

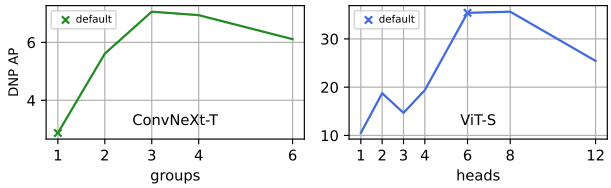


Figure 1: DNP performance of ConvNeXt and ViT features wit different numbers fo groups and heads respectively.

outperform cDNP. The first is NFlowJS, which uses synthetic outliers from a generative model, and the second is DaCUP [6], which is based on inpainting reconstruction error. It should be noted that NFlowJS performs worse than cDNP on SMIYC-Anomaly (oriented towards semantic anomalies, reported in the main paper), and DaCUP is specifically designed for road obstacle detection.

### 3.2. Feature Partitioning - Fishyscapes Lost&Found

In this section we extend the results of Section 5.2 of the main paper. In Figure 1 we report the performance on Fishyscapes Lost&Found-val of the modified ConvNeXt-T and ViT-S backbones, with different numbers of convolutional groups and transformer heads respectively. The models evaluated here are the same as those evaluated in the main paper.

The average precision (AP) of DNP in both cases follows the same behavior as on RoadAnomaly, i.e. the same optimal number of groups/heads and performance that decreases rapidly when fewer greoups/heads are used.

### 3.3. Feature Partitioning - Segmentation Performance

In Table 4 we report the in-distribution segmentation performance of the models involved in the feature partitioning ablation (Section 5.2 of the main paper). The results show that the two architectures (ConvNeXt and ViT) behave differently in terms of mIoU, and confirm that there is no direct relation between segmentation and OoD detection performance.

It should also be noted that the segmentation and OoD

| Groups/ Heads | ConvNeXt mIoU | ConvNeXt AP | ViT mIoU | ViT AP |
|---|---|---|---|---|
| 1 | 76 | 31 | 63 | 45 |
| 2 | 76 | 39 | 66 | 69 |
| 3 | 75 | 44 | 67 | 71 |
| 4 | 75 | 37 | 69 | 69 |
| 6 | 75 | 34 | 71 | 80 |

Table 4: Segmentation (mIoU) and OoD detection (AP) performance for the models of the feature partitioning ablation, on Cityscapes/RoadAnomaly. For ConvNeXt the segmentation performance doesn't change substantially, and is inversely proportional to the number of groups. For ViT the segmentation performance increases with the number of heads.

detection performances are overall negative affected by the ablation protocol, which involves discarding the pre-trained weight initialization for the last stages.

### 3.4. Qualitative Results

**Feature Partitioning** Figure 2 shows a qualitative comparison between ViT with 1 head and 6 heads, respectively the worst and best versions of the network in the ablation study.

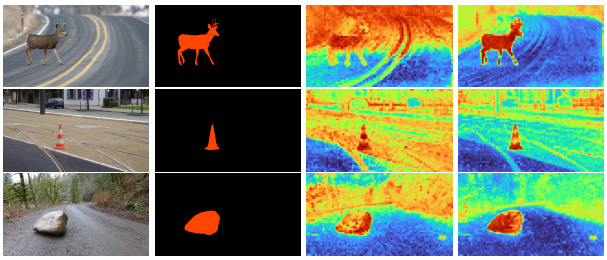


Figure 2: Qualitative examples for the feature partitioning ablation, showing results on RoadAnomaly examples, for ViT-S with 1 and 6 heads. Anomalous pixels are shown in red in the ground truth. The single head model struggles both with false negatives and false positives, and assigns a higher anomaly score to edges and backgrounds.

**State-Of-The-Art** In Figure 3 we show additional qualitative results for our approach, compared with recent approaches DenseHybrid and PEBAL. In general, compared to cDNP, the other approaches suffer from more false negatives and false positives respectively.

