Karlsruhe Institute of Technology

Camera-based Anomaly Detection with Generative World Models

Bachelor Thesis

Noël Ollick

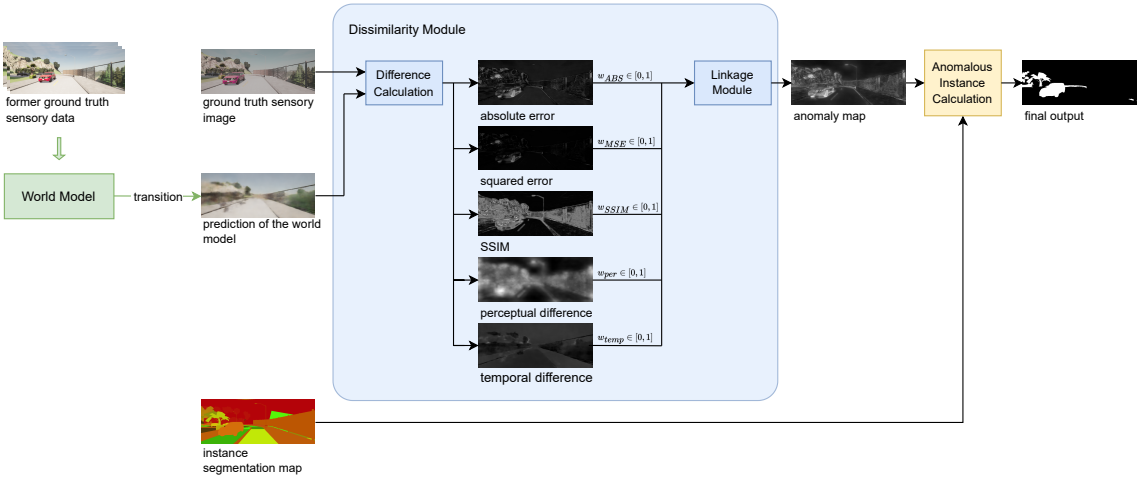
Department of Informatics
Institute of Applied Informatics and Formal Description Methods
and
FZI Research Center for Information Technology

Reviewer:	Prof. Dr. R. Reussner
Second reviewer:	Prof. Dr.–Ing. J. M. Zöllner
Advisor:	M.Sc. Daniel Bogdoll

Research Period: 01. December 2023 – 31. March 2024

Camera-based Anomaly Detection with Generative World Models

by
Noël Ollick



Bachelor Thesis
March 2024

FZI

Bachelor Thesis, FZI
Department of Informatics, 2024
Reviewers: Prof. Dr. R. Reussner, Prof. Dr.–Ing. J. M. Zöllner

Abteilung Technisch Kognitive Assistenzsysteme
FZI Research Center for Information Technology

Affirmation

Ich versichere wahrheitsgemäß, die Arbeit selbstständig angefertigt, alle benutzten Hilfsmittel vollständig und genau angegeben und alles kenntlich gemacht zu haben, was aus Arbeiten anderer unverändert oder mit Abänderungen entnommen wurde.

Karlsruhe, *Noël Ollick*
March 2024

Abstract

Although huge improvements in the field of autonomous driving have been made in recent years, dealing with unexpected situations remains a challenging task. Anomaly detection techniques aim to detect those unknown cases. While generative world models have shown promising results regarding the perception of the environment and predicting future driving conditions, they are rarely utilized in anomaly detection for autonomous driving. This thesis presents a novel anomaly detection method which leverages the advantages of world models and uses feature extraction, reconstructed observations, and predictions of future observations in order to detect corner cases in automated driving. The proposed anomaly detection model works fully unsupervised and does not require anomalies in training data.

Contents

1 Introduction

While recent advancements in automated driving led to more reliable systems in self-driving cars, dealing with abnormalities is still a challenging task [4]. In order to develop autonomous driving techniques which function dependably even in most complex driving scenarios, it is crucial for those systems to detect anomalous situations [5]. Overconfident systems could pose a major threat to road safety, for instance by causing a traffic accident due to their failure to accurately detect an animal crossing the street.

Generative world models have shown promising results in perceiving even elaborate driving scenarios correctly in automated driving [16]. Hu et al. summarize the importance of world models for reliable autonomous driving system as follows: "World models represent a crucial step towards achieving autonomous systems that can understand, predict, and adapt to the complexities of the real world. Furthermore, by incorporating world models into driving models, we can enable them to better understand their own decisions and ultimately generalize to more real-world situations." [18] In embodied artificial intelligence architectures which are for instance used in autonomous vehicles, normality is not solely specified by data, but also by taken actions. Despite the capability of world models to predict future states based on taken action and to generate reconstructions in the observation space [4], most anomaly detection methods do not deploy world models for corner case detection [4]. In this thesis, an anomaly detection model for autonomous driving which takes advantage of world models is presented. Furthermore, a novel approach for anomaly detection which combines predictions of future states, feature embedding, and reconstructions of observations in order to detect corner cases is introduced. The proposed anomaly detection technique works fully unsupervised and does not require labeled data. In addition, it does not re-train underlying models and thus works without anomalous training data.

The main research questions of this thesis can be summarized as follows:

i. How can the advantages of world models be leveraged in the context of anomaly detection?

ii. Which anomaly detection techniques are suitable for camera-based anomaly detection in autonomous driving systems which use generative world models?

The thesis is structured as follows: In chapter 2, I define the terms corner case and anomaly in the context of autonomous driving. Furthermore, I introduce world models and common image comparison metrics. In chapter 3, I introduce state-of-the-art world models and anomaly detection techniques in autonomous driving. Additionally, I present state-of-the-art image segmentation approaches, one of which is used by the proposed anomaly detection model in order to allocate individual instances. Chapter 4 then gives a detailed explanation of how the anomaly detection model works. Here, I explain, how predictions are generated by the world model and how the

world model's perception is compared to ground truth sensory image data with the intend to detect anomalies. Also, I elaborate how the image comparison metrics are calculated and how single instances are classified as anomalous. The evaluation and experimental results for different configurations of the anomaly detection model are shown in chapter 5. Finally, the main results and their possible effects on future work are summarized in chapter 6.

2 Background

2.1 World Models

While the term *world model* was initially introduced in the context of reinforcement learning [16], world models are nowadays used in numerous fields in computer science, such as in computer vision [16] and autonomous driving [4, 16, 18]. Bogdoll et al. define a world model as a model which "embeds sensory observations into a latent state, predicts action-conditioned state transitions, and is able to decode into observation space." [4]. Figure 2.1 depicts a world model during inference and its essential components: a *representation model*, an *observation model*, and a *prediction model* [4].

The *representation model* embeds an observation o_t, the last taken action a_{t-1}, and the former latent state s_{t-1} into a new state s_t [4]. Given this state s_t, a reconstructed observation $\hat{o}_t$ can be created using the *observation model* [4]. The *prediction model* allows predicting future states s_{t+n} given the current state s_t and action a_t [4] and is therefore also referred to as transition model [4]. The two denotations "prediction model" and "transition model" will be used interchangeably in this thesis.

By this specific structure, world models can create predictions of future observations $\hat{o}_{t+n}$ using the prediction model and the observation model in combination.

2.2 Anomalies in Autonomous Driving

Dealing with complex driving scenarios in autonomous driving remains a challenging task [8]. Recognizing abnormal situations incorrectly can lead to a catastrophic outcome [15], such as causing an accident because the system was not capable of detecting an unknown object on the street.

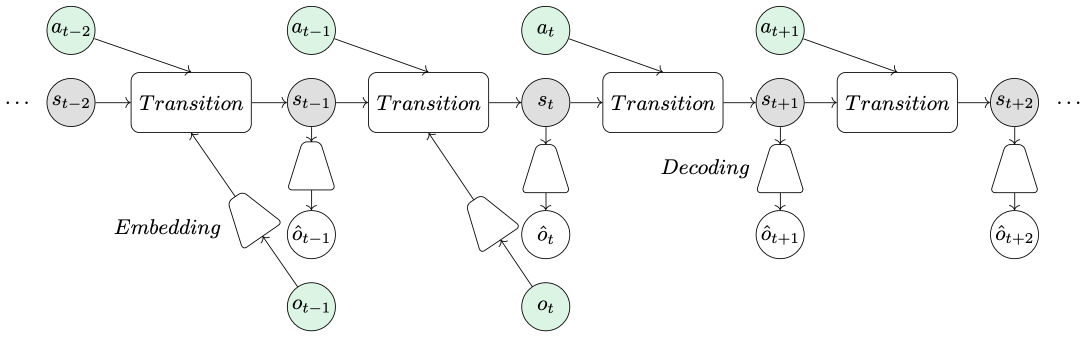


Figure 2.1: A world model during inference. (figure reprinted from Bogdoll et al. [4])

Heidecker et al. describe corner cases and anomalies as follows: “For corner cases in automated driving, there is a deviation from normality that is manifested in non-conform behavior or patterns. The terms anomaly and corner case are almost used synonymously. Anomalies describe a deviation from normality.” [15] The terms corner case and anomaly are sometimes differentiated in literature and multiple definitions for an anomaly can be found [15]. For the sake of readability, I define both terms as applicable when there is an abnormality in a driving scenario and therefore use the terms corner case and anomaly interchangeably in this thesis.

For categorizing corner cases, Heidecker et al. [15] have extended the categorization of Breitenstein et al. [7] into corner cases in the sensor layer, the content layer, and the temporal layer. The sensor layer contains anomalies regarding hardware or the physical state of a sensor, such as sensor failure or using a soiled camera. The content layer contains corner cases regarding the domain, objects or contextual abnormalities in a single frame. Examples for such anomalies are unknown traffic signs, abnormal objects on the street or recognizing known objects in posters. The third layer is the temporal layer which contains scenery anomalies which become present when considering multiple frames. Such an anomaly is for example a driver breaking a traffic rule.

2.3 Metrics for Image Comparison

Camera-based anomaly detection techniques which reconstruct an input image in order to detect abnormalities often use image comparison metrics to find discrepancies between the input image and the reconstruction [5, 13, 24]. In the following, I will introduce common image comparison metrics which are utilized in the proposed anomaly detection model. In chapter 4.3, I will then give a more detailed explanation of how they are calculated.

The underlying world model MUVO of the proposed anomaly detection model uses the common L1 loss as training loss for RGB images [6]. The L1 loss uses the absolute error between the input image and the target image. Since the observation model of the world model was trained with this loss [6], I use the absolute error as one of the metrics for comparing images in the anomaly detection method.

As pointed out by Wang and Bovik [27], the mean squared error (MSE), sometimes also referred to as L2 loss when used as training loss, is a very popular comparison metric in signal processing. Similar to the absolute error, the MSE compares to signals element-wise. However, it squares the difference of those elements instead of using its absolute value. Due to squaring the difference, the MSE weights high outliers more heavily than small discrepancies in comparison to the absolute error. The MSE is also used as the loss function for point clouds in MUVO [6].

