Entropy Maximization for Anomaly Detection in complex driving scenes using a UNet architecture

Said Harb, Ryan Benvenuti, Arindam Laha

CSC 592: Algorithms for Big Data

December 21, 2023

Abstract

Alongside the development of self-driving cars an urgency for anomaly detection software has risen. Anomaly detection is crucial in the context of autonomous driving, as an undetected anomaly can have dangerous consequences. To solve this problem, different approaches have been developed to extend the semantic segmentation of images and videos with anomaly detection. One of these approaches is to retrain semantic segmentation models to maximize the entropy of anomalous objects in an image and thereby being able to discern between anomaly and non-anomaly. What this approach has in common with others is that it is applied to state-of-the-art machine learning models. These require heavy computational costs and are therefore not suited for scenarios with limited computing power. For this reason, the application of entropy maximization to a lightweight UNet is researched in this paper with promising results. Even lightweight models like the UNet strongly benefit from entropy maximization, which enables them to detect anomalies, although this comes with a decrease in semantic segmentation performance.

Keywords— Anomaly Detection, Entropy Maximization, Autonomous Driving, Semantic Segmentation, Machine Learning

1 Introduction

1.1 Computer Vision

In the modern world, the necessity of the analysis and monitoring of vast amounts of data and information is ever-increasing. One notable example is the field of computer vision, a branch of computer science that aims to create computer algorithms that can better analyze and understand images and videos. Such information can greatly help in many fields such as manufacturing to detect faulty products, medicine to more quickly and accurately analyze and diagnose patients, and related to the main focus of this paper in surveillance to analyze and label objects for each frame of a video [1], [2].

Machine learning algorithms can be applied in the field of computer vision in different ways. Objects in images can be classified and localized by bounding boxes. The next step is object detection, where different instances of objects are located and categorized. The classification of different objects is also possible on pixel level, which is called semantic segmentation. The method of locating different instances of objects pixelwise is called instance segmentation. This research is focused on the field of semantic segmentation of complex urban driving scenes. In our case there will be a label assigned to each pixel from a predefined set of labels [1], [2].

One notorious issue that frequently appears in computer vision is anomalies. These are objects on which the model was not trained on, therefore it cannot recognize the object. For autonomous driving vehicles, this can lead to dangerous situations. Semantic segmentation models react differently to anomalies in images. Anomalies can be detected as one class or as a blend of many different classes. For example, a dog could have its head read as a person, its body the road, and various parts around it to be a bicycle. In other occasions the anomaly is not detected at all, which is a dangerous case in a driving scene, when anomalies on the street are recognized as road. If a self-driving car were to mislabel debris in the middle of the road as more road, that could result in a collision. Since training a model on all possible objects that exist in real life is not possible, anomaly detection is an important topic in computer vision. [3]–[6].

1.2 Research Question

Our goal in this paper is to retrain two semantic segmentation models, a state of the art DeeplabV3+ and a more lightweight UNet, for entropy maximization to detect anomalies in complex urban driving scenes. The idea of entropy maximization is about encouraging the neu-

ral network to assign high entropy to out-of-distribution objects, i.e. anomalies. This is done by modifying the loss function of the training so that the models are trained to assign uniform probabilities for anomalies, while still giving in-distribution object high probabilities. After this, a binary threshold is better suited to classify anomalies than before training [3]. In the context of autonomous driving, this decreases the risk of not detecting an anomaly. By retraining two models of different complexity, our goal is to determine if entropy maximization also works on a less complex neural network. This can be important to reduce computational effort and costs in self-driving cars.

1.3 Models

There are many different models that are deployed for the purpose of analyzing images. In our case, the DeepLabV3+ is supported by a WideResNet50 and the UNet is supported by a ResNet18. Wide Residual Networks (WideResNet or WRN) act as feature extractors and consist of increased width and decreased depth compared to conventional residual networks. Using this architecture, high performance can be achieved. In this network the data is processed by Convolutional Neutal Networks (CNN)which are combined with skip connections from the residual network architecture [7]. In CNN's multiple kernels with specified dimension and trainable weights iterate over images. Thereby, they create feature maps, on which a ReLU function is applied. Then often a max pooling layer is implemented, where again a kernel iterates over the feature maps, but this time only saving the maximum value. This way, the size of the feature maps are compressed and features of an image can be extracted. A deep neural network after the CNN is then able to perform for example semantic segmentation with the input picture.

UNet employs an encoder-decoder structure and encodes an image through multiple pooling and convolution layers, and then decodes it by going through the same layers only in reverse. The contracting path captures high level features while reducing the spatial dimension of the input. The bottleneck retains high level semantic information in a compact representation and the decoder path increases the spatial dimension again. Between the layers skip connections are implemented and after the decoder a final layer produces the semantic segmentation scores for all classes pixelwise. UNet models also work well with smaller training data pools, which helps to making the program more lightweight. Usually it is applied in medical imaging, but a use case in driving scenes is also possible [8].

The DeepLabv3+ model combines a similar encoder-decoder structure from the UNet with a spatial pyramid pooling module. Also, Atrous Spatial Pyramid Pooling is applied in the network. The DeepLabV3+ model is a state-of-the art model with high scores on popular benchmarks

like Pascal VOC or Cityscapes [9]. The authors of [3] already proved good results for entropy maximization using the DeepLabV3+. This is why we will first replicate their findings and then try to adapt their approach to a UNet, to see if similar performance can be reached.

1.4 Evaluation Metrics

To measure the performance of both the DeeplabV3+ and the UNet before entropy maximization training and after, four metrics are going to be used. The first one is the area under receiver operating characteristic (AUROC), which depicts for a binary threshold the true positive rate (TPR) over the false positive rate (FPR) threshold independent. The TPR shows what proportion of anomalies were found and the FPR indicates what proportion of non-anomalies were classified as anomalies. The area under this characteristic can be seen as a representative for the probability that a binary classifier is able to detect anomalies or as the degree of separability of the system [3]. A small AUROC for example means, that the differences between the softmax entropy between anomalies and non-anomalies are rather small, making it hard to discern between the two. On the other hand, a large AUROC can indicate large differences in the softmax entropy between anomalies and non-anomalies, making it easy to threshold them. A AUROC of 0.5 indicates that the binary classifier is no better than random chance and a value of 1 represents a perfect classifier.

The second metric is the area under precision recall curve (AUPRC), that depicts the precision (TP/(TP+FP)) over the recall or the true positive rate (TPR). The precision basically measures what proportion of anomalies detected were indeed anomalies. Because the AUROC does not consider class imbalances, the AUPRC is an alternative metric, that focuses more on the detection of false positives. A AUPRC value of 1 represents a perfect classifier and a value of 0.5 a classifier that is not better than random chance.

Another metric for evaluation will be the false positive rate at 95% true positive rate (FPR95), which can be see as the probability that an inlier sample will be removed when the 95% of out-of-distribution samples are rejected. Therefore, it can be used to analyze imbalanced datasets because the FPR at a set TPR can be observed [10].

The last metric used is the mean intersection over union (mIoU) and in contrast to the other metrics it measures the performance of the semantic segmentation instead of anomaly detection. This metric derives from the intersection over union metric, that shows proportion of the overlap of prediction and ground truth over the union of both.

$$IoU = \frac{TP}{TP + FP + FN} \tag{1}$$

The mean IoU is the average IoU over multiple samples and the closer it is to 1, the better the segmentation [11].

2 Related Work

In consideration of the need for anomaly detection software, much research has been done in that field. One approach is presented by Di Biase et al. [4], who propose a method that combines uncertainty and resynthesis techniques. In this method, uncertainty measurement maps, such as softmax entropy and softmax distance, are employed to identify differences between the input image and an image re-synthesized from the predicted semantic map. By combining both techniques, dissimilarities with the input image can be recognized.

