Local and Global Information in Obstacle Detection on Railway Tracks

Author(s):
Brucker, Matthias; Cramariuc, Andrei; von Einem, Cornelius; Siegwart, Roland; Cadena, Cesar

Publication date:
2023

Permanent link:
https://doi.org/10.3929/ethz-b-000624526

Originally published in:
https://doi.org/10.1109/IROS55552.2023.10342174

Local and Global Information in Obstacle Detection on Railway Tracks

Matthias Brucker[2,*], Andrei Cramariuc[1,*], Cornelius von Einem[1,*], Roland Siegwart[1], and Cesar Cadena[1]

***Abstract*— Reliable obstacle detection on railways could help prevent collisions that result in injuries and potentially damage or derail the train. Unfortunately, generic object detectors do not have enough classes to account for all possible scenarios, and datasets featuring objects on railways are challenging to obtain. We propose utilizing a shallow network to learn railway segmentation from normal railway images. The limited receptive field of the network prevents overconfident predictions and allows the network to focus on the locally very distinct and repetitive patterns of the railway environment. Additionally, we explore the controlled inclusion of global information by learning to hallucinate obstacle-free images. We evaluate our method on a custom dataset featuring railway images with artificially augmented obstacles. Our proposed method outperforms other learning-based baseline methods.**

I. INTRODUCTION

With rising global demand for transportation by rail, both due to increasing global trade and changing consumer behavior, railway networks are reaching their operational capacities. To ensure the safe operation of trains on increasingly busier tracks, upgrades to the railway control systems are required in the form of reliable communication, continuous and accurate localization [1], and environmental awareness in the form of obstacle detection systems. Life-threatening risks to humans and extensive disruptions to the railway network operation can be prevented through the reliable detection of humans, animals or any unknown objects on the train tracks.

High vehicle speeds, low braking forces, and the great weight of trains result in braking distances exceeding several hundreds of meters, out of the range of common obstacle detection sensor systems, such as LiDAR or stereo vision. We thus propose a novel active long-range obstacle detection system, consisting of a zoomable high-focal length camera on an actuated platform [2], to detect potential obstacles, even at great distance [3] as shown in Figure 1. A critical aspect of such a system is the accuracy of the visual detection of known and unknown entities on the railway. So far, most research has focused on detecting pre-defined categories, such as humans or other trains, by training object detection networks on custom railway obstacle datasets [4]–[8]. These systems are, however, by design limited to this pre-defined set of categories and fail to generalize to unknown obstacle types. Moreover, the models must be trained on datasets containing artificially inserted obstacles or datasets out of the railway domain, as real images containing obstacles on rails are challenging to obtain. This leads to strong biases in the detectors and raises doubts about their capabilities in real-world applications.

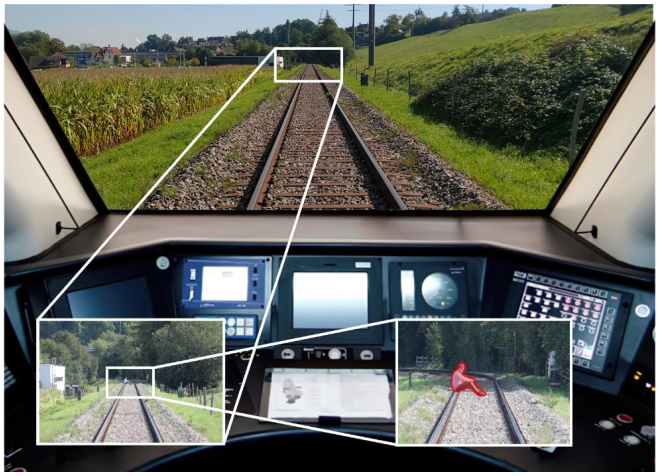

Fig. 1. Concept of a zoomable view enhancement with obstacle detection.

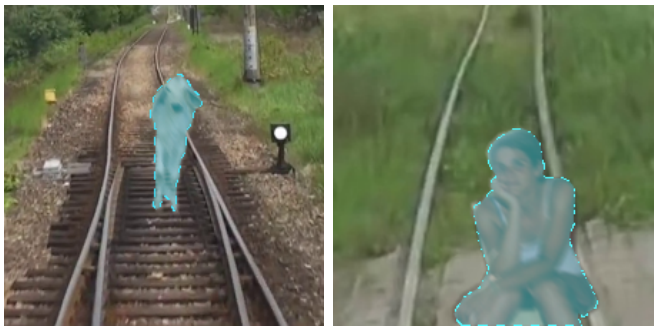

Fig. 2. Railway obstacle detection on the *FishyRails* dataset. Annotated in blue are the synthetically placed obstacles and our detection of them.

In this work, we focus on the task of data-driven anomaly detection, where instead of learning to detect a predefined set of obstacle categories, we learn to recognize the railway environment and by exclusion everything that does not belong, *i.e.* an anomaly, is also a potential obstacle. From here on, we will use the term anomaly detection to refer to generic obstacle detection, where the object category is not known at training time. Beyond the ability to detect generic obstacles, another advantage of our approach is that the training does not require real or artificially created examples of obstacles on railways. While visual anomaly detection has been extensively studied in the context of industrial inspection [9]–[11] and autonomous driving [12], [13], Boussik *et al.* [14] are the

*Authors contributed equally to this work, ordered alphabetically.