While the MSE is very prominent in signal processing, it sometimes exhibits deficiencies in performance when comparing images regarding their perceptual fidelity [27]. For instance, differently distorted images can have a similar MSE value when compared to the original image although they look very different in the eyes of humans [27]. To address this issue, Wang et al. [28] introduced the Structural Similarity Index (SSIM) which compares two images based on their spatial structure, e. g. their texture, rather than their pixel-wise differences. The SSIM achieves this by taking into account that proximate pixels are more likely to be dependent on each other

than pixels with a large spacial distance [24]. The SSIM is also used as a metric in the anomaly detection technique for autonomous driving by Vojir et al. [24].

The perceptual difference leverages the ability of pre-trained VGG networks to extract complex features from images [13]. In contrast to the comparison of low level features such as pixel-wise RGB data, this metric allows the model to evaluate the similarity of two images based on their content and spatial discrepancy [13]. The perceptual difference is also used as a comparison metric by the anomaly detection method of Di Biase et al. [13].

3 State of the Art

3.1 World Models for Autonomous Driving

World models have shown promising results in perceiving even complex driving scenarios and predict future driving conditions based on prior latent states and taken action [17, 18, 29]. In the following, I will present state-of-the-art world models for automated driving systems.

Hu et al. have developed MILE which is a "Model-based Imitation LEarning approach" [17]. It contains an inference model, which is capable of mapping observations and actions into latent states and a generative model. The generative model is able to predict future states and has a bird's-eye view decoder.

GAIA-1 is another world model which is used in the context of autonomous driving [18]. In contrast to MILE, GAIA-1 also contains a high resolution video decoder [18]. GAIA-1 takes video, information about the action, and text as input and is capable of decoding its predictions into video frames [18].

Wang et al. [26] have developed a world model with a video decoder called DriveDreamer. Similar to GAIA-1, DriveDreamer was trained on real-world data. DriveDreamer uses a two-stage learning pipeline, where structural traffic conditions are learned before video frame predictions.

Zhang et al. [29] have developed a world model which utilizes point clouds. Their method first uses a tokenizer to encode sensory data into bird's-eye view tokens. The world model then operates on those generated tokes through discrete diffusion.

MUVO [6] is a multimodal world model with a 3D occupancy voxel representation of the environment. It takes images, lidar data, and data of the taken action as input, learns a voxel representation of the world, embeds the fused sensor data into a latent state, and then predicts future states using a transition model. Those predicted states can then be converted into high resolution RGB representations, occupancy grids, and point clouds. MUVO is the underlying world model for the anomaly detection method proposed in this thesis.

While world models are used in autonomous driving to model the environment, predict future states, and generate observations given states and actions, they are rarely used to detect anomalous data which the world model might not be capable of perceiving correctly in the first place [4]. World models can predict action-conditioned future latent states [4] and therefore do not solely consider observed sensory data, but also taken actions. For embodied agents such as autonomous driving vehicles, both action and sensory data are relevant to define normal driving behavior. World models thus demonstrate significant potential for anomaly detection for automated driving [4], despite being utilized rarely in prior corner case detection techniques.

3.2 Camera-based Anomaly Detection for Autonomous Driving

Breitenstein et al. have categorized methods for anomaly detection in different approaches: "reconstruction, prediction, generative, confidence scores, and feature extraction" [8]. *Reconstruction* and *generative* approaches make use of the reconstruction error of autoencoders. Those approaches are common for camera based anomaly detection methods for autonomous driving [5]. Di Biase et al. [13] developed a reconstructive method [5] which uses re-synthesized images and segmentation uncertainty for pixel-wise anomaly detection. Their method requires anomalous training data [5], which is also referred to as auxiliary data [5]. Vojir et al. [24] use the features of a semantic segmentation network as input for a reconstruction module in order to recreate the initial input image. In contrast to Di Biase et al. [13], their anomaly detection method does not require auxiliary data [5]. *Predictive* approaches try to predict future frames and then compare them to the true frame. In recent work for camera based anomaly detection for autonomous driving, however, they are not very common [5]. *Feature-based methods* extract features from input data using neural networks. For example, Bai et al. [2] use feature extraction and a one-class Support Vector Machine (SVM) to detect anomalies in urban road scenes. Methods which utilize *confidence scores* are very common in anomaly detection for autonomous driving [5]. Maskomaly is for instance an anomaly detection method which uses post-processing on masked-based segmentation modules, while not requiring auxiliary data [1].

Most of the state-of-the-art anomaly detection techniques solely focus on sensory data and do not leverage action-conditioned predictions of future observations. In driving scenarios with fast moving cars, however, the taken action heavily impacts what is considered either normal or anomalous. Utilizing models which can understand the complexities of real world driving situation, such as world models [18], therefore might be beneficial for anomaly detection techniques.

Furthermore, world models inherently offer sub-models which provide similar functionality than some parts of prior anomaly detection models: State-of-the-art reconstructive anomaly detection models rely on modules which are trained to reconstruct into observation space. For instance, Vojir et al. [24] trained a reconstruction module which takes features from the semantic segmentation as input. Di Biase et al. [13] trained a synthesis network which takes the semantic segmentation map as input in order to generate a reconstruction of the observation. World models, on the other hand, contain an observation model which is inherently capable of reconstructing an observation and can be trained separately from the anomaly detection technique.

Similarly it is possible to predict future frames by using the observation model in combination with the prediction model. This is then comparable to the approach of predictive anomaly detection models.

The embedding into latent sates by utilizing the world model's representation model can furthermore be seen as a form of feature extraction [4]. World models thus inherently provide extracted features from sensory data which potentially could be used by feature-based anomaly detection techniques.

3.3 Unsupervised Image Segmentation

For partitioning an image into multiple regions, there are different categories of classes for those regions: Semantic segmentation aims to partition images into semantic classes [14]. A semantic segmentation approach on camera data for autonomous driving could for instance map classes such as “car”, “traffic sign” or “bus” to individual pixels in an observation. Such approaches, however, do not differentiate between instances which can be mapped to the same semantic class: In this example, the semantic segmentation method would give different cars on the road the same semantic label and would not differentiate between different vehicles. For anomaly detection in safety-critical systems such as autonomous driving cars, it is often not sufficient to only comprehend the semantic classes in order to correctly classify objects as anomalous. For instance, one car in the observation might be driven normally, while the driver of another car in the same observed image breaks a traffic rule. To address this, instance segmentation approaches classify into individual instances [14] instead of semantic classes. In this example, two different cars would be labeled differently with an instance segmentation model, however, one would also lose the classification into the semantic class “car” when using instance segmentation. Approaches which simultaneously output instance and semantic labels in one output format are called panoptic image segmentation approaches [14, 21]. Finally, amodal segmentation tasks do not solely annotate the visible parts of objects in an image, but also mark hidden parts of instances [30]. For example, if a pedestrian walks in front of a parked car, an amodal image segmentation approach also estimates which hidden pixels are part of the car.

While many image segmentation techniques require training data which is labeled by humans [21], unsupervised image segmentation models like CutLER [25] or U2Seg [21] aim to achieve similar results without requiring labeled data. Those models are therefore especially appealing for camera-based anomaly detection models which do not utilize auxiliary data, since labeled anomalous instances are not available in the training dataset [5].

CutLER [25] allows unsupervised object detection and image segmentation. This approach first generates multiple masks by utilizing a model called MaskCut. Then, CutLER trains a detector with drop lossing and finally utilizes a self-training approach. Wang et al. [25] have used the unlabeled ImageNet [12] dataset to train CutLER.

U2Seg [21] is another image segmentation approach which demonstrated even better experimental results than CutLER on unsupervised instance segmentation tasks. The distinctive feature of U2Seg is the capability of forming instance, semantic, and panoptic segmentation masks with panoptic segmentation being the combination of instance and semantic segmentation in one output format. The proposed anomaly detection model in this thesis utilizes the panoptic output of U2Seg to distinguish instances in observations. For this thesis, weights which are pre-trained on both the ImageNet [22] and the COCO [19] dataset are used. U2Seg also utilizes MaskCut from CutLER to generate pseudo instance mask. It then generates semantic masks and fuses instance masks and semantic masks. Finally, U2Seg uses self-learning to train a universal segmentation model with the objective to generate instance-level and semantic masks on unlabeled data.

4 Method

In the following, I describe the methodology of the proposed anomaly detection model in detail: First I explain how predictions for future observations are created using the world model. Those predictions are then compared to sensory image data using multiple metrics. I also elaborate how those metrics and temporal differences in reconstructions are calculated and, if desired, combined. Finally, I describe how the anomaly detection model classifies single instances as anomalous. A general overview of the model is illustrated in figure 4.1.

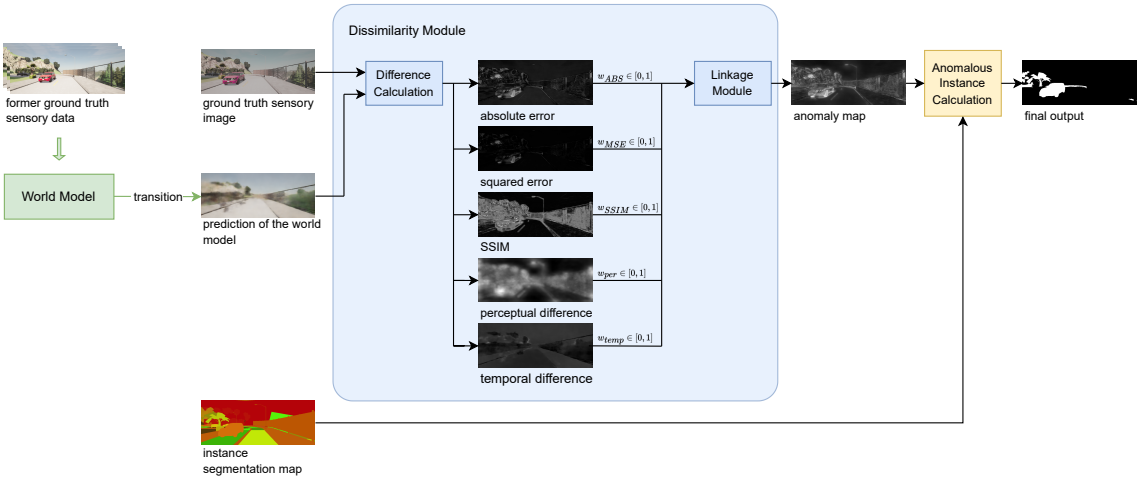


Figure 4.1: **An overview of the proposed anomaly detection method.** After generating the world model's prediction for the future observation, the model compares this reconstruction to the ground truth sensory image data using multiple error metrics. Those metrics can then be weighted for the linkage module. This allows the linkage module to either combine multiple metrics to an anomaly map or pass one single metric through. If desired, the model finally iterates through each instance in the observation and calculates its individual anomaly score using the anomaly map. This anomaly score can then be used to classify an instance as either normal or anomalous.