Another method that Grcić et al. [5] introduces is an innovative method designed to address the intricate challenges of anomaly detection in dense environments. It does so by combining generative modeling of regular training data and discrimination against negative training data with distinct failures. The DenseHybrid model tackles this challenge by incorporating an unnormalized joint distribution based on a shared convolution representation. A significant aspect of the research involves adjusting the method to enable training on mixed-content images. This adaptation streamlines the training process by eliminating the need for backpropagation. The model also addresses the open-set recognition challenge by constructing a model that employs thresholding on the hybrid anomaly score. This results in the model's output not only identifying anomalies but also recognizing inliers, making it a valuable tool for comprehensive scene understanding.

Research by Tian et al. [6] primarily delves into the concept of pixel-wise energy-biased abstention learning (PEBAL). This novel approach combines pixel-wise abstention learning (AL) with an adaptive pixel-level anomaly class, utilizing an energy-based model (EBM). The energy-based model involves the training of deep learning models, where the energy value serves to compute the probabilities that a sample belongs to the inlier distribution. While EBMs typically necessitate the computation of the partition function, which may not yield accurate high-resolution images, the proposed methodology employs the logsumexp operator for estimating the energy score. By minimizing the energy of inliers and employing an Outlier Exposure (OE) strategy to maximize the energy of outliers, the partition function calculation is circumvented. By combining the strengths of energy-based models and abstention learning, PEBAL aims to enhance anomaly segmentation on complex urban driving scenes. The joint training approach ensures that the model effectively identifies and integrates high-energy anomaly pixels, providing

a more nuanced and adaptive solution compared to existing methods.

In the paper from Chan et al. entropy maximization is applied to the input images, causing anomalies to be assigned uniform probabilities, which leads to them having high entropy. The second step for this procedure is a meta classification which applies linear models on a set of handcrafted metrics to remove false positives from the first step. [3]. In this research, the first step of the Chan et al. paper is reproduced on the DeepLabV3+ and then tested out on the UNet. The reason we chose this approach is that the model is described as a lightweight approach to anomaly detection. Part of the big data problem is the ability to use algorithms to process large amounts of data without having access to large computational power. While using complex CNN's and deep neural networks produces good results, having the processing power and storage space to train and run these networks is not always feasible, especially in certain situations where a neural network like this could be useful. For example, in self-driving cars having a stable connection to the cloud or installing a large model into the car costs too much, but this more lightweight approach would be easier and less resource intensive to implement.

Due to time constraint and the scope of this project, the meta classifier will not be implemented. The entropy maximization as a means of anomaly detection is tested and the meta classification would improve the results, but our focus is on entropy maximization. Though despite that, the entropy maximization we're implementing will still preform the main task we're attempting to do, that being detecting anomalies.

3 Methodology

3.1 Datasets

For this research, different datasets need to be used according to the approach described in [3]. For the semantic segmentation the Cityscapes dataset, for the entropy maximization the Cityscapes and COCO dataset and for evaluation the Fishyscapes dataset will be used.

The Cityscapes dataset was made to be used as semantic segmentation training data. It consists of 30 classes and 5000 annotated images. All the pictures were taken in urban scenes from 50 German cities. The angle of the pictures is from the hood of the car facing the street, therefore you can see the ego vehicle. In our research we focus on the following classes and ignore all other labels: road, sidewalk, building, wall, fence, pole, traffic light, traffic sign, vegetation, terrain, sky, person, rider, truck, bus, train, motorcycle and bicycle. This choice is made with respect to the available computing power and the scope of the project. The split of the dataset

is 2975 test images, 500 validation images and 1525 test images [12].

The COCO (Common Objects in Context) dataset was also made for computer vision tasks such as object detection or segmentation. In contrast to the Cityscapes dataset, it contains images from complex everyday scene with common objects. These objects are in their natural context (for example sheep on grass, cup on a table). There are 91 different object types which are said to be easily recognizable. There are over 328 000 images with over 2.5 million instances [13].

The third dataset is the Fishyscapes benchmark, which is the third dataset being used for evaluation of the anomaly detection. Fishyscapes was again made for computer vision tasks related to autonomous driving vehicles in complex urban scenes. In particular, we are using the Fishyscapes Static dataset, which is made out Cityscapes validation images with overlayed objects from the Pascal VOC dataset [14]. These are only objects which are not instances in Cityscapes, like airplanes, birds, cast or sofas. The size of the dataset is 30 images [15].

3.2 Semantic Segmentation

The Cityscapes dataset is used for the initial model, i.e. the model which is only used for semantic segmentation and which was not trained yet for anomaly detection. In this paragraph a description on how the training for semantic segmentation works is given. This is how the UNet can be trained for semantic segmentation.

In semantic segmentation training, the inputs are train images and their respective labels. During the process of training, the model assigns scores for each class pixelwise to the image. Then the softmax function is applied on the result, which turns the scores into predicted probabilities for each class per pixel. The loss function used is the Cross-Entropy loss, which one-hot-encodes the target and compares it with the predicted probabilities of the image. In detail, the predicted probabilities of the input image per pixel for each class are between zero and one. The target for this pixel consists of a probability of zero for all classes, except for the class which is the ground truth. This class has a probability of one, because this class is known to be the correct class. Therefore, the model is trained to assign high scores to the correct class and low scores to wrong classes. This is done through backpropagation and the use of the Adam optimizer, which serves as a proxy for stochastic gradient descent. Using the Adam optimizer, the learning rate can be adapted dynamically [16]. For our UNet architecture we will use pretrained weights from a ResNet18 architecture publicly available in [17]. The architecture behind the UNet can be found in [18].

3.3 Anomaly Detection

A working semantic segmentation model is the basis of an anomaly detection model. In this research the approach presented in [3] will be followed, as it is an efficient, yet lightweight method for anomaly detection. The authors present a two-step solution consisting of entropy maximization and meta classification, of which we are implementing only the first one due to the scope of this project. The code of [3] can be retrieved from [19]. Our research focuses on the aspect of entropy maximization and applying this to a UNet-architecture.

Therefore, entropy maximization will be applied to the state of the art DeepLabV3+ model with a WideResNet38 backbone, which is the model used by the researchers. Then the entropy maximization will be applied to a UNet architecture supported by a ResNet-18 backbone. This is not such a strong model like the DeepLabV3+ and it was not tested by the researchers. For this reason, the comparison of the effect of entropy maximization on the DeepLabV3+ and the UNet will be compared and a conclusion on the applicability of a UNet architecture for anomaly detection will be drawn. First, the process of entropy maximization will be explained.

To be able to train a semantic segmentation model for entropy maximization, an appropriate dataset needs to be created. In this case, a 10 to 1 mix of the Cityscapes and COCO dataset will serve as training data. The labeling policy of Cityscapes stays the same as for the semantic segmentation training, but the labeling of COCO is changed. In particular, images from COCO are filtered so that no instances of Cityscapes instances show up in COCO and there are only two classes: Anomaly and road. This is because the COCO dataset contains images in their natural context, with no instances of Cityscapes instances. Therefore, the object in its natural context is labeled as the anomaly and the natural context must be road in context of Cityscapes [3].

With the mixed dataset at hand, the retraining of the semantic segmentation model (DeepLabV3+ and UNet respectively) can begin. The difference to conventional semantic segmentation is, that a modified cross entropy loss function is applied, which enables multitask learning of the model. These tasks are the maximization of entropy for out-of-distribution objects and semantic segmentation. The formulation of the loss function is as follows, where $(x, y(x)) \sim D_{in}$ are the in-distribution pixels, $y(x)\epsilon C$ is the ground truth label class of the input x, C are the classes and $x' \sim D_{out}$ are the out-of-distribution pixels [3]:

$$L = (1-\lambda)\mathbb{E}_{(x,y(x)\sim D_{in}}\left[l_{in}(f(x), y(x))\right] + \lambda\mathbb{E}_{x'\sim D_{out}}\left[l_{out}(f(x'))\right], \lambda\epsilon[0,1] \tag{2}$$

$$l_{in}(f(x), y(x)) := -\sum_{j\epsilon C} 1_{j=y(x)} \log(f_j(x)) \tag{3}$$

$$l_{out}(f(x')) := -\sum_{j\epsilon C} \frac{1}{q} \log(f_j(x')) \tag{4}$$

In these equations, q is the number of classes and $f(x)\epsilon(0,1)^q$ is the softmax probability of a class after the input was processed. The term $1_{j=y(x)}$ is only one for the ground truth class and otherwise it is zero. The factor λ controls the weighting between entropy maximization and image segmentation. The softmax entropy resembles the amount of uncertainty of the model prediction can be calculated according to the following formula [3]:

$$E(f(x)) = -\sum_{j\epsilon C} f_j(x) \log(f_j(x)) \tag{5}$$

In equation (2) the principle behind multitask learning of entropy maximization of out-of-distribution and semantic segmentation of in-distribution objects is revealed. By designing the loss function as the sum of two entropies and weighting them, multitask learning is achieved. The first term adds the softmax entropy of the ground truth class to the loss function, i.e. the higher the softmax probability of the correct class, the less the addition to the loss is. Thereby, by minimizing the loss function, the model is encouraged to assign high scores to correct classes and learns image segmentation. The second term is responsible for the entropy maximization by penalizing the model for assigning high scores to out-of-distribution objects. This is done by adding the normalized entropy for every class to the loss and thereby encouraging the model to assign low scores for all classes to out-of-distribution objects. This way, the softmax entropy of those objects is increased. The application of the modified loss function is realized through encoding the targets as specified above and using manual one-hot-encoding for the cross-entropy loss. The factor λ is responsible for weighting both task. For the training of the DeepLabV3+ we followed the researchers recommendations of $\lambda = 0.9$, the use of Adam optimizer, a learning rate of 10^{-5} and training the DeepLabV3+ model for 4 epochs. For a UNet there are no specific recommendations to choose the parameters, which is why our initial guess was to use the same parameters as for the DeepLabV3+ model. The difference in the training of the two models lies in the number of epochs and in the resolution of the images and labels. We soon realized that the UNet was not able to work with high resolution pictures (1024x2048) like the DeepLabV3+ which is why we scaled down the resolution of the training data to (256x512) for the UNet.

We also evaluate the UNet on images with that resolution. Then, the UNet will be trained for a considerable amount of time (76 epochs) longer than the DeepLabV3+. This is because the UNet is not as capable as the DeepLabV3+ and therefore needs more time to learn the tasks. In the Chan et al. paper, this was already recognized when the researchers tried out a less powerful model.

After the training for entropy maximization and semantic segmentation is done, a threshold is introduced, that thresholds the resulting softmax entropy of the output image. Any pixel above that threshold is considered and any pixel below that threshold is considered non-anomalous. The code for our project is attached.

4 Results

To analyze the results, first the results of the UNet will be analyzed and then they will be compared to the DeepLabV3+. In table 1 the metrics for both models before and after entropy maximization training are given.

Metric	DeepLabV3+ untrained	DeepLabV3+ trained	UNet untrained	UNet trained
AUROC	0.9443	0.9764	0.7835	0.8774
FPR95	0.1896	0.1218	1.0	0.3853
AUPRC	0.2751	0.6566	0.03972	0.1906
mIoU	-	-	0.6553	0.4885

Table 1: Metrics for the UNet and the DeepLabV3+ before and after entropy maximization training.

The results show a major improvement of the ability of the UNet to detect anomalies, which comes with the penalty of a worse performance in semantic segmentation. These results are in line with the findings of Chan et al. and show, that entropy maximization can be successfully applied to a UNet architecture. In particular, the AUROC of the UNet rose by 11.98%, which means that the separability of anomalies and non-anomalies in the output of the UNet is increased. The FPR95 sank by 61.47%, which means that when the model has a 95% TPR, the UNet now only classifies 38.53% of non-anomalies as anomalies. Before it was 100% which implies that the decrease is a significant improvement of the model. The AUPRC was very low in the untrained model, but after training it increased by 379.86% which is another significant increase in performance, as the AUPRC resembles the ability of the model to detect anomalies while keeping false positive seldom. This is especially crucial in the field of autonomous driving, because false positives can have dangerous consequences. While all the metrics for anomaly

detection showed a better performance of the UNet, the mIoU sank by 25.45% which shows a decrease in the perfomance of the UNet in the original task of semantic segmentation. This result was also observed in the research by Chan et al. which is why they adapted their training parameters and the number of epochs so that the mIoU is retained at a high level. Because our main goal was the anomaly detection, increasing the mIoU is a topic for future work.

Similar improvements were recorded for the DeepLabV3+ which is also in line with the findings of Chan et al. The AUROC increased by 3.40%, the FPR95 decreased by 35.76% and the AUPRC increased by 138.68%. Unfortunately we were not able to calculate the mIoU scores for the trained and untrained DeepLabV3+ with high resolution images. The mIoU could be calculated on low resolution images, but for high resolution images the computational costs were too high, which is why our computing cluster (Unity Cluster from the University of Massachussets Amherst [20]) ran out of memory frequently when trying to get the mIoU. Because the DeepLabV3+ was trained on high resolution pictures, applying it to low resolution pictures actually worsened the results. But accroding to the paper from Chan et al. [3] and our own findings from the UNet, a slight decrease in semantic segmentation performance of the DeepLabV3+ can be recorded. This decrease is by far not as large as the decrease in the UNet, it is only some percents less accurate.

What is noticeable is the fact that the performance increase in anomaly detection as well as the performance decrease in semantic segmentation was significantly higher in the UNet than the DeepLabV3+. Reasons for that will be given in the conclusion section. To better understand the results, consider figure 1, which compares the semantic segmentation, softmax entropy and anomaly detection for a threshold of 0.7 on the softmax entropy of a semantic segmentation UNet (UNet untrained) and a semantic segmentation with anomaly detection UNet. The image is a sample from the Fishyscapes dataset and the decrease in performance in the segmentation is clearly visible, in particular the amount of details in the segmentation is not as much as before entropy maximization training. On the other hand, a significant increase of the softmax entropy on the anomaly (the cat) can be observed. If the softmax entropy gets thresholded at 0.7 then every pixel with a softmax entropy of over 0.7 is considered an anomaly and is marked with the red color. Before entropy maximization, there were no anomalies detected, which is a false negative. After entropy maximization, the UNet recognizes the cat as an anomaly and a few spots in the background which are false positives. But in other images which can be found in the appendix, anomalies were not detected or non anomalous objects were detected as an anomaly and the threshold varies between 0.6 and 0.8. Generaly the observation can be made that significantly more anomlies get detected, but still there are some false positives and false

negatives. In figure 2 a sample from Cityscapes is depicted, as an example for a typical complex driving scene with no anomaly. In this case the delivered the wanted result and did not detect an anomaly, only few false positives in the background. In the appendix more predictions from the UNet are given where anomalies are detected, partially detected or not detected.