[1]Authors are members of the Autonomous Systems Lab, ETH Zurich, Switzerland; {firstname.lastname}@mavt.ethz.ch

[2]Author is with Magazino, Germany but the work was done while the author was a member of [1];

This work was supported by the ETH Mobility Initiative under the project *LROD-ADAS*. The code is available at https://github.com/ethz-asl/railway-anomaly-detection

only ones addressing the problem of anomaly detection for railway environments. They proposed training a series of Auto-Encoders (AEs) and using their reconstruction errors as a metric for anomalies. In our experiments, we find that using reconstruction errors for obstacle detection fails, for example, for small obstacles or obstacles with colors that match the background.

We propose a novel data-driven approach to anomaly detection in railway environments by reformulating the problem as a local segmentation task, where the global information available to the network is restricted. This bottlenecking limits over-confident predictions in the very well-defined and structured railway environment. We train a network with a limited *receptive field* (the patch size around a pixel the network sees) to segment railways from the background. We also include random (non-railway) images as negative examples in the training procedure. Inspired by Boussik *et al*. [14], we also study the incorporation of limited global information through obstacle-free images hallucinated by a neural network. Our models are trained on an obstacle-free training set consisting of a subset of *RailSem19* [15] and evaluated on an obstacle-enhanced version of it (see Figure 2), following a procedure proposed by Blum *et al*. [12]. We summarize our contributions as follows:

- A novel approach using shallow networks to perform visual anomaly detection that is ideal for the highly structured environment of railways. Additionally, our method does not require hard-to-obtain images of real anomalies on railways.
- We explore the inclusion of global information through the use of hallucinated obstacle-free reconstructions and reformulate the anomaly detection problem as a semantic difference detection task.
- An extensive evaluation on an object-enhances version of the *RailSem19* dataset [15] comparing our solution to multiple baselines and evaluating the advantages and disadvantages of each method.

II. RELATED WORK

A. Visual Anomaly Detection

In our case, visual anomaly detection is the task of detecting abnormalities or unknown entities in images based on the expected situation of the environment for normal operations. According to Yang *et al*. [16], works on visual anomaly detection can be classified into five categories. **Probabilistic methods** such as Gaussian (mixture) models, kernel density estimation [17], or variational Auto-Encoders (AEs) [18] aim at estimating the probability distribution of normal images and detecting the ones that fall out of the distribution. **One-class classification methods** follow a similar approach by constructing a decision boundary either through support vector machines [19], support vector data descriptors [20] or deep learning [21], [22]. **Reconstruction-based methods** commonly use AEs to learn a low-dimensional representation from which the original input can be reconstructed [10], [23]–[25]. Reconstruction errors at test time are then used as a metric for anomalous image regions. **Self-supervised anomaly detection methods** aim to learn significant and high-level features of normal samples, for example, through an auxiliary learning task [26], [27]. This avoids the need to precisely define what an anomaly is and does not require example data of anomalies. **Feature modeling methods** do not detect anomalies in the image space, but in a hand-crafted or learned feature space, such as features of pre-trained Convolutional Neural Networks (CNNs) [11], [28], [29]. A good example is the student teacher distillation framework proposed by Bergmann *et al*. [11], which achieved very good results on industrial images from the *MVTec AD* anomaly detection dataset [9]. The utilization of both deep and shallow CNNs has been explored for various of these methods [30], and it has been shown that shallow methods can outperform their deep counterparts in certain scenarios [31].

B. Visual Obstacle Detection on Railways

When developing new railway obstacle detection systems, it is important to consider the highly structured nature of the railway environment. The unique shape and color of rails and rail ties, the well-defined trajectories, and the limited number of object categories on the train tracks facilitate the use of classical image processing methods. This contrasts the more complex case of anomaly detection on roads, where learning-based methods are more essential [13]. Ristić-Durrant *et al*. [32] have performed an extensive review of existing obstacle detection systems with a wide variety of sensors. We focus on camera-based systems, which are directly comparable to our proposed approach. Rüder *et al*. [33] introduce a system for track and obstacle detection using edge detection, optical flow, and statistics of texture in an approach that is heavily tailored towards domain-specific grayscale images. However, evaluations regarding different object categories remain limited. Mukojima *et al*. [34], [35] propose a background subtraction-based method, which is, however susceptible to changes in lighting or environment. Rodriguez *et al*. [36] address this problem on the railway tracks themselves, which are observed using a Hough transform and a Canny edge detector. Discontinuities in this detection process signify obstacles on the track, though with limited robustness, as assumptions about the track geometry cause this approach to fail in scenarios with curved tracks. Uribe *et al*. [37] follow a similar approach with the same shortcomings. Learning-based object detection methods have also shown some success in detecting humans, trains, or luggage [4]–[8], [38]–[41], but they rely on custom datasets and are limited to a fixed set of object categories.