4.1 Generating Predictions

In order to detect anomalies, the proposed model compares the world model's perception of the environment to ground truth observations. An overview of the underlying world model MUVO [6] is given in Figure 4.2. The anomaly detection method first constructs an image which represents the world model's prediction of a future observation:

Initially, the world model fuses camera images and lidar point clouds and embeds the output into an one-dimensional vector o_t [6]. Together with a deterministic historical state h_t and an embedding of the last taken action a_t, MUVO generates a latent hidden state s_t [6]. This process can be seen as a form of feature extraction [4] where the state s_t embeds information about current

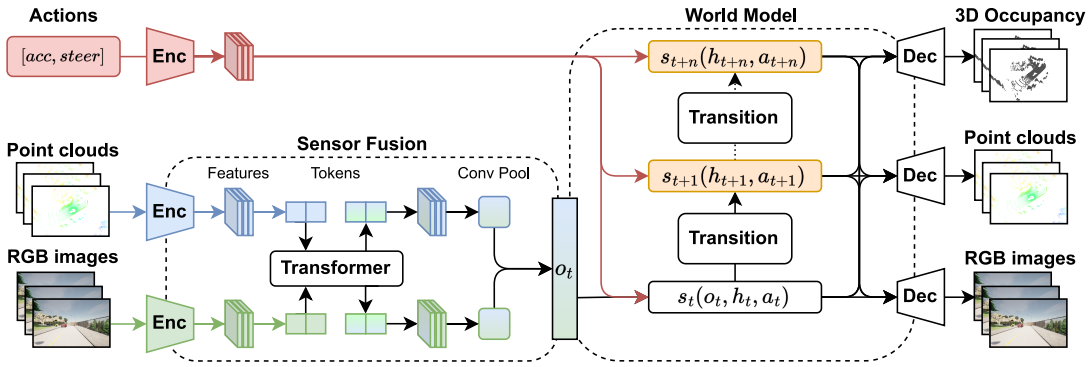


Figure 4.2: Overview of MUVO. (figure reprinted from Bogdoll et al. [6])

observations, taken actions, and the historical state of the environment.

The world model then determines the probability for the next prior hidden state $p(s_{t+1}|h_{t+1}, a_t)$ with its transition model [6]. The historical state $h_{t+1} = f_\theta(h_t, s_t)$ is calculated using the former deterministic historical state h_t and hidden state s_t [6]. MUVO uses a Gated Recurrent Unit (GRU) for modelling f_θ [6]. Note that s_{t+1} can be solely calculated by using the prior variables s_t, h_t, and a_t and is therefore a prediction for the next latent state of the world model. Finally, the observation model reconstructs a RGB image [6] by using the state s_{t+1}. In the following, this output image will be compared to the next ground truth sensor image in order to find discrepancies between the wold model's perception of the environment and the sensory image data.

4.1.1 Time Delay in Prediction

Instead of just predicting the next state, one can generate a batch of predictions for future states by using the transition model repetitively. By this, the world model does not only reveal its current perception of the environment, but how it predicts changes in the environment. Since the predictions are an additional disclosure of the world model's perception, the model further generates a prediction for an observation with a small time delay, stores it temporarily, and finally compares it to the sensory data once it is available. Figure 4.3 illustrates this process with an exemplary delay of two frames: Here, the preceding state for reconstruction is not the latest prior state, but rather one which emerges through two successive iterations in the transition model. In chapter 5, I assess the performance of the prediction for the next time step, as well as the delayed prediction.

4.2 Computer Vision Datasets

The proposed anomaly detection technique uses pre-trained models for calculating the perceptual difference and generating image segmentation maps. Di Biase et al. [13] use a network which is pre-trained on the ImageNet [12, 22] dataset in order to calculate the perceptual difference. For generating image segmentation maps, the proposed anomaly detection model utilizes the model U2Seg which was developed by Niu et al. [21]. The anomaly detection model uses weights for U2Seg which are pre-trained on the ImageNet [22] and the COCO [19] dataset. In the following,

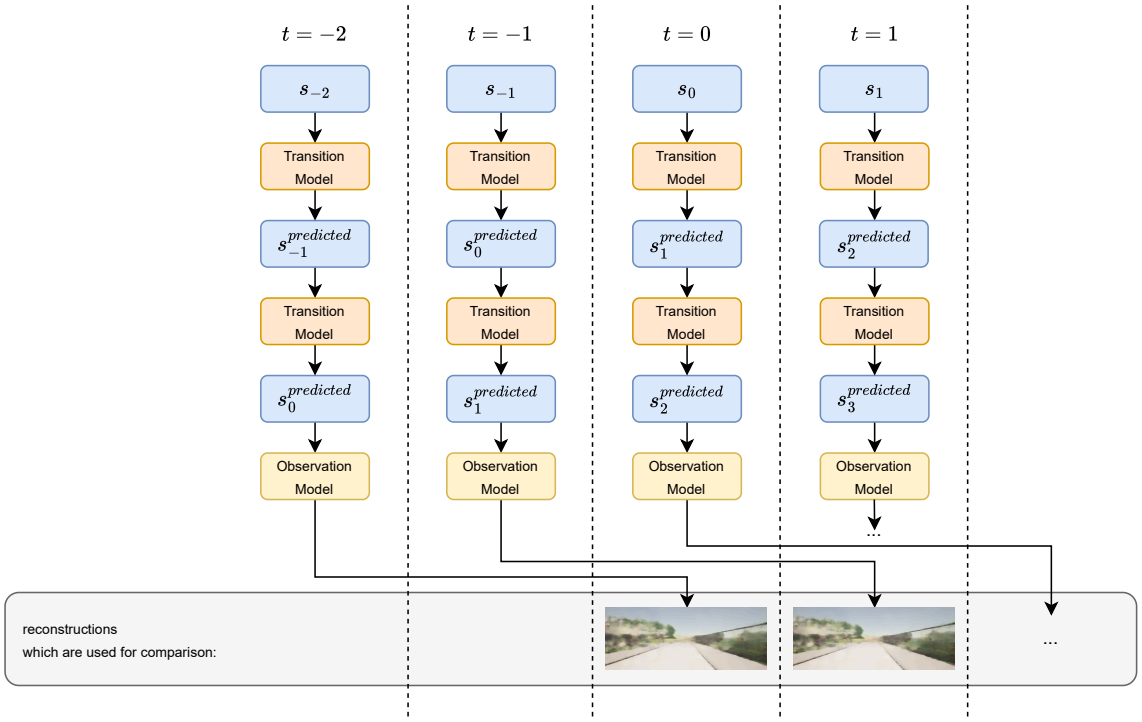


Figure 4.3: An overview of generating predictions with a time delay. Here, the delay is two frames.

I will illustrate the structure and content of both training datasets.

ImageNet [22] contains over one million annotated training images and has annotations about the object classes and also bounding boxes. The dataset is structured after the WordNet schema and was labeled by humans. ImageNet is not only largely scaled in respect to the amount of training images, but also regarding the amount of object classes: The image classification dataset contains 1000 classes.

With around 328,000 images, COCO [19] contains less images than ImageNet. In contrast to many prior image datasets however, COCO focuses on segmenting single object instances rather than just classifying images or finding bounding boxes. In total, COCO consists of 2.5 million labeled instances and 91 object classes.

4.3 Dissimilarity Module

The dissimilarity module detects discrepancies between the predicted observation from the world model and the ground truth observation. First, it constructs multiple metrics for comparison. Then, it either combines those metrics to an anomaly map similar to Di Biase et al. [13] or it passes one metric through and uses just this single metric as anomaly map. In chapter 5, I evaluate the effectiveness of single metrics and their combination for anomaly detection. The objective of the dissimilarity module is to generate an anomaly map which represents for each pixel its likelihood to be part of an anomaly in the image space. In the following, I will explicate the different metrics for comparison.

4.3.1 Absolute Error

The absolute error is calculated using raw data from the RGB channels of the images to be compared. Let $r_{pred}(x,y), g_{pred}(x,y), b_{pred}(x,y) \in [0,1]$ be the RGB channels of the world model's prediction o_{pred} at the pixel with the position (x,y) and $r_{gt}(x,y), g_{gt}(x,y), b_{gt}(x,y) \in [0,1]$ be the RGB channels of the ground truth observation o_{gt} respectively. The absolute error $ABS(x,y)$ for the pixel at position (x,y) is calculated as follows:

$$ABS(x,y) = \frac{|r_{gt}(x,y) - r_{pred}(x,y)| + |g_{gt}(x,y) - g_{pred}(x,y)| + |b_{gt}(x,y) - b_{pred}(x,y)|}{3}$$

4.3.2 Squared Error

Similar to the absolute error, the squared error is calculated for each pixel independently and measures the error in the RGB channels. In contrast to the absolute error, the squared error weights larger errors more heavily than smaller errors in the RGB channels due to squaring the error rather than using its absolute value. The pixel-wise squared error is calculated as follows:

$$MSE(x,y) = \frac{(r_{gt}(x,y) - r_{pred}(x,y))^2 + (g_{gt}(x,y) - g_{pred}(x,y))^2 + (b_{gt}(x,y) - b_{pred}(x,y))^2}{3}$$

4.3.3 Structural Similarity

While the absolute error and squared error calculate the reconstruction error for each pixel independently, the Structural Similarity Index (SSIM) introduced by Wang et al. [28] takes into account that pixel which are closer to each other are likely more dependent on each other than pixel with a larger spatial distance [24]. By using such an index, it is possible to identify differences in the structure of two images and thus identify differences in e. g. texture. The SSIM is calculated with the following equation [28]:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

x and y are image signals and in the proposed anomaly detection model sliding spatial window patches of the images to be compared. μ_x and μ_y are the means and σ_x and σ_y are the variances of the pixel values of those patches. σ_{xy} is the covariance. To achieve numerical stability, two constants C_1 and C_2 are added in the equation [24, 28]. C_1 is calculated as follows: $C_1 = (K_1 L)^2$ with $K_1 \ll 1$ being a small constant and L being the pixel value range [28]. $C_2 = (K_2 L)^2$ with $K_2 \ll 1$ is calculated similarly [28].