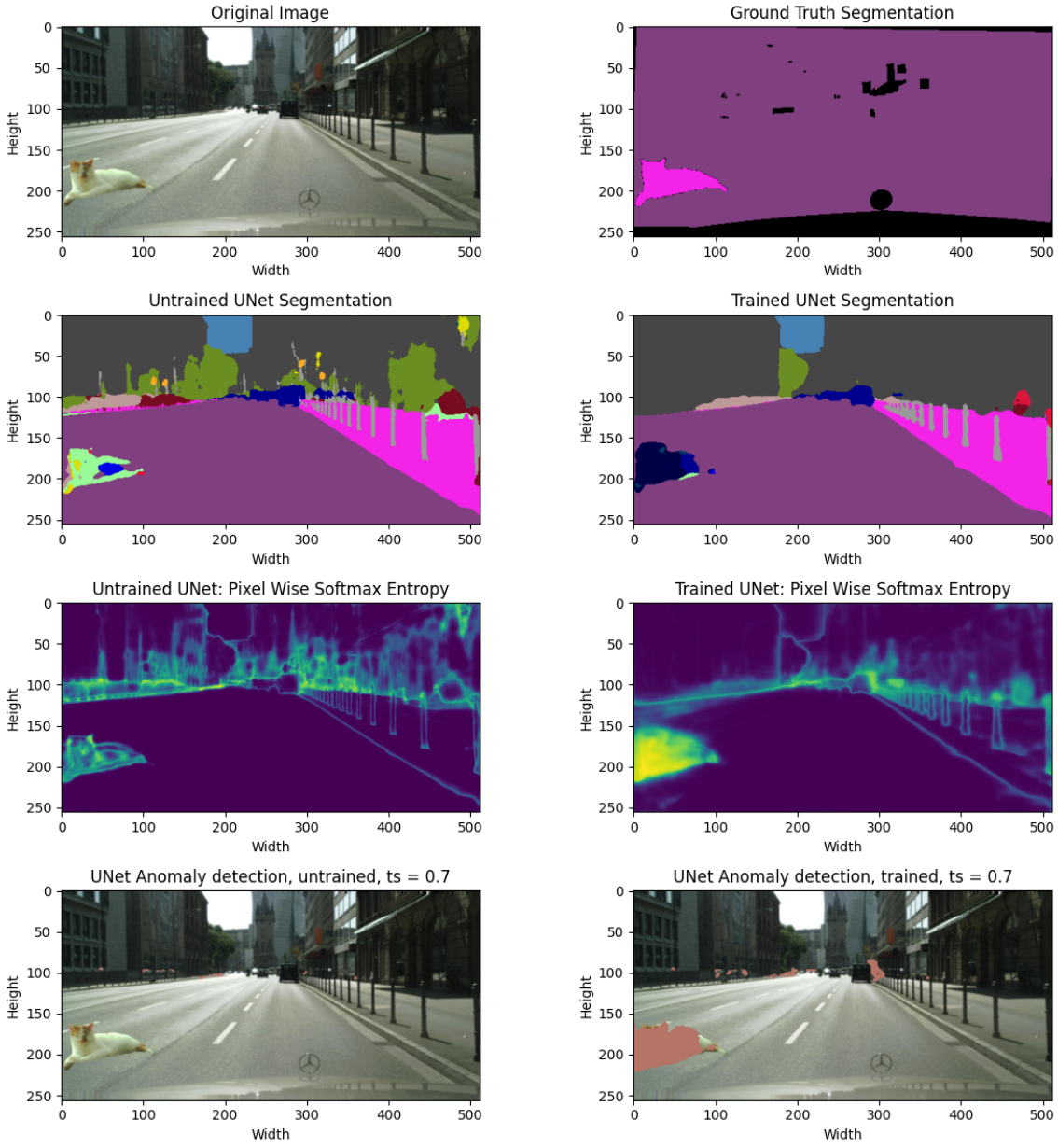

Figure 1: Fishyscapes image processed by the UNet with ground truth, the semantic segmentation as well as the softmax entropy before and after entropy maximization training and the identified anomalies (marked in red) in both cases.

Looking at the results for the same image, but processed by the DeepLabV3+ in figure 3, a clear difference can be seen, as the DeepLabV3+ has far superior segmentation results with less performance decrease after entropy maximization training. In this case a threshold of 0.8 was chosen in contrast to the threshold of 0.7 like in the UNet, which is also recommended by

Chan et al. The anomaly got detected well and the number of false positives seems to be lower than in the UNet. Overall the DeepLabV3+ has a better performance than the UNet, which was expected.

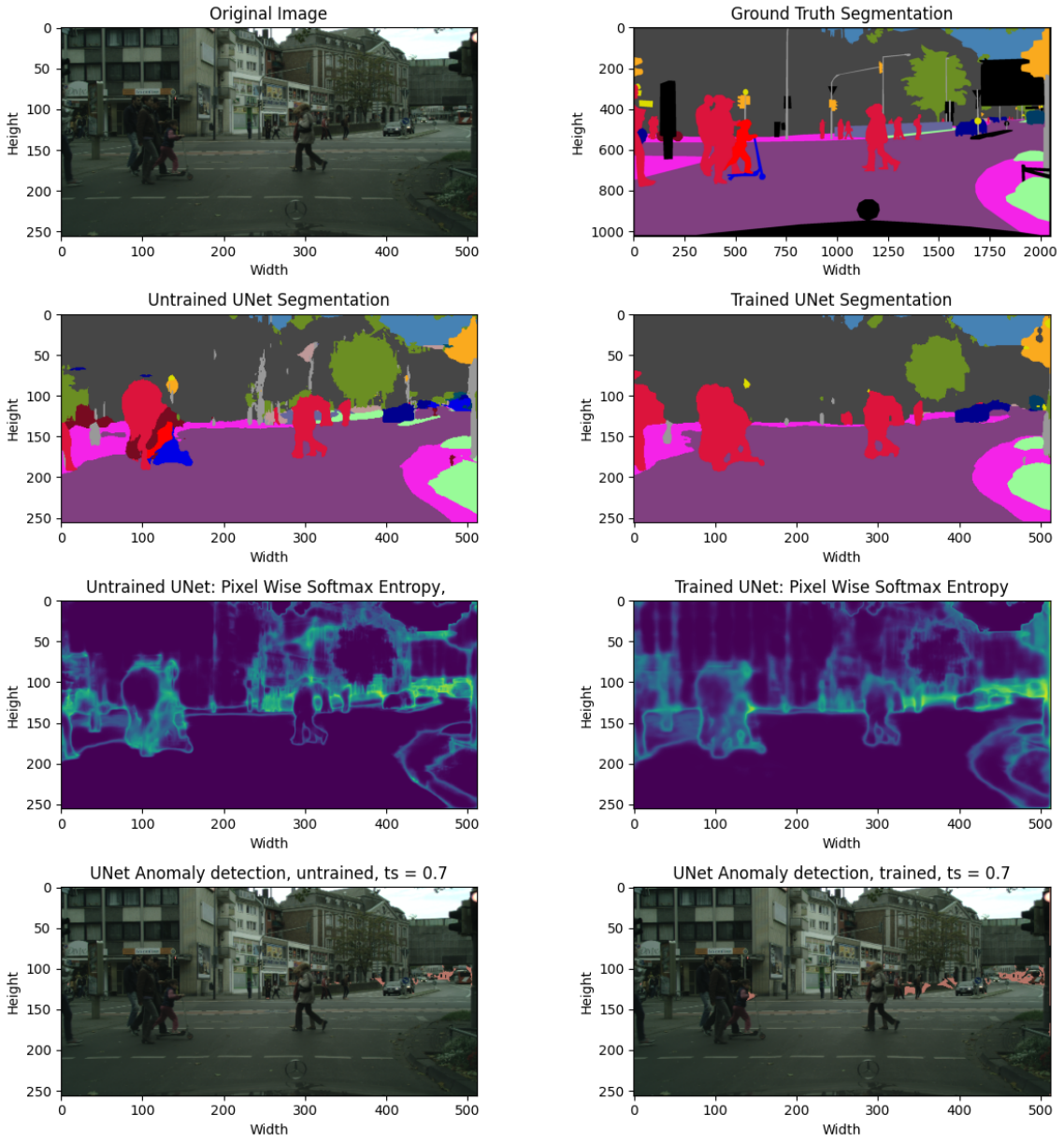

Figure 2: Cityscapes image processed by the UNet with ground truth, the semantic segmentation as well as the softmax entropy before and after entropy maximization training and the identified anomalies (marked in red) in both cases.