To this point, there is only little research applying visual anomaly detection methods to railway obstacle detection. Gasparini *et al*. [42] combine unsupervised image reconstruction with supervised detection of anomalies for nighttime railway inspection but are thus limited to thermal cameras. Boussik *et al*. [14] perform a grid search over AE structures with different optimizers, activations, and loss functions and evaluate them on a custom test dataset with artificially inserted obstacles and one real-world scenario. In our experiments,

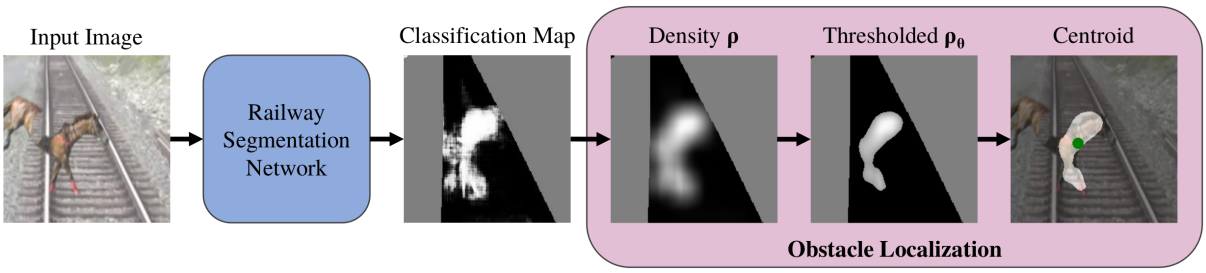

Fig. 3. Visual anomaly detection and localization using local segmentation. As input, we have a railway image with a synthetically added object. In the resulting classification map, the intensity of white represents the confidence that a pixel is anomalous, black denotes the railway, and the gray borders represent the background which is ignored. The final output is a detected object mask and its centroid.

these AE-based methods fail to detect small obstacles or those with colors common in the railway environment. Wang *et al*. [43] similarly train an AE, but detect anomalies by directly analyzing the distribution in the latent space and solely utilize the reconstruction for localizing detected anomalies.

III. Anomaly Detection Through Local Segmentation

A challenge for data-driven methods is the lack of existing training data featuring obstacles on railways and the difficulty of obtaining these images in such a safety-critical environment. This limits us to obstacle-free railway images and to other datasets featuring possible obstacles in non-railway scenarios. Additionally, we cannot limit detection to a fixed set of object categories, as we cannot predict in advance all possible obstacles we might encounter. Our approach to visual anomaly detection allows us to detect obstacles implicitly by exclusion. If it is on the railway but does not conform to known railway patterns, *i.e.* an anomaly, it is a potential obstacle.

We train an anomaly detection network by segmenting railways from the background as an auxiliary task. What sets our method apart is how we exploit the very structured and repetitive nature of our environment. Railway tracks have a very distinctive structure, which is easily identifiable even when looking at small patches taken from a larger image. We use a network with a small *receptive field* (the size of the patch around a pixel the network sees) as a means to restrict the global (*i.e.* contextual) information we provide to the anomaly detection network. This information bottleneck prevents issues of overconfidence in segmentation and classification tasks [44], examples of which we also show later in Section VI-A.

To provide negative examples, we include random (non-railway) images featuring a large variety of objects and scenes. Note that, in contrast to supervised object detection, we do not take an opinion on the semantic classes of obstacles but we label the entire image as background. The network is trained by minimizing the Binary Cross-Entropy (BCE) loss $\mathcal{L}_{BCE}$ for every pixel. We exclude the loss outside the track masks for railway images to avoid boundary issues and mislabeled pixels. Thus, we let all the negative examples come from non-railway images.

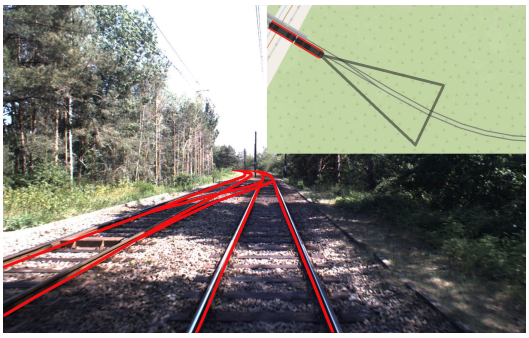

Fig. 4. Railway tracks being re-projected into the camera frame based on the known location and orientation of the train from an RTK-GPS.

A. Obstacle Localization With Classification Maps

An overview of the system is shown in Figure 3. During deployment, the previously trained network provides a classification map $\{\texttt{railway}, \texttt{background}\}$ covering each pixel in the image. Anomalies are then pixels that we know are railway but are classified as background. However, this requires a ground truth rail track segmentation mask to compare against. We assume prior knowledge about the location of the railway tracks in the image, which in practice can be obtained by having a prior map of the railway network and knowing the current pose of the train [45]–[47]. Using the map, the position of the train, and the camera intrinsics, the rail map can be projected into the image frame to obtain a ground truth mask as shown in Figure 4.