The SSIM outputs values between -1 and 1. The higher the value, the more similar are the two signals to be compared [27]. Since the other metrics do not measure similarity, but errors and differences, the SSIM also has to be adjusted to values in $[0,1]$ where a higher value implies more dissimilarity regarding the SSIM. To achieve this, values are normalized for the linkage module using the following equation:

$$SSIM_{adjusted}(x,y) = 1 - \frac{SSIM(x,y)+1}{2}$$

4.3.4 Perceptual Difference

Perceptual difference leverages the capability of pre-trained VGG networks to extract complex features from images [13]. VGG networks are very deep convolutional networks with small convolutional filters [23]. For instance, the proposed model uses a pre-trained VGG network with 19 layers and a 3×3 convolutional filter which was introduced by Simonyan and Zisserman [23]. In order to calculate the perceptual difference, the model uses the methodology of Di Biase et al.: "For every pixel x of the input image and corresponding pixel r from the synthesized image, the perceptual difference is calculated as follows:

$$V(x,r) = \sum_{i=1}^{N} \frac{1}{M_i} \| F^{(i)}(x) - F^{(i)}(r) \|_1$$

where $F^{(i)}$ denotes the i-th layer with M_i elements of the VGG network and N layers. For consistency these dispersion measure is also normalized between $[0,1]$." [13]

In the experimental setups of this thesis, the input image is the ground truth image from the CARLA simulator and the synthesized image is the reconstructed image from the latent prior state of the world model. The VGG network uses weights which are pre-trained on the ImageNet [22] dataset.

4.3.5 Temporal Difference

In the section Time Delay in Prediction, I already introduced the idea of generating predictions emerging from various historic latent states by using the transition model repetitively. The temporal difference calculates the difference between those historic predictions and the current reconstruction. To achieve this, the anomaly detection model first calculates for each historic prediction its per-pixel absolute difference to the current reconstruction and then the pixel-wise mean of all those differences. This process is illustrated in figure 4.4. In contrast to the other metrics, the temporal difference does not compare the world model's reconstruction to ground truth sensory data, but rather considers differences in reconstructions of the world model internally.

4.4 Locating Anomalous Instances

In order to locate single instances which are anomalous, the proposed model uses an instance segmentation map, which maps each pixel to an instance in the observation. In chapter 5.4.5, an approach that does not utilize image segmentation, but uses the raw output of the dissimilarity module, is evaluated. This, however, results in anomaly maps which sometimes only classify parts of instances as anomalous or where corners of instances are blurry. Such an example is depicted in figure 5.2. In safety-critical systems such as autonomous driving cars, it is important that anomalous objects can clearly be located in an observation, for instance to start a driving

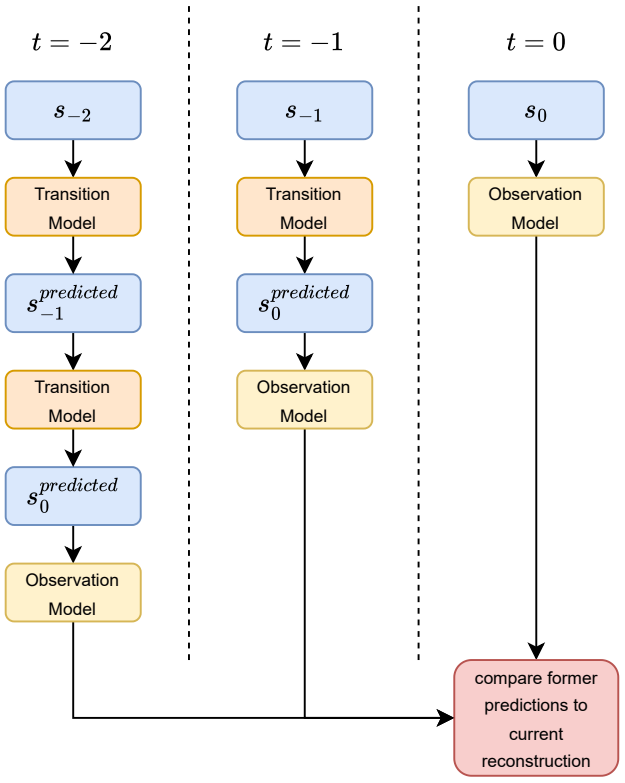


Figure 4.4: Illustration of the temporal difference calculation. Here, the temporal difference is two: The two former predictions for the observation at $t = 0$ are compared to the current reconstruction.

maneuver in order to drive around them. If the corners of the anomalous instance cannot be identified clearly, such a maneuver cannot be conducted safely.

In chapter 5, I use two types of instance segmentation maps for setups which utilize image segmentation: First, I use the ground truth instance segmentation map given by the evaluation dataset. Although instance segmentation maps are usually not inherently given by sensory data in real-world driving scenarios, this evaluating the anomaly detection model independently from errors introduced by instance segmentation methods. Another instance segmentation map is generated using the unsupervised image segmentation model U2Seg by Niu et al. [21]. Their model is capable of generating instance, semantic, and panoptic segmentation maps [21]. Although the maps generated by U2Seg are not as reliable as ground truth instance segmentation maps, they could be generated online on an autonomous driving system.

Given the anomaly map and one of the two introduced image segmentation maps, it is now possible to calculate the anomaly score for each instance in the observation. To achieve this, the proposed model iterates through each instance in the instance map and masks the anomaly map by extracting only pixels which are part of the respective instance. Afterwards, the mean anomaly score is calculated for that instance. The anomaly score for a pixel is in this case the value in the anomaly map at that pixel.

5 Evaluation

5.1 Evaluation Data

For the evaluation, the AnoVox dataset which contains numerous abnormal driving scenarios and was created using the CARLA 0.9.14 simulator is used. This dataset contains data about the taken action, a depth map, point cloud data, a route map, semantic point cloud data, information about the anomalies in the scenarios, instance segmentation maps, the RGB images, and semantic segmentation maps. AnoVox therefore contains the necessary input data for the underlying world model MUVO.

The evaluation dataset contains 16 scenarios with static anomalies such as an animal standing on the road. To achieve more variety in the environment, the driving scenarios take place in different towns in the CARLA simulator, under various weather conditions, and at different times of the day. Each scenario contains 200 frames. Therefore, the dataset contains 3200 RGB images of observations. Some anomalies are not visible in each frame of the scenario. The AnoVox dataset therefore also contains frames which do not represent abnormal driving situations. This is beneficial for testing anomaly detection models regarding falsely classified normal instances.

5.2 Evaluation Metrics

There are several evaluation metrics which are used in the experimental setup for the evaluation. The following metrics compare the predictions of the anomaly detection model to a ground truth map which was created using the ground truth semantic segmentation map of the AnoVox dataset.

The Average Precision **(AP)** and the False Positive Rate at 95 % True Positive Rate **(FPR95)** are very common metrics for evaluating anomaly detection models. For instance, they are also used as metrics in the popular Fishyscapes Benchmark by Blum et al. [3] which benchmarks anomaly detection methods on driving scenarios with real-world data. The Area under the Receiver Operating Characteristic curve **(AuROC)** is another evaluation metric which is used in this evaluation. The AP, FPR96 and AuROC scores do not require binary classification in either anomalous or not, but can also be used on anomaly scores in $[0, 1]$.

Similar to the Average Precision, the **F1** score considers both precision and recall. However, it calculates the harmonic mean of precision and recall. The True Negative Rate **(TNR)**, sometimes also referred to as Specificity, is a metric which considers true negative values in regard to both true negative and false positive values. This metric is the only metric which is also used when the scene does not contain an anomaly and therefore additionally measures how well the model deals with normality. Finally, the Positive Predictive Value **(PPV)** is used as an evaluation metric. The PPV measures the amount of true positives in respect to the amount of both true and false positives

in a classification.

5.3 Experimental Setup

MUVO allows the configuration of a receptive field which determines the amount of transitions the world model does per iteration. In this evaluation, a receptive field of twelve frames is chosen for the evaluation in order to calculate a temporal difference and a temporal delay of ten frames. The first transition outputs an unusable prior hidden state, since MUVO could not generate historical states prior to the first transition, resulting in eleven frames which the model uses for generating the differences between the last ten prior reconstructions and its current reconstruction. Ten frames were chosen as a distance since the evaluation dataset was generated with a fixed difference of 100 milliseconds between two observations, resulting in a time horizon of one second with this configuration. From the evaluation dataset, the data loader extracts 188 data points per scenario, making a total of 3008 data points in the evaluation dataset. Note that only 188 data points can be sampled since MUVO is configured with a receptive field of 12 frames, resulting in $200 - 12 = 188$ data points. The data loader samples each 10^{th} data point from this dataset. The argumentation for this sample rate is again the resulting difference of one second when using a fixed delta of 100 milliseconds in the configuration of the evaluation dataset. The sampling at different moments in the same driving scenario allows assessing the performance of anomaly detection models on the same anomaly having different distances to the car and being at different positions in the observation image.

5.4 Experimental Results

5.4.1 Evaluation of Average Anomaly Scores

First, the aptitude for anomaly detection of the generated average anomaly scores which are calculated for each instance in the observation independently were evaluated. The experimental results are presented in table 5.1. For this setup, the ground truth instance segmentation map of the evaluation dataset is chosen to find individual instances in the observations. Examples for outputs with this experimental setup are depicted in figure 5.1.

Initially, I ran the anomaly detection model on each image comparison metric individually. The perceptual difference outperforms all other configurations notably in each evaluation metric. This indicates that utilizing pre-trained very deep convolutional networks is highly beneficial for camera-based anomaly detection models which compare ground truth data to reconstructions of world models. A possible explanation for those good results might be that anomalies often occur in form of anomalous instances in an observation: Pre-trained VGG network are capable of extracting complex features, allowing to compare two images based on their content and the objects which are contained [13]. The first example in figure 5.1 depicts an example in which the perceptual difference was used as a comparison metric. The SSIM achieves a lower FPR95 and a higher AP and AuROC score than the pixel-wise calculated absolute and squared error metrics. One possible explanation for this could be the spatial characteristics of anomalies: Anomalies mostly affect