**Lost&Found** Figure 4 contains qualitative examples of parametric, DNP, and cDNP scores on Fishyscapes

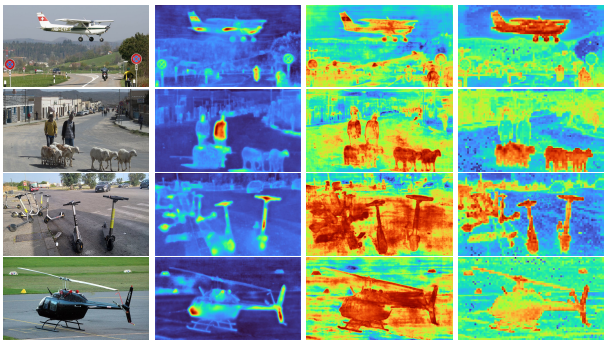

Image DenseHybrid PEBAL cDNP

Figure 3: Qualitative comparison between cDNP and most recent state-of-the art methods DenseHybrid and PEBAL. On the first three examples, cDNP is the only method that correctly and entirely identifies the anomalous samples (airplane, sheep, and scooters): DenseHybrid misses many parts of the objects, and PEBAL suffers from false positives. The forth example is challenging for all approaches: PEBAL is the only approach to detect the helicopter entirely, still producing false positives.

Lost&Found samples. As seen in other examples, the parametric (LogSumExp) scores suffer from frequent false positives, especially in correspondence of unusual terrain textures, which don't affect the nearest-neighbor based scores.

## 4. Comparison of Parametric OoD Scoring Functions

Here we compare the known parametric scoring functions which are available in the literature for dense OoD detection: maximum-softmax-probability [4] (MSP), prediction entropy [3] (H), maximum-logit [3] (ML), and LogSumExp [2, 5] (LSE).

The results of the comparison, reported in Table 5, reveal the superiority of maximum-logit and LogSumExp scores, with the latter outperforming the former.

## 5. Alternative Distance Functions for kNNs

In this section we report the performance of kNNs/DNP using other distance functions than the $L_2$/Euclidean used throughout the paper. In particular, we consider $L_1$ distance and cosine similarity. We choose the former since low-order distance functions have been reported to mitigate the effects of the curse of dimensionality, and the latter because the transformer features we consider are part of the scaled-dot-product attention mechanism.

The results, shown in Table 6 for Segmenter-ViT-B on RoadAnomaly, reveal that $L_1$ and $L_2$ perform very similarly, both significantly better than cosine similarity.

| Model | Scores | Parametric | | cDNP | |
|---|---|---|---|---|---|
| | | AP | $FPR_{95}$ | AP | $FPR_{95}$ |
| UperNet-ConvNexT-T | MSP | 23.53 | 66.05 | 29.54 | 45.52 |
| UperNet-ConvNexT-T | H | 28.34 | 65.79 | 34.33 | 45.67 |
| UperNet-ConvNexT-T | ML | 39.31 | 59.50 | 43.53 | 41.12 |
| UperNet-ConvNexT-T | LSE | 40.04 | 59.43 | 44.02 | 40.83 |
| Segmenter-ViT-S | MSP | 35.23 | 41.63 | 63.22 | 27.18 |
| Segmenter-ViT-S | H | 45.10 | 40.26 | 70.44 | 25.76 |
| Segmenter-ViT-S | ML | 51.58 | 35.16 | 78.25 | 20.59 |
| Segmenter-ViT-S | LSE | 56.39 | 34.54 | 79.42 | 19.74 |

Table 5: Comparison results for different parametric scoring functions on RoadAnomaly. We report the parametric performance and the final combined one (cDNP). LogSumExp (LSE) performs best, followed by maximum-logit (ML).

| Dist./sim. | AP | $FPR_{95}$ |
|---|---|---|
| cosine | 80.71 | 13.93 |
| $L_1$ | 85.29 | 8.32 |
| $L_2$ | 85.83 | 8.26 |

Table 6: Results for DNP on RoadAnomaly, using Segmenter-ViT-B features and different distance/similarity functions in the embedded space.

| Model | optimizer | LR | mIoU | | cDNP-AP | |
|---|---|---|---|---|---|---|
| | | | CS | SH | RA | SH |
| UperNet-ResNet50 | SGD | $10^{-2}$ | 78 | 66 | 34 | 25 |
| UperNet-ConvNeXt-T | AdamW | $10^{-4}$ | 81 | 72 | 47 | 27 |
| SegFormer-MiT-B3 | AdamW | $10^{-4}$ | 72 | 69 | 78 | 37 |
| Segmenter-ViT-S | SGD | $10^{-3}$ | 72 | 61 | 80 | 44 |
| SETR-Naive-ViT-L[1] | SGD | $10^{-2}$ | 80 | - | 86 | - |

Table 7: Overview of the optimization details – algorithm and learning rate – and semantic segmentation (in-distribution) performance for the considered architectures in terms of mIoU on Cityscapes (CS) and StreetHazards validation (SH). We also report the out-of-distribution detection performance on RoadAnomaly (RA) and StreetHazards test (SH).