From a given classification map, obstacles are localized by computing a density map $\boldsymbol{\rho}$ via uniform filtering with a filter size of K. Subsequently, a threshold θ is applied, where all pixel values smaller than θ are zeroed to obtain the final obstacle mask $\boldsymbol{\rho_\theta}$. Decisions on the existence of the object can then be made based on the number of anomalous pixels or the size of the connected anomalous regions, as obtained, for example from a clustering algorithm.

IV. Incorporating Global Information

We hypothesize that incorporating global information into our, so far, purely local approach would lead to better performance. Inspired by the reconstruction-based approaches by Boussik *et al*. [14], we use a neural network to hallucinate obstacle-free images. However, instead of looking at the reconstruction error, we compute the semantic class difference

between regions in the original image and the hallucinated obstacle-free reconstruction (see Figure 5). The semantic differentiation network predicts in which pixels in the two images the semantic class is different without ever having to explicitly predict the class. For the same reasons as stated in Section III, we design the semantic differentiator network with a limited receptive field. This allows us to maintain a local informational bottleneck, which we have seen to be beneficial, while also allowing an earlier network to incorporate global information.

A. Obstacle-Free Railway Image Generation

For generating synthetic, obstacle-free railway images, we train the network with an Auto-Encoder (AE) structure to reconstruct the input image through a low-dimensional bottleneck. Choosing a small bottleneck prevents the network from simply replicating its input and forces the decoder to encode in its weights repeating patterns in the training data. As the training data does not include obstacles, the decoder should never learn how to reconstruct them, and the resulting images should be obstacle-free versions of the original input [14]. We test in total four different combinations of reconstruction losses that enforce different properties.

1) The **Mean Squared Error (MSE)** loss $\mathcal{L}_{MSE}$ computes the mean per-pixel L_2 distance between the original and reconstructed image. MSE pushes the network's output toward the dataset average, which causes the network to only preserve low-frequency components in the image at the cost of finer structural details.

2) The **Structural Similarity (SSIM)** loss $\mathcal{L}_{SSIM}$ as proposed by Wang *et al*. [48] forces patches in the original and reconstructed image to have similar luminance, contrast, and structure. This results in sharper images, but still, the network can only preserve a limited amount of realism in the reconstructed image.

3) Images generated from the previous methods can easily be identified as synthetic, which leads us to **Generative Adversarial Networks (GANs)** for more realistic image generation [49], [50]. Inspired by Isola *et al*. [51], we use a conditional GAN architecture [52] with a loss $\mathcal{L}_{GAN}$ to generate obstacle-free railway images. In an adversarial setting, the auto-encoding generator's objective is to reconstruct realistic (obstacle-free) railway images, while a discriminator aims at distinguishing real images from fake generated ones. Both are conditioned using our original segmentation mask, to promote a semantic similarity and to conserve the rail trajectories. Important to note is that apart from a semantic similarity constraint, the adversarial loss $\mathcal{L}_{GAN}$ does not promote visual similarity, such as color or structure, between the two images.

4) To preserve visual similarity, we expand the adversarial loss by applying the two **Histogram Losses** proposed by Avi-Aharon *et al*. [53]. As suggested by the authors, we combine both their proposed mutual information and Earth Mover's Distance losses as $\mathcal{L}_{HIST}$ into the GAN training setup. These two loss functions aid the GAN in retaining structural and color information, respectively, instead of pure semantic similarity.

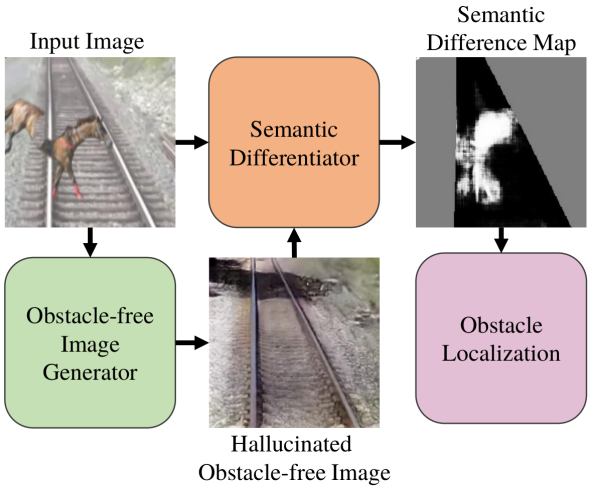

Fig. 5. Overview of the obstacle detection using hallucinated obstacle-free images. A semantic differentiator network does a pixel-wise prediction of where the original and obstacle-free images have different semantic meanings. This difference map can then be used to localize anomalies, as shown in Figure 3.

B. Object Localization with Semantic Difference Maps

The image reconstructed by our network provides global information about how each pixel would look like in an obstacle-free image. We leverage this information by training a network to compute the pixel-wise semantic difference map between the original and reconstructed obstacle-free image, as shown in Figure 5. We train the network with semantically similar and different image pairs. In the semantically similar case, we choose original and hallucinated versions of the same obstacle-free railway image and label every pixel as similar. In the semantically different case, we simulate an obstacle by replacing the original image with a random image from a dataset of non-railway images as in Section III and label every pixel as different, thus not focusing on specific classes. As before, we train the network by minimizing the BCE loss for every pixel and ignoring pixels outside the ground truth railway. Obstacles can then be localized using the same procedure as in Section III-A and Figure 3 on the difference map instead.