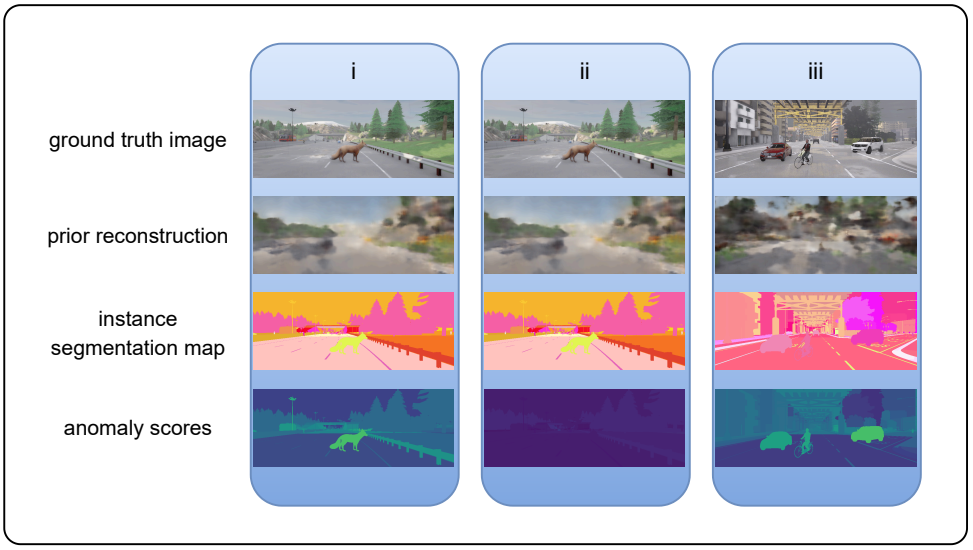


Figure 5.1: Examples for the output of the experimental setup with a ground truth instance segmentation map and average anomaly scores for each instance. i) In this case, the perceptual difference is used and the anomaly is correctly detected. The animal has a high anomaly score while other instances have a relatively low anomaly score. ii) In this example, the temporal difference is used. The model does not detect the anomaly correctly and generally outputs low anomaly scores. iii) In this case, the model maps a high anomaly score to cars and a cyclist, because they are not correctly reconstructed by the world model.

multiple proximate pixels. The SSIM takes into account that neighboring pixels tend to be more dependent on each other than pixels with a large spatial instance, whereas the absolute and squared error metrics are calculated element-wise. After evaluating metrics which compare sensory data to reconstructions, I assessed the performance of the temporal difference. For this experimental setup, I used a temporal difference of ten frames, meaning that ten former prior reconstructions are compared to the current prior reconstruction. The anomaly detection model classified only few instances as anomalous when using the temporal difference, indicating that the differences of varying temporal predictions tend to be very low. An example for an anomaly map which was generated with the temporal difference and only contains very low anomaly scores is shown in figure 5.1. While this might imply that differences from various temporal predictions do not enhance the performance of image-based anomaly detection models, one may view it as a quality indicator for the performance of the transition model: There are only small discrepancies between the predictions for the driving situation, suggesting that the transition model accurately predicted future latent states. Afterwards, I combined multiple metrics to an anomaly map. For the different weights, I chose the following structure: First, I combined each pixel-wise calculated error metric individually with the SSIM and perceptual loss. Then, I used the mean of all image comparison metrics which compare sensory data to reconstructions of MUVO. Then, I combined each pixel-wise calculated image comparison metric individually with the other metrics and the final combination is the mean of all five metrics. The second-best results are achieved when combining the squared error, the SSIM, and the perceptual difference, suggesting that combining multiple metrics can be beneficial for camera-based anomaly detection methods with world models.

weights					evaluation metrics		
w_{ABS}	w_{MSE}	w_{SSIM}	w_{per}	w_{temp}	AP ↑	FPR95 ↓	AuROC ↑
1.0	0.0	0.0	0.0	0.0	17.68	35.56	65.23
0.0	1.0	0.0	0.0	0.0	19.05	38.92	63.61
0.0	0.0	1.0	0.0	0.0	19.77	21.26	79.03
0.0	0.0	0.0	1.0	0.0	**29.90**	**16.93**	**83.18**
0.0	0.0	0.0	0.0	1.0	11.41	52.70	49.15
$\frac{1}{3}$	0.0	$\frac{1}{3}$	$\frac{1}{3}$	0.0	26.21	18.16	82.07
0.0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0.0	<u>27.50</u>	<u>17.81</u>	<u>82.47</u>
0.25	0.25	0.25	0.25	0.0	23.47	19.04	81.34
0.25	0.0	0.25	0.25	0.25	24.34	19.39	81.27
0.0	0.25	0.25	0.25	0.25	25.28	18.53	81.83
0.2	0.2	0.2	0.2	0.2	22.38	19.71	80.74

Table 5.1: Mean of evaluation metrics with all object classes considered, a ground truth instance segmentation map, and average anomaly scores. The respective best result is bold and the second-best result is underlined. (evaluation metrics in %)

Despite MUVO's multimodal approach improving prediction tasks on camera and lidar data [6], it is noteworthy that MUVO sometimes cannot reconstruct instances of some object classes correctly [6]. This leads to a failure class where cars or cyclists are for instance not in the output of the observation model. An example for such a case is given in figure 5.1. The presented anomaly detection model then tends to classify even "normal" instances as anomalous, since they are not in the reconstruction of the world model. Those false positives in the output of the anomaly detection model are not referable to the corner case detection model itself, but the utilized world model. In the next evaluation setup, semantic labels of pedestrians, busses, cars, bicycles with their cyclists, motorcycles with their motorcyclists, guardrails, and street lights were extracted and instances of those semantic classes were filtered out for the evaluation. A more detailed explanation of this filtering process including the semantic labels of object classes which are not considered in this part of the evaluation is given in A.1.

The experimental results without those often incorrectly reconstructed object classes are shown in table 5.2. The performance of the model improved in regard to most metrics. Those improved results thus support the hypothesis that the anomaly detection model depends on the reconstruction quality of the underlying world model. When comparing the configurations to each other, however, only minor disparities can be found: The perceptual loss achieves once again the best experimental results. The effectiveness of using pre-trained VGG networks for image comparison is therefore again substantiated. Second-best results are once again achieved when combining multiple metrics.

5.4.2 Evaluation of Maximum Anomaly Scores

In this subsection, not the average anomaly score of an instance is calculated, but the maximum anomaly score of a pixel in the observation is taken. The evaluation results are given in table 5.3. In general, the average anomaly score achieved better experimental results than this setup. This

weights					evaluation metrics		
w_{ABS}	w_{MSE}	w_{SSIM}	w_{per}	w_{temp}	AP ↑	FPR95 ↓	AuROC ↑
1.0	0.0	0.0	0.0	0.0	19.29	34.69	66.09
0.0	1.0	0.0	0.0	0.0	20.90	38.12	64.50
0.0	0.0	1.0	0.0	0.0	24.60	19.91	80.39
0.0	0.0	0.0	1.0	0.0	**35.94**	**15.34**	**84.76**
0.0	0.0	0.0	0.0	1.0	11.85	52.28	49.64
$\frac{1}{3}$	0.0	$\frac{1}{3}$	$\frac{1}{3}$	0.0	31.23	16.46	83.76
0.0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0.0	<u>33.02</u>	<u>16.14</u>	<u>84.12</u>
0.25	0.25	0.25	0.25	0.0	28.89	17.44	82.95
0.25	0.0	0.25	0.25	0.25	29.44	17.71	82.94
0.0	0.25	0.25	0.25	0.25	31.23	16.78	83.56
0.2	0.2	0.2	0.2	0.2	27.69	18.02	82.43

Table 5.2: Mean of evaluation metrics without incorrectly reconstructed object classes, a ground truth instance segmentation map, and average anomaly scores. The respective best result is bold and the second-best result is underlined. (evaluation metrics in %)

indicates that using the maximum anomaly score weights outliers too much and does not represent the likelihood of instances to be anomalous adequately. The first example in figure 5.2 depicts such a case. For the absolute error, however, using the maximum score led to a better Average Precision. With the maximum anomaly score, the perceptual loss is less eligible for anomaly detection: In this setup, an Average Precision of 17.26 % and a FPR95 score of 57.68 % was achieved when using the perceptual difference.

5.4.3 Evaluation of Average Anomaly Scores with Time Delay

The idea of using a prediction emerging from a prior state instead of the current reconstruction of the world model was introduced in chapter 4.1.1 and illustrated in figure 4.3. In this section, a time delay of 10 frames was chosen. The experimental results with this setup are given in table 5.4. In comparison to the respective experimental results with the current reconstruction, which are given in table 5.1, the quality of the predictions from the proposed model decreased when using the time delay. The differences, however, were minor. This again demonstrates that prior predictions and the current reconstruction present only small discrepancies.

5.4.4 Evaluation of Average Anomaly Scores with Generated Image Segmentation Map

In this subsection, I present the experimental results when the ground truth map is substituted with a panoptic segmentation map which was generated with U2Seg [21]. Such a map could, in contrast to ground truth maps, be generated online on an autonomous driving system. The experimental results with all object classes considered are given in table 5.5 and the results without object classes which are often not correctly reconstructed by MUVO are depicted in table 5.6. Examples for image segmentation maps, which were generated with U2Seg are shown in figure 5.3.

weights					evaluation metrics		
w_{ABS}	w_{MSE}	w_{SSIM}	w_{per}	w_{temp}	AP ↑	FPR95 ↓	AuROC ↑
1.0	0.0	0.0	0.0	0.0	19.00	59.68	40.68
0.0	1.0	0.0	0.0	0.0	18.86	59.52	40.76
0.0	0.0	1.0	0.0	0.0	10.87	67.03	33.30
0.0	0.0	0.0	1.0	0.0	17.26	57.68	42.55
0.0	0.0	0.0	0.0	1.0	11.01	74.23	25.97
$\frac{1}{3}$	0.0	$\frac{1}{3}$	$\frac{1}{3}$	0.0	18.71	<u>52.63</u>	<u>47.85</u>
0.0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0.0	**20.97**	**52.01**	**48.57**
0.25	0.25	0.25	0.25	0.0	19.86	53.36	47.22
0.25	0.0	0.25	0.25	0.25	15.44	57.88	42.90
0.0	0.25	0.25	0.25	0.25	16.69	56.77	43.76
0.2	0.2	0.2	0.2	0.2	<u>20.01</u>	56.84	43.82

Table 5.3: Mean of evaluation metrics with all object classes considered, a ground truth instance segmentation map, and the maximum anomaly score. The respective best result is bold and the second-best result is underlined. (evaluation metrics in %)

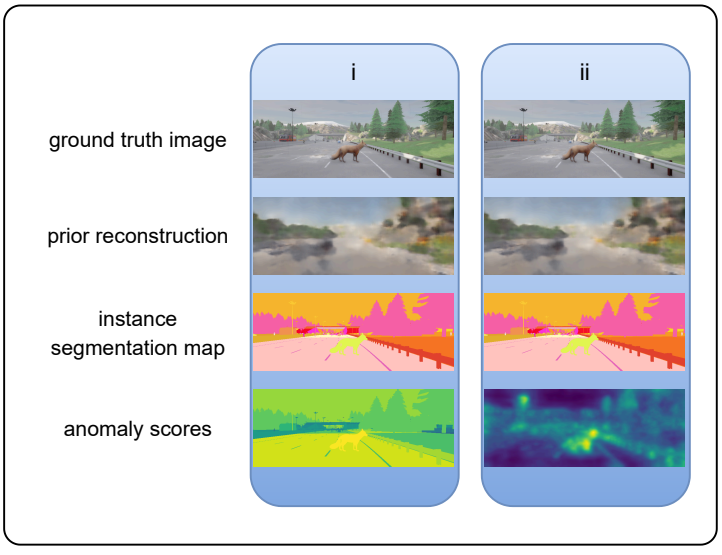


Figure 5.2: Examples for alternative experimental setups. The configuration of the examples is comparable to the configuration in the first example of figure 5.1: The perceptual difference is used and the moment in the driving scenario is the same. i) Here, the maximum anomaly score in an instance is used to calculate the respective anomaly score of the instance. This results in generally high values in the output. ii) In this example, the anomaly map from the dissimilarity module is used for the evaluation. An image segmentation map is not required in this setup.