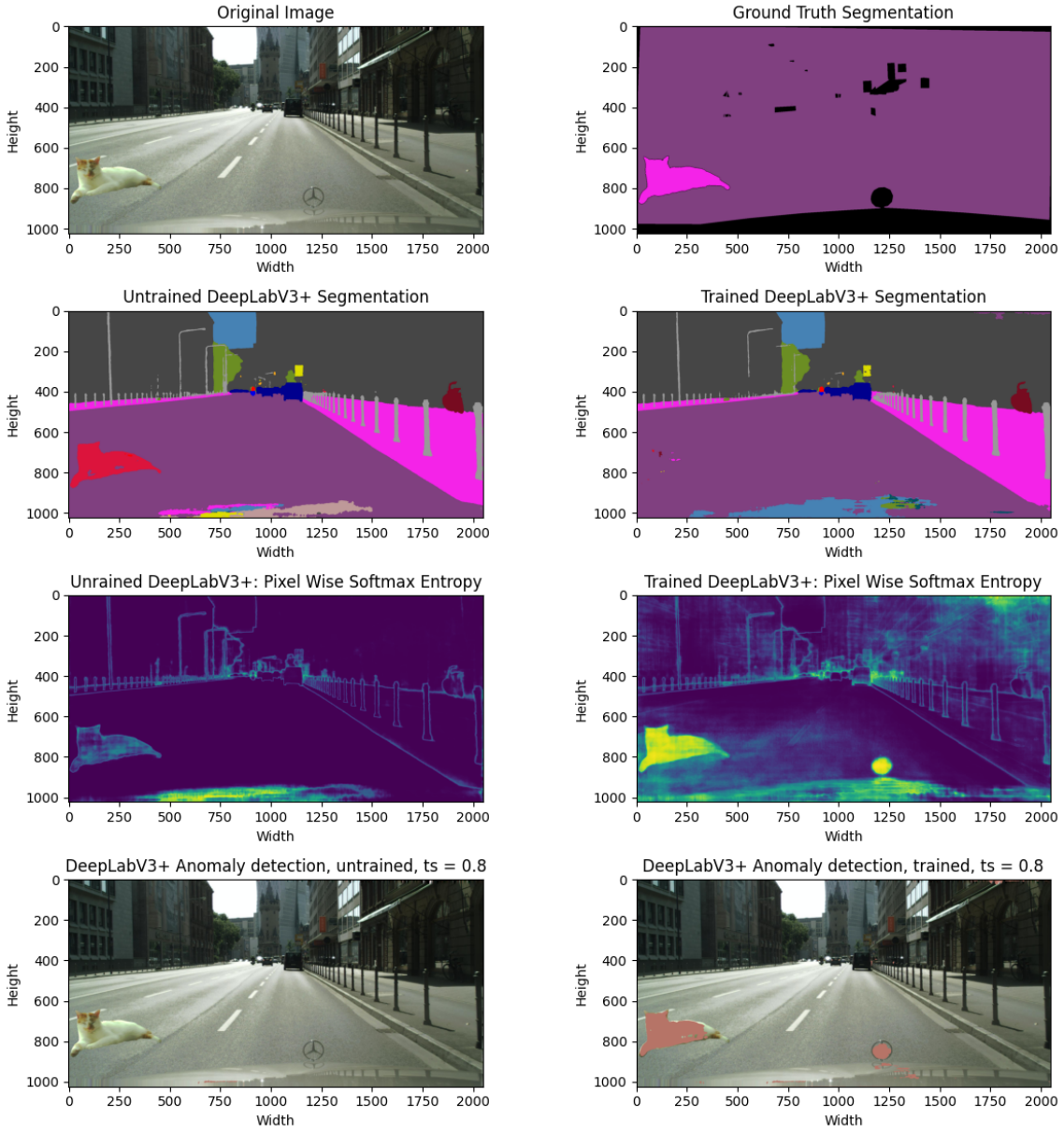

Figure 3: Fishyscapes image processed by DeepLabV3+ with ground truth, the semantic segmentation as well as the softmax entropy before and after entropy maximization training and the identified anomalies (marked in red) in both cases.

5 Conclusion

The initial research goal of applying entropy maximization on a lightweight UNet architecture was successful and it can be shown, that the UNet's anomaly detection performance can be significantly increased by training the UNet for entropy maximization. This comes at the cost of a considerable decrease in semantic segmentation performance. Another observation is, that lightweight models like the UNet benefit more from entropy maximization than stronger models, like the DeepLabV3+ when looking at certain metrics. On the other hand, state of the art models like the DeepLabV3+ also benefit from entropy maximization as it also enables them to detect anomalies. It also had less decrease of semantic segmentation performance, than the UNet, while recognizing the anomaly as well as the UNet. To improve an already strong model is harder than to improve a weaker one. This was expected, but the important takeaway is, that even light models can be trained to detect anomalies and these less computationally expensive models can be applied in scenarios with limited computing power. In this research, this effect was already visible when working with both models. The calculations on the UNet were faster, which shines an optimistic light on the possibility of a real life application of anomaly detection in UNet's.

6 Discussion

One of the first points to discuss is the decrease of semantic segmentation performance in the UNet after entropy maximization training. The decrease was a lot higher than in the DeepLabV3+ which can have several reasons. One reason can be that the researchers in the paper from Chan et al. had more time and more computational power at hand to analyze how to optimally set the training parameters, which we did not have time to do. Another reason can be that we trained the UNet for too long on entropy maximization, which results in the UNet being overfitted on anomaly data. Also, in this case the timespan for this project was too small to test out different approaches.

Another aspect worth discussing is the presence of false positives in the anomaly detection. This problem was already discussed and improved by Chan et al. using a meta classification stage. The scope of this project was too small to realize this too, but by mentioning it we want to show that there are already solutions for this problem. How well the meta classification works could not be verified by us.

The third point which is discussed is the datasets used for training. The Cityscapes dataset contains many pictures of complex urban driving scenes. But these were all taken from the hood

of the car facing the street from the same perspective for all images. Also, the location of all images is Germany, which is why we wanted to try and test the models on our own pictures. After taking the pictures and processing them by the DeepLabV3+ with anomaly detection, it was quickly realized that the model is very uncertain for pictures that are not taken from the typical camera frame, even when only well known instances are in the picture (like persons, trees or cars). Also, elements like shade confused the model and upon further research, many Cityscapes images were taken in cloudy daylight with almost no shadows on the street. This reveals the question, if a more diverse dataset maybe more useful.

One aspect of the UNet is, that it can handle low resolution pictures well, but with high resolution pictures the semantic segmentation performances decreases by a large amount. Normally the UNet is used for smaller datasets in medical imaging, but a future research question can be to adapt the UNet to higher resolutions.

The last point to discuss is the missing information on the mIoU for the DeepLabV3+. In our research, we frequently ran into problems with not enough memory being available for computing, even though GPU's and Cuda were used. Fortunately, the mIoU calculation for the UNet worked and in Chan et al. the mIoU of the DeepLabV3+ was already researched.

7 Future Work

An idea to continue working on this research is the implementation of the meta classification as a second step after entropy maximization on the DeepLabV3+ and the UNet. This will improve the anomaly detection performance. Also the approach presented by Chan et al. can be combined with other techniques like bayesian neural networks to achieve similar goals [3].

We evaluated our models on the Fishyscapes datset. There are other datasets like the Lost&Found dataset [21] which also consists of road images with hazardous objects. Here we could verify our results on a different benchmark. Also in this dataset the anomalies are not PascalVOC pictures added to Cityscapes images like in Fishyscapes, but the images of Lost&Found contain the anomalies for example as boxes on the road which were actually there when the image was taken.

The training for entropy maximization on the UNet showed a strong decrease in semantic segmentation performance, which is why in the future the parameters of the training can be fine tuned in order to prevent this problem. In this project, the timeframe for such an endeavor was too small, as the training of the models take a long time.

In the related works section, other approaches to anomaly detection were presented. After

testing entropy maximization on the UNet one could think of applying these state-of-the-art approaches to the UNet and document the results. We also tested the DeepLabV3+ model with our own created images. A small dataset of self created driving scenes could help to understand if the application of the anomaly detection software works in real life scenarios.

References

[1] A. Voulodimos, N. Doulamis, A. Doulamis, E. Protopapadakis, *et al.*, "Deep learning for computer vision: A brief review", *Computational intelligence and neuroscience*, vol. 2018, 2018. DOI: 10.1155/2018/7068349.

[2] V. Wiley and T. Lucas, "Computer vision and image processing: A paper review", *International Journal of Artificial Intelligence Research*, vol. 2, no. 1, pp. 29–36, 2018. DOI: 10.29099/ijair.v2i1.42.

[3] R. Chan, M. Rottmann, and H. Gottschalk, "Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation", in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 5128–5137. DOI: 10.48550/arXiv.2012.06575.

[4] G. Di Biase, H. Blum, R. Siegwart, and C. Cadena, "Pixel-wise anomaly detection in complex driving scenes", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 16 918–16 927. DOI: 10.48550/arXiv.2103.05445.