V. EXPERIMENTAL SETUP

A. Datasets

As a training dataset, we use the annotated subset of 7500 training images from the public *RailSem19* dataset [15], excluding tramway images. From these images, various regions of interest crops are obtained at a resolution of 224×224, resulting in a total of $26,810$ images. As additional non-railway images $\mathcal{X}_I$, we take $1,281,167$ crops from the public *ImageNet* dataset [54].

For our evaluation dataset *FishyRails*, we take the official annotated test set of *RailSem19*. We populate the images with objects from the public *PascalVOC* object segmentation dataset [55], resulting in 7142 images and segmentation

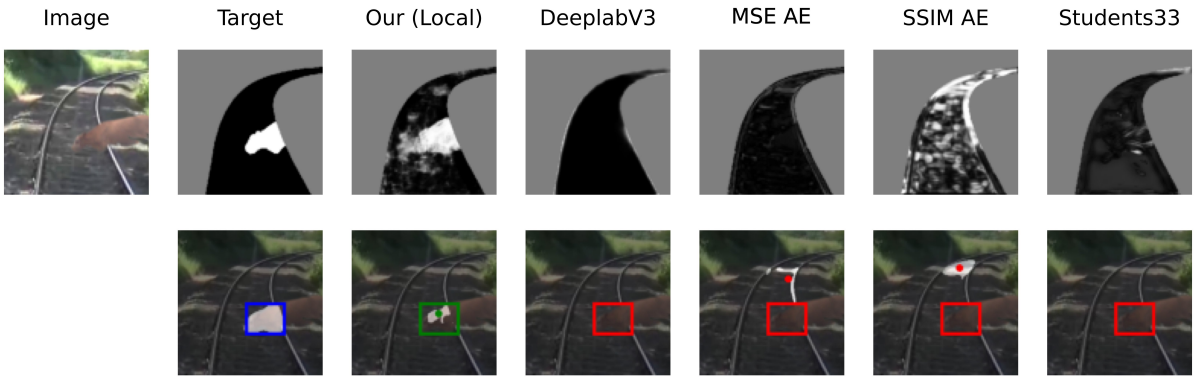

Fig. 6. Example image where only our local method localizes the obstacle correctly, while all baseline methods fail. The first row shows the predicted obstacle segmentation maps. Gray represents the areas outside the railway tracks where no detections are done, in black is the railway and white represents the anomaly detection mask. The second row displays localization results obtained via centroid computation.

masks. The pasting process is inspired by *Fishyscapes* [12], an object-enhanced dataset for measuring segmentation blind spots in traffic images, and involves image border smoothing, brightness correction, motion blur, depth blur, and Gaussian noise. Similar augmentations have also been proposed by [34], [37], [38], [56] for use in railway anomaly detection, as there are no public real-world railway obstacle detection datasets. Our synthetic obstacle testing method is also in line with the one proposed by Boussik *et al*. [14], who instead use a GAN to blend images of railways and obstacles on the *RailSem19* dataset.

B. Baselines

We compare against a set of state-of-the-art baselines selected from the related works. A standard semantic segmentation network, *DeeplabV3* [57], trained on obstacle-free railway images, to simply differentiate railway from background, serves as a first baseline. The second and third baselines are based on the work of Boussik *et al*. [14], consisting of two Auto-Encoders (AEs) for railway images, trained with either a Mean Squared Error (MSE) or Structural Similarity (SSIM) loss. The reconstruction error between the output and the original image is then used as a measure of anomalousness. We call these two baselines *MSE AE* and *SSIM AE* respectively. As our last baseline, we consider a patch-wise student teacher method[1] proposed by Bergmann *et al*. [11]. The authors report impressive results for visual anomaly detection on the *MVTec AD* industrial image dataset [9], and we expect comparable results in our uniform and equally well-structured railway environment.

C. Metrics

In order to assess the quality of the classification or difference map, we use the Area Under the Receiver Operator Characteristic Curve (AUROC). The AUROC is computed over all pixels of interest in the dataset given a ground-truth segmentation mask. Even though this is a good threshold-independent metric, it does not reflect well the applicability for our target use case of obstacle localization.

[1]Our re-implementation, as the original source code is not available.

TABLE I
THE AUROC AND LOCALIZATION F1 SCORE ON *FishyRails*.

Method	AE loss	AUROC	F1
DeeplabV3 [57]	-	0.817	0.535
MSE AE [14]	$\mathcal{L}_{MSE}$	0.686	0.498
SSIM AE [14]	$\mathcal{L}_{SSIM}$	0.737	0.429
Students33 [11]	-	0.594	0.458
Our (Local)	-	0.926	**0.863**
Our (Global)	$\mathcal{L}_{MSE}$	0.917	0.825
Our (Global)	$\mathcal{L}_{SSIM}$	0.915	0.812
Our (Global)	$\mathcal{L}_{GAN}$	0.935	0.857
Our (Global)	$\mathcal{L}_{GAN} + \mathcal{L}_{HIST}$	**0.936**	0.838

We say an obstacle is correctly localized if and only if the predicted centroid lies within the bounding box of the ground truth obstacle. Based on this localization, the F1 score can be computed as a second evaluation metric. To compute the centroid, we calculate the mean of the coordinates of all non-zero pixels (as seen in Figure 3). For all methods and baselines, we individually grid search for the optimal K and θ (see Section III-A) that maximize the F1 score. To note is that we implicitly assume there is at most one object in the image. This assumption holds in our experiments and provides a useful comparison metric. In practice, clustering could be used to distinguish multiple objects or an alert could be triggered based on the amount of anomalous pixels.