weights					evaluation metrics		
w_{ABS}	w_{MSE}	w_{SSIM}	w_{per}	w_{temp}	AP ↑	FPR95 ↓	AuROC ↑
1.0	0.0	0.0	0.0	0.0	17.90	39.85	60.78
0.0	1.0	0.0	0.0	0.0	17.08	42.37	60.38
0.0	0.0	1.0	0.0	0.0	17.29	23.70	76.67
0.0	0.0	0.0	1.0	0.0	**28.97**	**17.99**	**82.23**
0.0	0.0	0.0	0.0	1.0	12.96	53.56	48.21
$\frac{1}{3}$	0.0	$\frac{1}{3}$	$\frac{1}{3}$	0.0	<u>23.20</u>	21.01	79.44
0.0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0.0	23.19	<u>19.73</u>	<u>80.73</u>
0.25	0.25	0.25	0.25	0.0	21.84	21.32	78.92
0.25	0.0	0.25	0.25	0.25	21.59	21.69	78.90
0.0	0.25	0.25	0.25	0.25	22.57	20.83	79.79
0.2	0.2	0.2	0.2	0.2	21.06	22.33	78.13

Table 5.4: Mean of evaluation metrics with all object classes considered, a ground truth instance segmentation map, average anomaly scores, and a time delay of 10 frames. The respective best result is bold and the second-best result is underlined. (evaluation metrics in %)

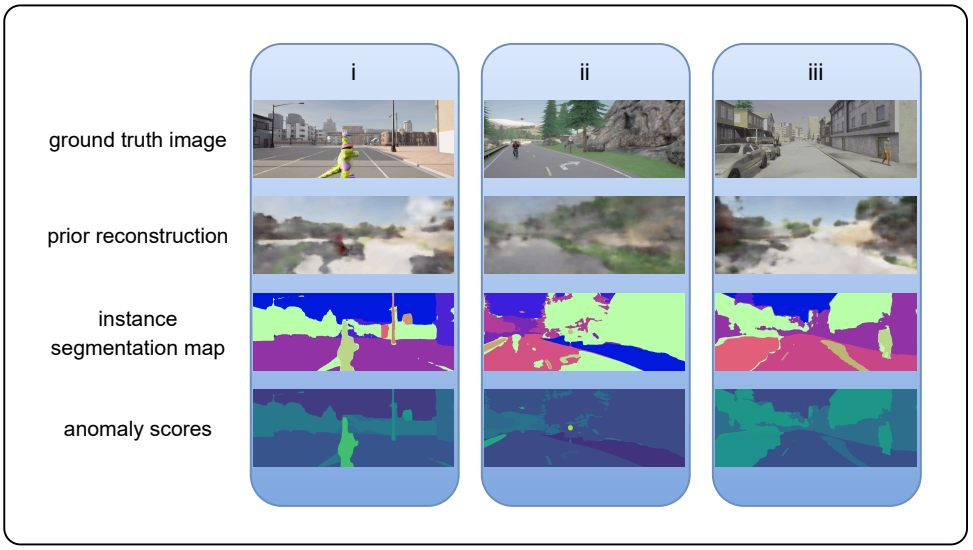


Figure 5.3: Examples for the setup with an image segmentation map which was generated with U2Seg. i) In this example, a large portion of the anomalous object is detected as an instance by the image segmentation method and correctly given a high anomaly score. ii) In this case, the road is fragmented into multiple instances. iii) Here, two cars are merged into one instance by the image segmentation approach.

weights					evaluation metrics		
w_{ABS}	w_{MSE}	w_{SSIM}	w_{per}	w_{temp}	AP ↑	FPR95 ↓	AuROC ↑
1.0	0.0	0.0	0.0	0.0	14.04	60.20	59.55
0.0	1.0	0.0	0.0	0.0	14.54	60.98	59.93
0.0	0.0	1.0	0.0	0.0	12.17	58.44	62.88
0.0	0.0	0.0	1.0	0.0	**18.88**	56.74	64.77
0.0	0.0	0.0	0.0	1.0	9.02	68.89	54.44
$\frac{1}{3}$	0.0	$\frac{1}{3}$	$\frac{1}{3}$	0.0	17.13	<u>56.73</u>	**65.50**
0.0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0.0	<u>17.70</u>	56.76	65.09
0.25	0.25	0.25	0.25	0.0	17.16	56.84	<u>65.41</u>
0.25	0.0	0.25	0.25	0.25	16.35	56.92	65.18
0.0	0.25	0.25	0.25	0.25	17.08	**56.54**	65.13
0.2	0.2	0.2	0.2	0.2	16.38	57.06	65.04

Table 5.5: Mean of evaluation metrics with all object classes considered, a generated image segmentation map, and average anomaly scores. The respective best result is bold and the second-best result is underlined. (evaluation metrics in %)

The Average Precision and AuROC are overall much lower and the FPR95 score much higher when the generated segmentation map is used. Such findings are to expected since the generated segmentation map is not as accurate as the ground truth instances segmentation map: As depicted in figure 5.3, U2Seg sometimes fuses multiple instances in the observation into one mask or contrary scatters single instances into multiple masks. These findings suggest that the anomaly detection model is highly dependant on the image segmentation model.

Despite observing differences when comparing both experimental setups with different image segmentation approaches to each other, statements regarding the suitability of separate combinations of image comparison metrics for camera-based anomaly detection can be made, regardless of the utilized segmentation map: The best results are achieved once again when choosing the perceptual difference or a combination of multiple metrics. In this setup, however, only using the perceptual difference has a slightly higher FPR95 and a slightly lower AuROC score than the respective best performing configuration.

5.4.5 Evaluation of Individual Anomaly Scores

Since substituting the ground truth instance segmentation map with a segmentation map which was generated with U2Seg led to inferior experimental results, the pixel-wise anomaly scores are evaluated instead of the average anomaly score of instances in this section. For this setup, the output of the dissimilarity module was used for the evaluation. This experimental setup does therefore not utilize image segmentation maps. The experimental results are given in table 5.7. An output image in which the pixel-wise anomaly scores are used is depicted in the second example in figure 5.2.

Generally, the experimental results are inferior to those of the setup with a ground truth instance segmentation map and average anomaly scores on the instance-level given in table 5.1. This indicates that loosing information about pixel regions of instances affects the quality of the anomaly

weights					evaluation metrics		
w_{ABS}	w_{MSE}	w_{SSIM}	w_{per}	w_{temp}	AP ↑	FPR95 ↓	AuROC ↑
1.0	0.0	0.0	0.0	0.0	15.02	59.12	60.36
0.0	1.0	0.0	0.0	0.0	15.41	59.95	60.68
0.0	0.0	1.0	0.0	0.0	13.20	57.11	63.96
0.0	0.0	0.0	1.0	0.0	**19.68**	55.57	65.67
0.0	0.0	0.0	0.0	1.0	9.41	68.25	54.84
$\frac{1}{3}$	0.0	$\frac{1}{3}$	$\frac{1}{3}$	0.0	18.08	55.60	**66.40**
0.0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0.0	<u>18.63</u>	<u>55.49</u>	66.08
0.25	0.25	0.25	0.25	0.0	18.26	55.70	<u>66.29</u>
0.25	0.0	0.25	0.25	0.25	17.35	55.76	66.08
0.0	0.25	0.25	0.25	0.25	18.05	**55.34**	66.05
0.2	0.2	0.2	0.2	0.2	17.37	55.92	65.91

Table 5.6: Mean of evaluation metrics without incorrectly reconstructed object classes, a generated image segmentation map, and average anomaly scores. The respective best result is bold and the second-best result is underlined. (evaluation metrics in %)

detection model.

Despite leading to lower average precision scores, the setup with pixel-wise anomaly scores outperforms the setup with the generated image segmentation map given in table 5.5 regarding the FPR95 and the AuROC scores. In respect to these two evaluation metrics, a setup without an image segmentation map is preferable to a setup with a segmentation map which was generated with U2Seg.

5.4.6 Evaluation of Instances with Highest Anomaly Scores

Prior experimental setups calculated anomaly scores for all instances in the observation or used all pixel-wise calculated anomaly scores from the dissimilarity module for the evaluation. As a consequence, the anomaly detection model also assigned an anomaly score to instances which are most likely not abnormal. In this section, an anomaly score greater than 0 is only assigned to the instances which have the largest average anomaly score. Pixels which are not part of those instances receive an anomaly score of 0.

There are two types of anomaly scores which are assigned to the instance with the largest average anomaly score: First, the average anomaly score of that instance is assigned to all pixels which are part of the respective instance. The first example in figure 5.4 exemplary depicts a case where this setup was used. Then, the pixel-wise raw anomaly scores of the instance were chosen for the evaluation. An example for such an anomaly map can be found in the second example of figure 5.4. The experimental results are given in table 5.8 and 5.9. Since all pixels which are not part of the instances with the largest anomaly score have an anomaly score of 0, the experimental results of the two different anomaly map calculations are nearly the same.