[5] M. Grcić, P. Bevandić, and S. Šegvić, "Densehybrid: Hybrid anomaly detection for dense open-set recognition", in *Computer Vision – ECCV 2022*, Oct. 2022, pp. 500–517. DOI: 10.48550/arXiv.2207.02606.

[6] Y. Tian, L. Yuyuan, G. Pang, F. Liu, Y. Chen, and G. Carneiro, "Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes", in *Computer Vision – ECCV 2022*, Oct. 2022, pp. 246–263. DOI: 10.48550/arXiv.2111.12264.

[7] S. Zagoruyko and N. Komodakis, "Wide residual networks", *CoRR*, vol. abs/1605.07146, 2016. arXiv: 1605.07146. [Online]. Available: http://arxiv.org/abs/1605.07146.

[8] O. Ronnenberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation", in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Nov. 2015, pp. 234–241. DOI: 10.48550/arXiv.1505.04597.

[9] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation", *CoRR*, vol. abs/1802.02611, 2018. [Online]. Available: http://arxiv.org/abs/1802.02611.

[10] J. Henriksson, C. Berger, M. Borg, L. Tornberg, S. R. Sathyamoorthy, and C. Englund, "Performance analysis of out-of-distribution detection on trained neural networks", *Information and Software Technology*, vol. 130, p. 106 409, 2021, ISSN: 0950-5849. DOI: https://doi.org/10.1016/j.infsof.2020.106409. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950584919302204.

[11] M. Everingham, S. M. A. Eslami, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zissermann, "He pascal visual object classes challenge: A retrospective", *International Journal of Computer Vision*, vol. 111, pp. 98–136, 2015, ISSN: 1573-1405. DOI: 10.1007/s11263-014-0733-5. [Online]. Available: https://doi.org/10.1007/s11263-014-0733-5.

[12] M. Cordts, M. Omran, S. Ramos, *et al.*, "The cityscapes dataset for semantic urban scene understanding", *CoRR*, vol. abs/1604.01685, 2016. arXiv: 1604.01685. [Online]. Available: http://arxiv.org/abs/1604.01685.

[13] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, "Microsoft coco: Common objects in context", in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Cham: Springer International Publishing, 2014, pp. 740–755, ISBN: 978-3-319-10602-1. [Online]. Available: http://arxiv.org/abs/1405.0312.

[14] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge", *International Journal of Computer Vision*, vol. 88, pp. 303–338, Nov. 2010. DOI: 10.1007/s11263-009-0275-4.

[15] H. Blum, P.-E. Sarlin, J. Nieto, R. Siegwart, and C. Cadena, "The fishyscapes benchmark: Measuring blind spots in semantic segmentation", *International Journal of Computer Vision*, vol. 129, pp. 1–17, Nov. 2021. DOI: 10.1007/s11263-021-01511-6.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization", in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA,*

USA, May 7-9, 2015, Conference Track Proceedings, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980.

[17] *Pytorch*, Last accessed on 12/19/2023, 2017. [Online]. Available: https://pytorch.org/vision/stable/models.html.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation", in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., Cham: Springer International Publishing, 2015, pp. 234–241, ISBN: 978-3-319-24574-4. [Online]. Available: http://arxiv.org/abs/1505.04597.

[19] *Robin chan meta-ood*, Last accessed on 12/20/2023, 2021. [Online]. Available: https://github.com/robin-chan/meta-ood.

[20] *Unity, umass amherst*, Last accessed on 12/20/2023, 2023. [Online]. Available: https://unity.rc.umass.edu/.

[21] P. Pinggera, S. Ramos, S. Gehrig, U. Franke, C. Rother, and R. Mester, "Lost and found: Detecting small road hazards for self-driving vehicles", in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 1099–1106. DOI: 10.1109/IROS.2016.7759186.

Appendix

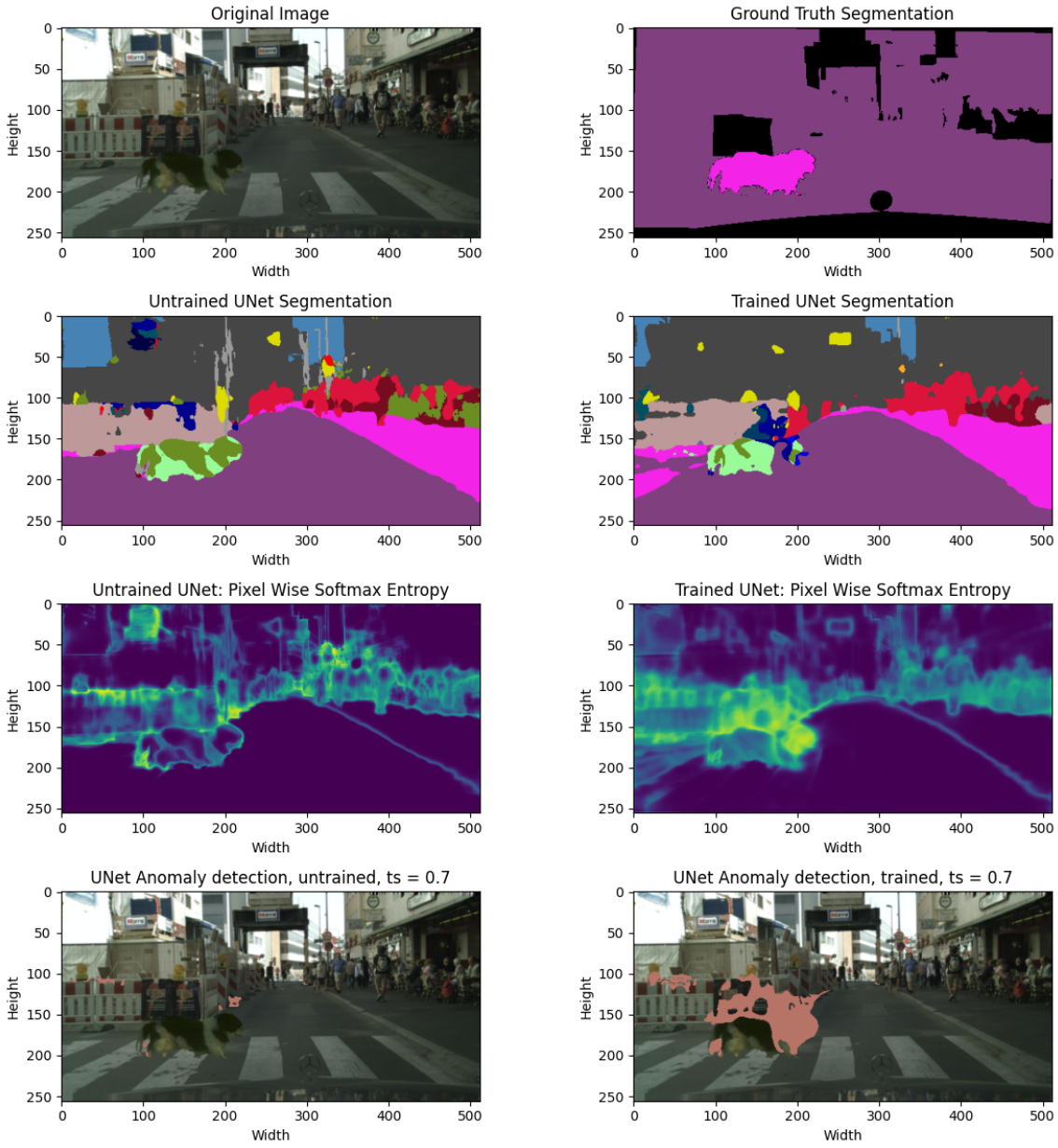

Figure 4: Fishyscapes image processed by the UNet with ground truth, the semantic segmentation as well as the softmax entropy before and after entropy maximization training and the identified anomalies (marked in red) in both cases. In this image the anomaly got identified.