VI. RESULTS

A. Obstacle detection on FishyRails

For each method separately we use a fixed ρ and θ (see Section III-A) across all experiments. We determine the two thresholds separately for each method by picking the best-performing value across the training set.

After evaluation on our *FishyRails* dataset, the AUROC and F1 scores, as reported in Table I, paint a clear picture: our methods outperform the baselines by a large margin. Both the local and best global version ($\mathcal{L}_{GAN} + \mathcal{L}_{HIST}$) of our proposed method achieve an AUROC of 0.926 or better. This is significantly better than the AUROC of the best baseline method (*DeeplabV3*) at 0.817. The difference in

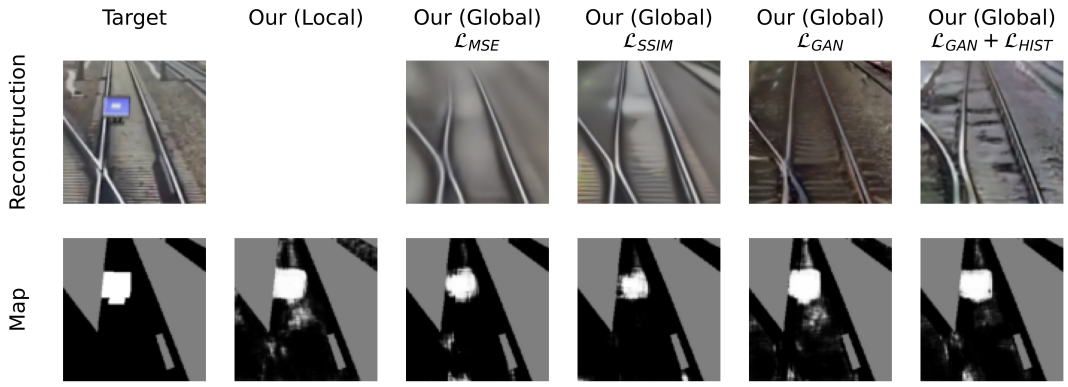

Fig. 7. Example image where both our local and global methods correctly segment the obstacle. The obstacle-free reconstructions are compared semantically with the original image on the left to find differences that could be anomalies. Note that the reconstructed images look different depending on the loss used to train the obstacle-free image generator. Nevertheless, in this example, all reconstructions successfully ignore the obstacle.

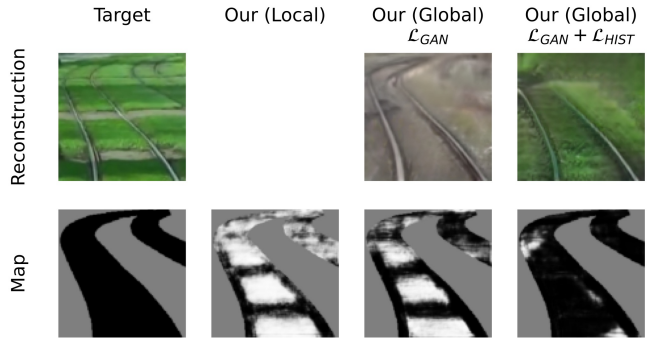

Fig. 8. Our global method with $\mathcal{L}_{GAN} + \mathcal{L}_{HIST}$ successfully reconstructs grass as non-anomalous. This allows the semantic difference network to correctly predict the images are similar and that there is no obstacle. The simpler GAN succeeds in reconstructing the railway but fails to match colors as it has not seen enough examples of grassy railways during training.

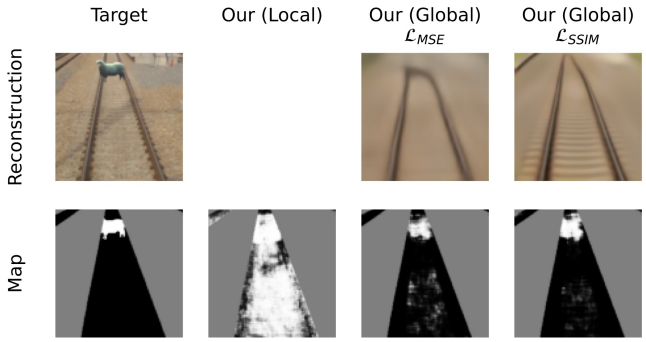

Fig. 9. Our local method misclassifies the sandy background as anomalous, because of lack of sand between the tracks during training. However, the $\mathcal{L}_{MSE}$ and $\mathcal{L}_{SSIM}$ networks succeed in removing the obstacle because of the distinct color difference to the railway.