In this setup, however, it often occurred that not the anomalous instance, but a “normal” instance achieved the highest anomaly score. Such a situation is depicted in the third example in figure 5.4. This once again demonstrates how the reconstruction quality of MUVO affects the predictions of

weights					evaluation metrics		
w_{ABS}	w_{MSE}	w_{SSIM}	w_{per}	w_{temp}	AP ↑	FPR95 ↓	AuROC ↑
1.0	0.0	0.0	0.0	0.0	6.80	78.19	60.19
0.0	1.0	0.0	0.0	0.0	7.03	78.49	60.68
0.0	0.0	1.0	0.0	0.0	4.72	50.87	73.02
0.0	0.0	0.0	1.0	0.0	**10.86**	**32.91**	**79.51**
0.0	0.0	0.0	0.0	1.0	4.09	73.37	53.05
$\frac{1}{3}$	0.0	$\frac{1}{3}$	$\frac{1}{3}$	0.0	9.29	38.99	78.24
0.0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0.0	<u>9.51</u>	<u>37.84</u>	<u>78.70</u>
0.25	0.25	0.25	0.25	0.0	8.83	40.07	77.62
0.25	0.0	0.25	0.25	0.25	8.14	40.26	77.17
0.0	0.25	0.25	0.25	0.25	8.29	39.37	77.51
0.2	0.2	0.2	0.2	0.2	8.11	41.12	76.69

Table 5.7: Mean of evaluation metrics with all object classes considered and pixel-wise anomaly scores. The respective best result is bold and the second-best result is underlined. (evaluation metrics in %)

weights					evaluation metrics		
w_{ABS}	w_{MSE}	w_{SSIM}	w_{per}	w_{temp}	AP ↑	FPR95 ↓	AuROC ↑
1.0	0.0	0.0	0.0	0.0	5.04	93.57	50.54
0.0	1.0	0.0	0.0	0.0	5.04	93.57	50.53
0.0	0.0	1.0	0.0	0.0	5.83	92.83	51.12
0.0	0.0	0.0	1.0	0.0	<u>10.40</u>	88.49	53.26
0.0	0.0	0.0	0.0	1.0	5.06	93.57	50.59
$\frac{1}{3}$	0.0	$\frac{1}{3}$	$\frac{1}{3}$	0.0	8.88	89.93	52.59
0.0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0.0	**12.66**	**86.31**	**54.48**
0.25	0.25	0.25	0.25	0.0	8.83	89.93	52.57
0.25	0.0	0.25	0.25	0.25	10.37	<u>88.48</u>	<u>53.35</u>
0.0	0.25	0.25	0.25	0.25	8.88	89.93	52.57
0.2	0.2	0.2	0.2	0.2	8.07	90.66	52.18

Table 5.8: Mean of evaluation metrics with all object classes considered, a ground truth instance segmentation map, and the average anomaly score of only the instance with the largest average anomaly score. The respective best result is bold and the second-best result is underlined. (evaluation metrics in %)

weights					evaluation metrics		
w_{ABS}	w_{MSE}	w_{SSIM}	w_{per}	w_{temp}	AP ↑	FPR95 ↓	AuROC ↑
1.0	0.0	0.0	0.0	0.0	5.04	93.57	50.54
0.0	1.0	0.0	0.0	0.0	5.03	93.57	50.54
0.0	0.0	1.0	0.0	0.0	5.83	92.83	51.12
0.0	0.0	0.0	1.0	0.0	<u>10.40</u>	88.49	53.26
0.0	0.0	0.0	0.0	1.0	5.06	93.57	50.59
$\frac{1}{3}$	0.0	$\frac{1}{3}$	$\frac{1}{3}$	0.0	8.88	89.93	52.59
0.0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0.0	**12.66**	**86.31**	**54.48**
0.25	0.25	0.25	0.25	0.0	8.83	89.93	52.57
0.25	0.0	0.25	0.25	0.25	10.37	<u>88.48</u>	<u>53.35</u>
0.0	0.25	0.25	0.25	0.25	8.88	89.93	52.57
0.2	0.2	0.2	0.2	0.2	8.07	90.66	52.18

Table 5.9: Mean of evaluation metrics with all object classes considered, a ground truth instance segmentation map, and the anomaly scores of only the instance with the largest average anomaly score. The respective best result is bold and the second-best result is underlined. (evaluation metrics in %)

weights					evaluation metrics		
w_{ABS}	w_{MSE}	w_{SSIM}	w_{per}	w_{temp}	AP ↑	FPR95 ↓	AuROC ↑
1.0	0.0	0.0	0.0	0.0	5.77	92.85	50.87
0.0	1.0	0.0	0.0	0.0	7.29	91.40	51.64
0.0	0.0	1.0	0.0	0.0	7.35	91.38	51.90
0.0	0.0	0.0	1.0	0.0	<u>16.48</u>	<u>82.68</u>	<u>56.38</u>
0.0	0.0	0.0	0.0	1.0	5.06	93.57	50.52
$\frac{1}{3}$	0.0	$\frac{1}{3}$	$\frac{1}{3}$	0.0	13.40	85.58	54.92
0.0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0.0	**17.18**	**81.95**	**56.83**
0.25	0.25	0.25	0.25	0.0	12.61	86.30	54.52
0.25	0.0	0.25	0.25	0.25	14.16	84.85	55.30
0.0	0.25	0.25	0.25	0.25	13.45	85.58	54.92
0.2	0.2	0.2	0.2	0.2	11.08	87.76	53.72

Table 5.10: Mean of evaluation metrics without incorrectly reconstructed object classes, a ground truth instance segmentation map, and the anomaly scores of only the instance with the largest average anomaly score. The respective best result is bold and the second-best result is underlined. (evaluation metrics in %)

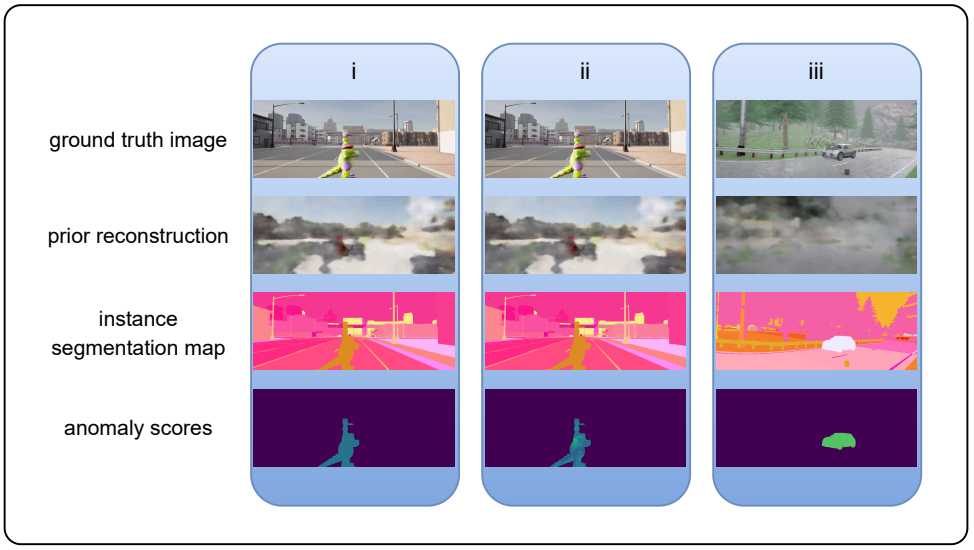


Figure 5.4: Examples of the experimental setup with only instances with the highest anomaly score. i) This example depicts the setup where the average anomaly score of the instance with the highest anomaly score is used. ii) In this case, the raw anomaly scores of the instance with the highest average anomaly score are used. iii) Here, the highest anomaly score is given to a normal instance in the observation which is not reconstructed correctly by the world model. This leads to ignoring the anomaly on the street.

the anomaly detection model. The falsely chosen instances could furthermore be a reason for prior experimental setups outperforming the setups of this section. Similar to prior experiments, instances of object classes which MUVO tends to reconstruct incorrectly were then extracted: In the setup of the experimental results depicted in table 5.10, instances with the highest anomaly score which are not part of such an incorrectly reconstructed object class were determined. Similar to prior experimental results, this setup achieved better experimental results than prior setups which considered all object classes.

5.4.7 Evaluation with Threshold

Prior experimental setups did not classify instances as anomalous or normal, but used an anomaly score in $[0, 1]$ for the evaluation. The following experimental setup uses a threshold to explicitly classify objects as normal or anomalous. In order to find a suitable threshold for the anomaly classification, the anomaly detection model was run with a dataset which does not contain anomalies. Afterwards, several percentiles of all calculated anomaly scores from the output of the dissimilarity module were calculated. By this, it is possible to ascertain a threshold, which distinguishes an instance being “normal” from an anomaly by using its average anomaly score. The dataset for determining this value must not contain anomalies, since auxiliary data would otherwise be required. For the ascertainment of the threshold, all metrics were weighted equally. After 50 iterations, the percentiles given in table 5.11 were calculated.

Since the largest distance is between the 80- and 90-percentiles, it seems that abnormal anomaly scores are especially over the 80-percentile and I therefore assessed a threshold of 0.1938 as appro-

60-percentile	0.1428
70-percentile	0.1664
80-percentile	0.1938
90-percentile	0.2351

Table 5.11: Different percentiles of anomaly scores.

priate in order to distinguish anomalous from normal instances. If an instance has a mean anomaly score of over 0.1938, it will be classified as anomalous in this experimental setup.

The experimental results are given in table 5.12. For this setup, the evaluation metrics which utilize binary classification are used. Similar to previous findings in this evaluation, combining multiple comparison metrics is beneficial: The best F1 score and the second-best PPV score are achieved when joining metrics. The temporal difference and the square error achieved a very high True Negative Rate. This is, however, not directly an indication for good performance: The squared error and the temporal difference simply classify only few pixels as anomalous, meaning that there are many true negatives with nearly no false positives. This leads to a high True Negative Rate, showing that the True Negative Rate should only be used in combination with other metrics as a performance indicator. With this setup, the absolute error, however, also led to promising results: In all metrics, the best or second-best results are achieved in a configuration where the absolute error is the only utilized metric.

5.4.8 Comparison to the State of the Art

Since the AnoVox dataset is rather new, only few anomaly detection models have been evaluated on this dataset. Therefore, it was necessary to first evaluate a state-of-the-art approach on the AnoVox dataset in order to compare the proposed model to state-of-the-art techniques. To achieve this, a converter was developed which converts the ground truth sensory images and segmentation maps of the AnoVox dataset to the structure of the SegmentMeIfYouCan [9] dataset. This then allows running an evaluation script of a state-of-the-art anomaly detection model for the SegmentMeIfYouCan dataset on the converted AnoVox dataset.

Ackermann et al. [1] have developed Maskomaly, which is a state-of-the-art anomaly detection model. In this evaluation, Maskomaly will be used for comparison to the proposed model of this thesis, since it has shown promising results on the SegmentMeIfYouCan benchmark and, similar to the proposed model of this thesis, does not require auxiliary data or additional training.