## 6. Training Details

For the experiments on Cityscapes we use a batch size of 8 and randomly crop the input image and ground truth to 769×769 pixels. For StreetHazards we use a batch size of 4 and a crop size of 512.

The optimization algorithms and learning rates are the same for both datasets, and are listed in Table 7, along with the segmentation performance of each architecture on the in-distribution validation sets. UperNet-ConvNeXt performs best on semantic segmentation. While SegFormer-MiT and Segmenter-ViT have inferior mIoUs, the other transformer-based model (SETR) has a competitive segmentation performance and a state-of-the-art OoD detection performance with cDNP.

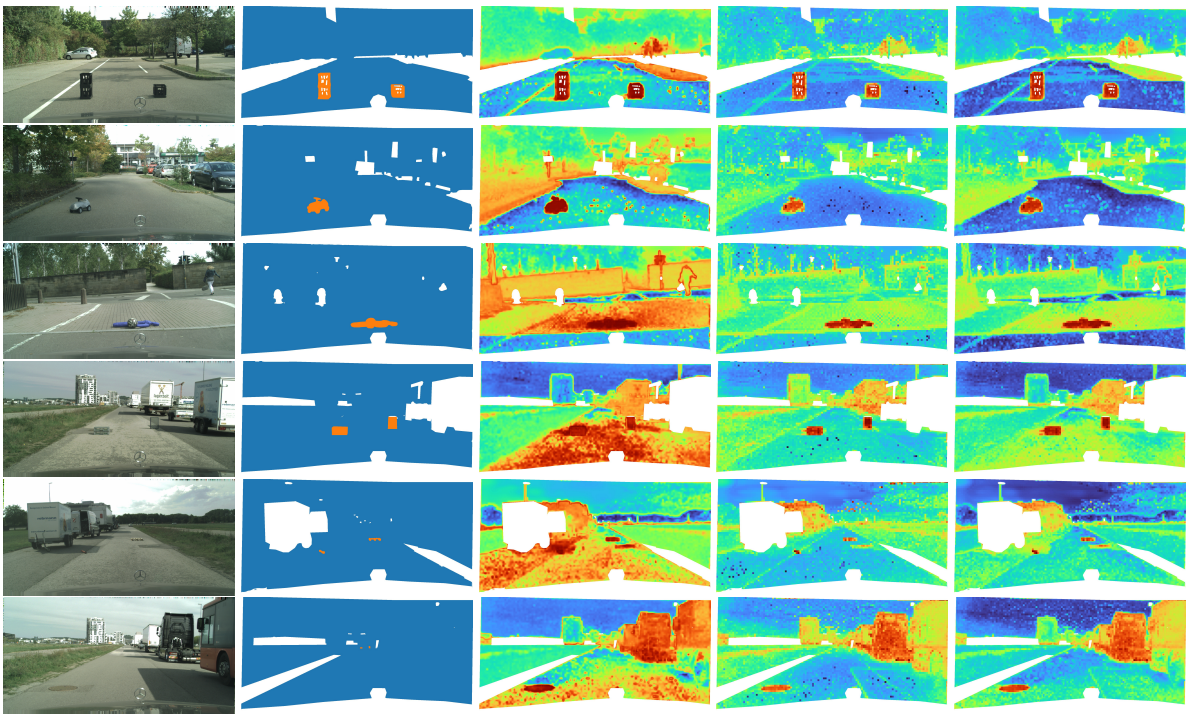


Figure 4: Qualitative examples of our approach on Fishyscapes Lost&Found, showing the parametric (LogSumExp), DNP and cDNP scores for the best model reported in Table 2c of the main paper. The ground truth shows the valid in and out of distribution pixels in blue and orange respectively. DNP and cDNP exhibit fewer false positives than LogSumExp on all examples, especially on unusual terrains (rows 3, 4, and 5). DNP/cDNP can also successfully identify the small obstacles in the 5th row example. All methods fail on the example in the last row.

## References


[1] 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. 1

[2] 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

[3] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. A benchmark for anomaly segmentation. *arXiv preprint arXiv:1911.11132*, 1(2):5, 2019. 3

[4] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. 3

[5] 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. 3

[6] Tomáš Vojíř and Jiří Matas. Image-consistent detection of road anomalies as unpredictable patches. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5491–5500, January 2023. 2

[7] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *International Conference on Learning Representations (ICLR)*, 2022. 1


[1] From: `https://github.com/open-mmlab/mmsegmentation/blob/master/configs/setr/README.md`