F1 score is even larger, as our purely local method achieves a score of 0.863, while *DeeplabV3* only reaches a value of 0.535. We highlight this also visually in Figure 6 which shows an example where our local method succeeds at detecting the obstacle, whereas baseline methods fail. The comparison to *DeeplabV3* is important as it mirrors our local approach, except for a deeper network and a much higher receptive field. This highlights the importance of reducing the receptive field of the network to restrict global information and experimentally validates our hypothesis from Section III.

Interestingly, we find that *DeeplabV3* outperforms both the student teacher method *Students33* [11] and the reconstruction-based methods*MSE AE* [14] and *SSIM AE* [14]. *DeeplabV3* tends to over-confidently classify obstacle pixels as railway. During training, the model seems to learn that areas in between train tracks tend to correspond to the railway class, as it has never seen obstacles during training. This leads to small objects or ones with similar color as the background being misclassified.

From our investigations, we find that both *MSE AE* and *SSIM AE* fail when obstacles are small or have similar color as the background, with *SSIM AE* performing slightly better on obstacles with very prominent structure and contrast. Among all methods, *Students33* produces the lowest AUROC and F1 scores, probably because it was designed and optimized for industrial images with even less variance in structure and color than our dataset. The student networks are capable of detecting obstacles with salient colors but are unable to detect obstacles of colors that were frequently observed during student training, such as brown or gray.

B. Should We Include Global Information?

While the results reported in Table I show that our methods outperform the baselines by a large margin, the difference between our local (Section III) and global (Section IV) methods, is much smaller. The best AUROC score of 0.936 is achieved by our global method with $\mathcal{L}_{GAN} + \mathcal{L}_{HIST}$, but our local method yields the best F1 score. Therefore, a more detailed comparison of the observed strengths and failure cases in our experiments is needed. All global methods manage to remove most of the obstacles in their reconstruction stage for the four different loss combinations we tested (see Figure 7 for examples). When trained on anomaly-free railway and random non-railway patches, our generative models trained with $\mathcal{L}_{MSE}$ and $\mathcal{L}_{SSIM}$ learn to focus on pixel-wise color differences instead of semantic differences between the original and generated image. This leads to problems where

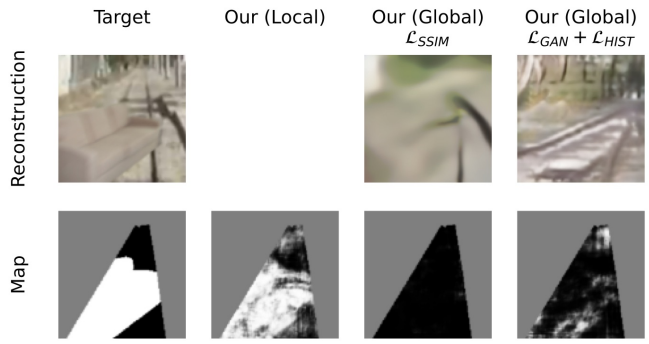

Fig. 10. Our global methods fail to reconstruct obstacle-free images with large occlusions, as they provide no prior on the potential tracks. The reconstruction failure also leads to the semantic difference map being meaningless. Our local method succeeds as it does not depend on the global context.

small obstacles or ones with similar background colors are reconstructed too well, instead of being removed.

The GAN based methods with losses $\mathcal{L}_{GAN}$ and $\mathcal{L}_{GAN} + \mathcal{L}_{HIST}$ succeed at rarely observed scenes by focusing on both visual and semantic reconstruction (see Figures 8 and 9 for examples). This leads to fewer false positives and is an instance where the inclusion of some global information into the process outperforms our purely local anomaly detection network. In cases where obstacles are too large, *i.e.* they cover too much of the image frame, the obstacle-free image generation process fails entirely, as seen in Figure 10. This is an instance where our purely local method outperforms global methods, as it does not rely on a reconstruction stage. Overall, the local and global versions of our methods have different weaknesses and strengths, while having similar overall performance according to our metrics. Ideally, a combination of multiple classifiers could be used. However, in practice, the local method is simpler and thus has fewer failure modes and might be preferable in a safety-critical application.

C. Ablation Study

We perform an ablation study on the receptive fields of both our local and global approaches using different loss functions. The results for the local method and the global ($\mathcal{L}_{GAN} + \mathcal{L}_{HIST}$) method are shown in Table II and highlight that according to the F1 score, a receptive field of 21 px is the optimum for us and has therefore been used in the evaluations of our method. For our global method, a similar pattern can be observed with a maximum at a receptive field of 29 px and utilizing the $\mathcal{L}_{GAN} + \mathcal{L}_{HIST}$ loss.