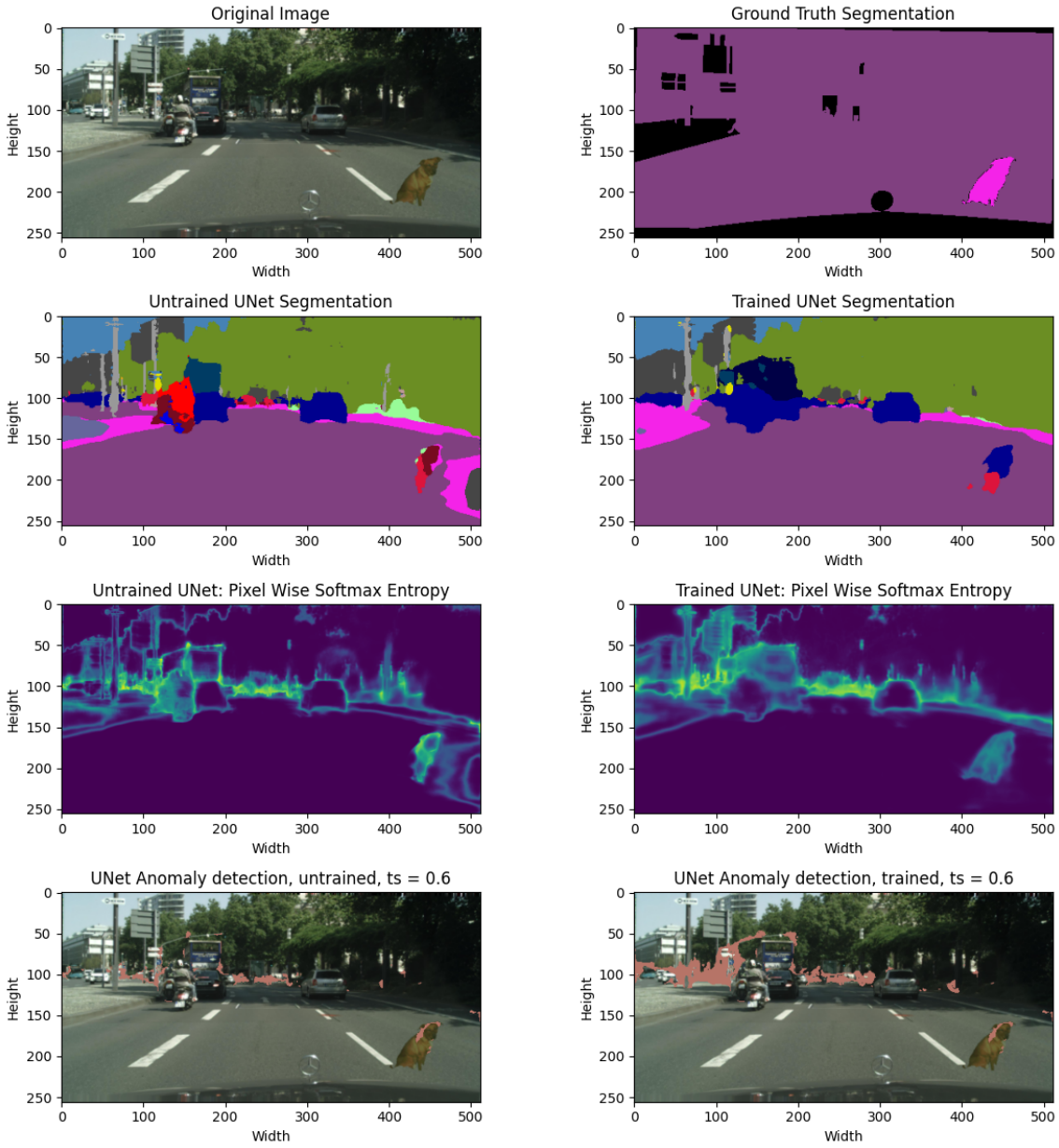

Figure 5: Fishyscapes image processed by the UNet with ground truth, the semantic segmentation as well as the softmax entropy before and after entropy maximization training and the identified anomalies (marked in red) in both cases. The anomaly was not identified.

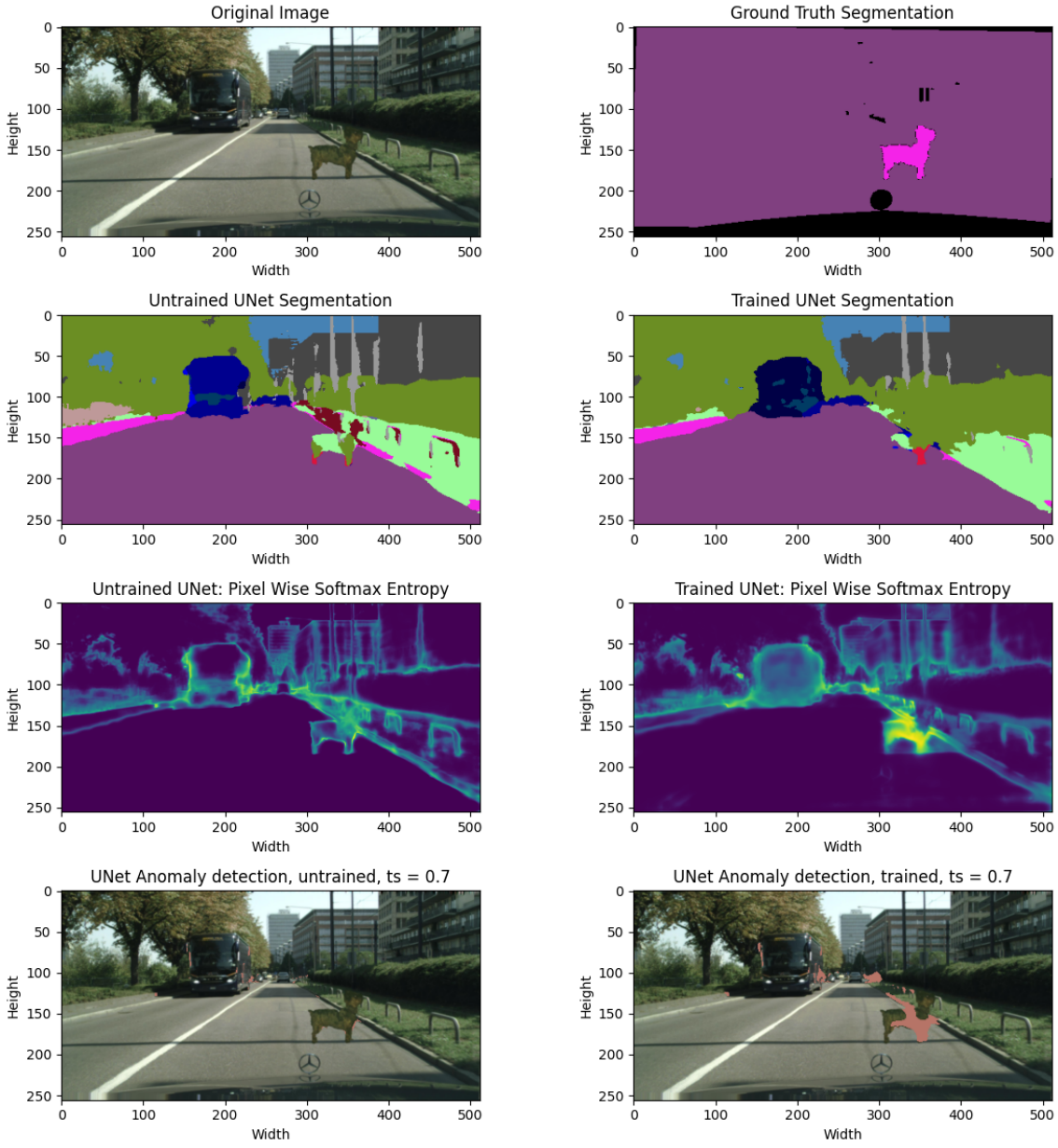

Figure 6: Fishyscapes image processed by the UNet with ground truth, the semantic segmentation as well as the softmax entropy before and after entropy maximization training and the identified anomalies (marked in red) in both cases. The anomaly was identified.