First, Maskomaly was run on the SegmentMeIfYouCan road anomaly validation dataset in order to assure that the setup in this evaluation achieves similar results to the evaluation of Ackermann et al. [1]. In this evaluation, Maskomaly is used with a Mask2Former [10] segmentation model with a Swin-L [20] backbone which is pre-trained on the Cityscapes [11] dataset. Mask2Former is an image segmentation model which uses in this setup a general-purpose Swin Transformer backbone. Cityscapes is a dataset for image segmentation tasks on urban scenes. With this setup, Maskomaly achieves an Average Precision of 94.10 %, a FPR95 score of 2.83 %, and an AuROC score of 98.93 % on the SegmentMeIfYouCan benchmark. Please note, that linear interpolation is

weights					evaluation metrics			
w_{ABS}	w_{MSE}	w_{SSIM}	w_{per}	w_{temp}	F1 ↑	PPV ↑	TNR ↑	$\text{TNR}_{filtered}$ ↑
1.0	0.0	0.0	0.0	0.0	<u>8.54</u>	**6.51**	76.52	74.86
0.0	1.0	0.0	0.0	0.0	2.40	2.12	**99.56**	**99.44**
0.0	0.0	1.0	0.0	0.0	7.97	4.86	46.11	47.05
0.0	0.0	0.0	1.0	0.0	7.52	4.76	17.58	16.37
0.0	0.0	0.0	0.0	1.0	2.30	1.76	<u>95.44</u>	<u>95.50</u>
$\frac{1}{3}$	0.0	$\frac{1}{3}$	$\frac{1}{3}$	0.0	7.73	4.68	40.86	40.24
0.0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0.0	8.21	5.04	53.09	51.63
0.25	0.25	0.25	0.25	0.0	8.38	5.17	58.43	55.99
0.25	0.0	0.25	0.25	0.25	7.87	4.79	50.61	48.69
0.0	0.25	0.25	0.25	0.25	8.41	5.25	62.30	59.37
0.2	0.2	0.2	0.2	0.2	**8.64**	<u>5.45</u>	66.09	64.04

Table 5.12: Mean of evaluation metrics with all object classes considered, a ground truth instance segmentation map, and a threshold for classification. Here, two types of a True Negative Rate are given: TNR includes all data points. $\text{TNR}_{filtered}$ was only calculated for observations which contain anomalous pixels in the sensory data. The respective best result is bold and the second-best result is underlined. (evaluation metrics in %)

used in this thesis to calculate the FPR95 score which slightly differs from the FPR95 calculation of Ackermann et al. [1]. To achieve comparable results, I evaluated Maskomaly with the FPR95 calculation which is used in this thesis.

Then, I evaluated Maskomaly on the AnoVox dataset. The results are given in table 5.13. The proposed model with the ground truth instance segmentation map achieves the best Average Precision and the best FPR95 score. With the generated image segmentation map which was built with U2Seg, however, Maskomaly outperforms the anomaly detection model of this thesis. Maskomaly also achieves the highest AuROC score. When using the pixel-wise anomaly score, the proposed anomaly detection model achieves nonetheless a much lower False Positive Rate at 95 % True Positive Rate than Maskomaly.

model	AP ↑	FPR95 ↓	AuROC ↑
Maskomaly [1]	<u>30.99</u>	53.16	**90.76**
the proposed model (ground truth, all classes, $w_{per} = 1$)	29.90	<u>16.93</u>	83.18
the proposed model (ground truth, restricted classes, $w_{per} = 1$)	**35.94**	**15.34**	<u>84.76</u>
the proposed model (U2Seg, all classes, $w_{per} = 1$)	18.88	56.74	64.77
the proposed model (U2Seg, restricted classes, $w_{per} = 1$)	19.68	55.57	65.67
the proposed model (anomaly score, $w_{per} = 1$)	10.86	32.91	79.51

Table 5.13: Experimental results of the evaluation of Maskomaly and the proposed model on the AnoVox dataset. The following configurations of the proposed model are depicted: i) all object classes with the perceptual difference as sole metric ii) without incorrectly reconstructed object classes and with the perceptual difference as sole metric iii) all object classes, with the perceptual difference as sole metric and with a generated image segmentation map iv) without incorrectly reconstructed object classes, with the perceptual difference as sole metric and with a generated image segmentation map v) with pixel-wise anomaly scores (output from the dissimiliarity module, no image segmentation map required) The respective best result is bold and the second-best result is underlined. (evaluation metrics in %)

6 Conclusion and Outlook

The proposed anomaly detection model demonstrates that the ability of world models to generate predictions of future observations by leveraging action-conditioned predictions of future latent states can be utilized for corner case detection. World models inherently provide useful features for anomaly recognition: In contrast to state-of-the-art anomaly detection techniques, the proposed model utilizes the observation model for reconstruction, the representation model for feature embedding, and the transition model for making predictions.

The experimental results of the evaluation furthermore indicate that utilizing pre-trained very deep convolutional networks is highly beneficial for camera-based anomaly detection approaches which leverage world models. Among different image comparison metrics, there is a great disparity between their eligibility for anomaly detection. Furthermore, it was illustrated that combining multiple image comparison metrics can lead to more reliable anomaly detection models by synthesizing the benefits of individual image comparison metrics. Moreover, this thesis presented that unsupervised image segmentation models can be leveraged for anomaly detection on camera-based approaches in order to detect anomalous instances in observations.

6.1 Future work

The primary constraint of the proposed model is its dependency on external systems such as image segmentation modules or models provided by the world model itself. For instance, if the world model reconstructs a “normal” object poorly, the proposed anomaly detection model tends to classify it as anomalous, since it encounters a difference between the reconstruction and sensory data. If the ground truth instance segmentation map is substituted by a segmentation map which was generated using an unsupervised image segmentation method, the model furthermore depends on the performance of this approach: If instances are falsely merged or fragmented, the anomaly detection model calculates an anomaly score for either multiple instances in combination or a subsection of an instance. In future work, it might be beneficial to test camera-based anomaly detection models with different instance segmentation approaches and possibly different world models in order to assess the impact of external models and to optimize the fit of the anomaly detection model with external systems. Additionally, it might be interesting to evaluate the approach on real-world data in contrast to data from the CARLA simulator to assess its operation in real driving scenarios.

A Appendix

A.1 Semantic Labels in Evaluation

The evaluation scripts of the corner case detection model outputs two maps which are relevant for the evaluation: One map contains the anomaly map of the anomaly detection model. The other map is a gray scale segmentation map, which depicts the object classes for each instance in the observation.

For the evaluation, I determined the value of static anomalies in the semantic segmentation map: Pixels, which are part of a static anomaly, have the value 156 in the gray scale semantic segmentation map. I therefore classified all pixels with this semantic value as anomalous.

Furthermore, I assess the performance of the proposed model without dynamic object classes in chapter 5. For this, I identified the semantic classes of pedestrians, busses, cars, bicycles with their cyclists, motorcycles with their motorcyclists, guardrails, and street lights. I determined the following values in the semantic map for those object classes: 243, 142, 137, 237, 141, 138, 242, 244, 238, 250, and 228. Pixels with those values in the semantic map were not considered in the evaluation when filtering poorly reconstructed object classes out.

B List of Figures

C List of Tables

D Bibliography

[1] J. Ackermann, C. Sakaridis, and F. Yu. Maskomaly: Zero-Shot Mask Anomaly Segmentation. In *The British Machine Vision Conference (BMVC)*, 2023.

[2] S. Bai, C. Han, and S. An. Recognizing Anomalies in Urban Road Scenes Through Analysing Single Images Captured by Cameras on Vehicles. *Sensing and Imaging*, 19, 2018.

[3] 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*, 2021.

[4] D. Bogdoll, L. Bosch, T. Joseph, H. Gremmelmaier, Y. Yang, and J. M. Zöllner. Exploring the Potential of World Models for Anomaly Detection in Autonomous Driving. In *2023 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2023.

[5] D. Bogdoll, M. Nitsche, and J. M. Zöllner. Anomaly Detection in Autonomous Driving: A Survey. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022.

[6] D. Bogdoll, Y. Yang, and J. M. Zöllner. MUVO: A Multimodal Generative World Model for Autonomous Driving with Geometric Representations. *arXiv preprint:2311.11762*, 2023.

[7] J. Breitenstein, J.-A. Termöhlen, D. Lipinski, and T. Fingscheidt. Systematization of Corner Cases for Visual Perception in Automated Driving. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, 2020.

[8] J. Breitenstein, J.-A. Termöhlen, D. Lipinski, and T. Fingscheidt. Corner Cases for Visual Perception in Automated Driving: Some Guidance on Detection Approaches. *arXiv preprint:2102.05897*, 2021.

[9] R. Chan, K. Lis, S. Uhlemeyer, H. Blum, S. Honari, R. Siegwart, P. Fua, M. Salzmann, and M. Rottmann. SegmentMeIfYouCan: A Benchmark for Anomaly Segmentation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

[10] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar. Masked-attention Mask Transformer for Universal Image Segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, 2022.

[11] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[13] G. Di Biase, H. Blum, R. Siegwart, and C. Cadena. Pixel-wise Anomaly Detection in Complex Driving Scenes. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[14] S. Hao, Y. Zhou, and Y. Guo. A Brief Survey on Semantic Segmentation with Deep Learning. *Neurocomputing*, 406, 2020.

[15] F. Heidecker, J. Breitenstein, K. Rösch, J. Löhdefink, M. Bieshaar, C. Stiller, T. Fingscheidt, and B. Sick. An Application-Driven Conceptualization of Corner Cases for Perception in Highly Automated Driving. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, 2021.

[16] A. Hu. *Neural World Models for Computer Vision*. PhD thesis, Wolfson College, 2023.

[17] A. Hu, G. Corrado, N. Griffiths, Z. Murez, C. Gurau, H. Yeo, A. Kendall, R. Cipolla, and J. Shotton. Model-Based Imitation Learning for Urban Driving. In *Proceedings of the 36th Conference on Neural Information Processing Systems*, 2022.

[18] A. Hu, L. Russell, H. Yeo, Z. Murez, G. Fedoseev, A. Kendall, J. Shotton, and G. Corrado. GAIA-1: A Generative World Model for Autonomous Driving. *arXiv preprint:2309.17080*, 2023.

[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*, 2014.

[20] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[21] D. Niu, X. Wang, X. Han, L. Lian, R. Herzig, and T. Darrell. Unsupervised Universal Image Segmentation. *arXiv preprint: 2312.17243*, 2023.

[22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115, 2015.

[23] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint: 1409.1556v6*, 2015.

[24] T. Vojir, T. Šipka, R. Aljundi, N. Chumerin, D. O. Reino, and J. Matas. Road Anomaly Detection by Partial Image Reconstruction with Segmentation Coupling. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[25] X. Wang, R. Girdhar, S. X. Yu, and I. Misra. Cut and learn for unsupervised object detection and instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[26] X. Wang, Z. Zhu, G. Huang, X. Chen, J. Zhu, and J. Lu. DriveDreamer: Towards Real-world-driven World Models for Autonomous Driving. *arXiv preprint:2309.09777*, 2023.

[27] Z. Wang and A. C. Bovik. Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures. *IEEE Signal Processing Magazine*, 26, 2009.

[28] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13, 2004.

[29] L. Zhang, Y. Xiong, Z. Yang, S. Casas, R. Hu, and R. Urtasun. Learning Unsupervised World Models for Autonomous Driving via Discrete Diffusion. *arXiv preprint:2311.01017*, 2024.

[30] Y. Zhu, Y. Tian, D. Metaxas, and P. Dollár. Semantic Amodal Segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.