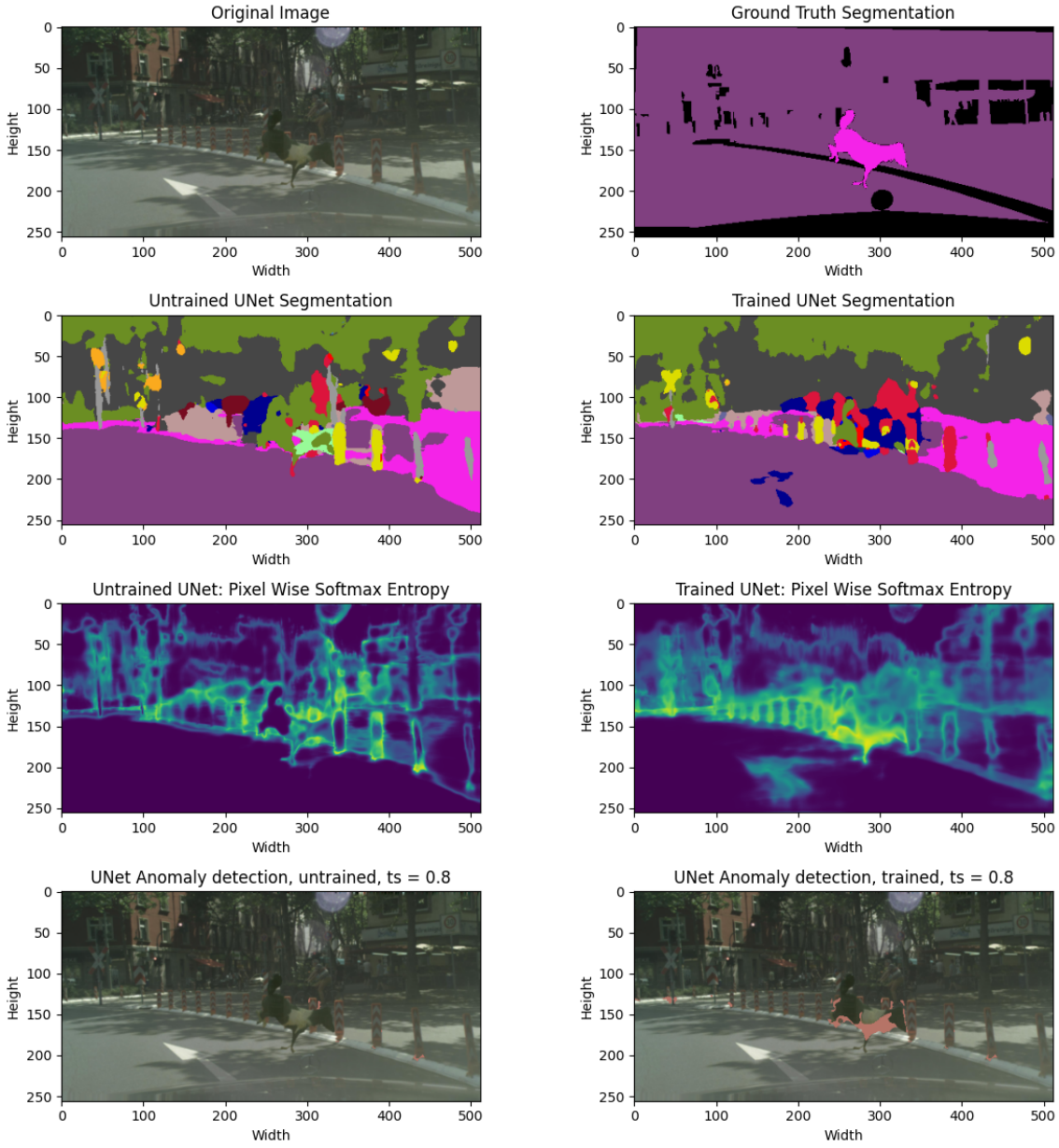

Figure 7: Fishyscapes image processed by the UNet with ground truth, the semantic segmentation as well as the softmax entropy before and after entropy maximization training and the identified anomalies (marked in red) in both cases. The anomaly was identified partially.

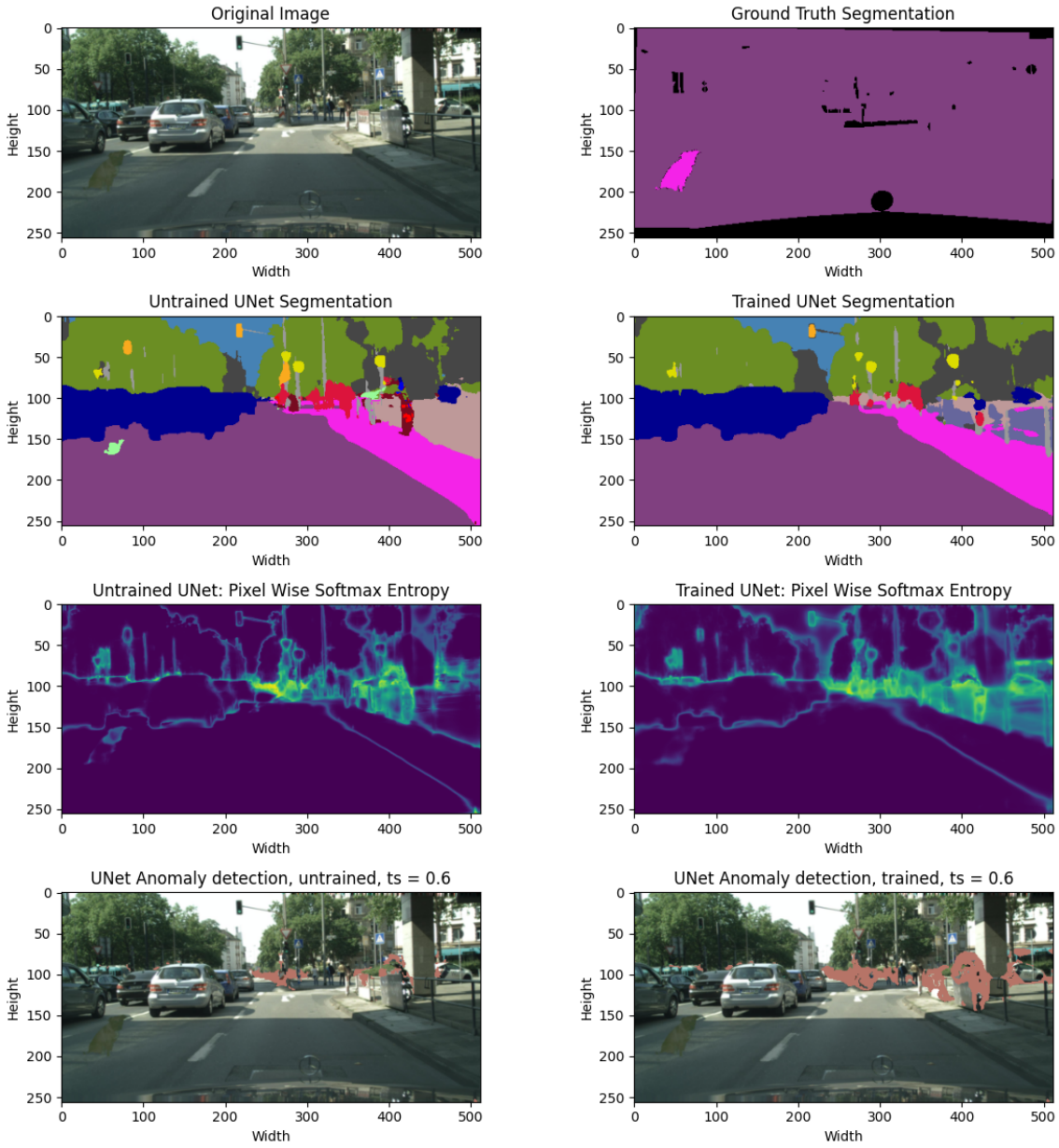

Figure 8: Fishyscapes image processed by the UNet with ground truth, the semantic segmentation as well as the softmax entropy before and after entropy maximization training and the identified anomalies (marked in red) in both cases. The anomaly was not get detected.