VII. CONCLUSION

In this paper, we have presented a novel approach to data-driven obstacle detection on railways. By training a network with a restricted receptive field on an auxiliary segmentation task, we are able to discern the well-structured railway background from any anomalies (obstacles). We are able to train the network without the need for hard-to-obtain data of obstacles on railways, and without having to restrict ourselves to a limited set of obstacle classes. Our method succeeds in cases where the benchmarks fail, *e.g.* small obstacles or obstacles with colors that blend into the background, but struggles with rarely seen railway environment types. In an extension, global information can be incorporated through hallucinated obstacle-free reconstructions. Successful reconstructions help detect anomalies even in rarely seen environments, though if the reconstruction fails, also the detection itself fails. Due to the limited availability of railway anomaly datasets, we evaluate our system on an obstacle-enhanced version of *Railsem19*, showing a significant improvement over state-of-the-art baselines.

TABLE II

ABLATION STUDY ON THE SIZE OF THE RECEPTIVE FIELD K_p [PIXELS] OF OUR LOCAL AND GLOBAL METHOD UTILIZING THE $\mathcal{L}_{GAN} + \mathcal{L}_{HIST}$ LOSS.

Method	K_p	ROC AUC	F1
Our (Local)	13	0.921	0.836
Our (Local)	21	0.926	**0.863**
Our (Local)	29	**0.928**	0.861
Our (Local)	35	0.925	0.839
Our (Local)	51	0.927	0.832
Our (Global)	13	0.931	0.839
Our (Global)	21	**0.936**	0.838
Our (Global)	29	**0.936**	**0.846**
Our (Global)	35	0.930	0.826
Our (Global)	51	0.914	0.782

REFERENCES

[1] P. Stanley, *ETCS for Engineers*, 1st ed. Roßdorf: TZ-Verl. & Print Gmbh, 2011.

[2] E. H. Assaf, C. von Einem, C. Cadena, R. Siegwart, and F. Tschopp, "High-precision low-cost gimballing platform for long-range railway obstacle detection," *MDPI Sensors*, vol. 22, no. 2, p. 474, 2022.

[3] M. Ukai, "Obstacle detection with a sequence of ultra telephoto camera images," in *Computers in Railways IX*, 2004, p. 10.

[4] M. Yu, P. Yang, and S. Wei, "Railway obstacle detection algorithm using neural network," *AIP Conference Proceedings*, vol. 1967, no. 1, 2018.

[5] J. Li, F. Zhou, and T. Ye, "Real-World Railway Traffic Detection Based on Faster Better Network," *IEEE Access*, vol. 6, 2018.

[6] Y. Xu, C. Gao, L. Yuan, S. Tang, and G. Wei, "Real-time Obstacle Detection Over Rails Using Deep Convolutional Neural Network," *IEEE Intelligent Transportation Systems Conference*, pp. 1007–1012, 2019.

[7] A. Chernov, M. Butakova, A. Guda, and P. Shevchuk, "Development of intelligent obstacle detection system on railway tracks for yard locomotives using CNN," *European Dependable Computing Conference*, pp. 33–43, 2020.

[8] D. Ristić-Durrant, M. A. Haseeb, M. Banić, D. Stamenković, M. Simonović, and D. Nikolić, "SMART on-board multi-sensor obstacle detection system for improvement of rail transport safety," *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 2021.

[9] P. Bergmann, "MVTec AD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9592–9600, 2019.

[10] P. Bergmann, S. Löwe, M. Fauser, D. Sattlegger, and C. Steger, "Improving unsupervised defect segmentation by applying structural similarity to autoencoders," *Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, vol. 5, pp. 372–380, 2019.

[11] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Uninformed Students: Student-Teacher Anomaly Detection with Discriminative Latent Embeddings," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4183–4192, 2020.

[12] 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, no. 11, pp. 3119–3135, 2021.

[13] Y. Yao, M. Xu, Y. Wang, D. J. Crandall, and E. M. Atkins, "Unsupervised Traffic Accident Detection in First-Person Videos," *IEEE International Conference on Intelligent Robots and Systems*, pp. 273–280, 2019.

[14] A. Boussik, W. Ben-Messaoud, S. Niar, and A. Taleb-Ahmed, "Railway obstacle detection using unsupervised learning: An exploratory study," *IEEE Intelligent Vehicles Symposium*, pp. 660–667, 2021.

[15] O. Zendel, M. Murschitz, M. Zeilinger, D. Steininger, S. Abbasi, and C. Beleznai, "RailSem19: A dataset for semantic rail scene understanding," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1221–1229, 2019.

[16] J. Yang, R. Xu, Z. Qi, and Y. Shi, "Visual Anomaly Detection for Images: A Systematic Survey," *Procedia Computer Science*, vol. 199, pp. 471–478, 2021.

[17] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

[18] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *International Conference on Learning Representations*, pp. 1–14, 2014.

[19] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[20] D. M. Tax and R. P. Duin, "Support Vector Data Description," *Machine Learning*, vol. 54, pp. 45–66, 2004.

[21] P. Oza and V. M. Patel, "One-Class Convolutional Neural Network," *IEEE Signal Processing Letters*, vol. 26, no. 2, pp. 277–281, 2019.

[22] L. Ruff, R. A. Vandermeulen, N. Görnitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, "Deep one-class classification," *35th International Conference on Machine Learning*, vol. 10, pp. 6981–6996, 2018.

[23] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "GANomaly: Semi-supervised anomaly detection via adversarial training," in *Computer Vision*. Springer International Publishing, 2019, pp. 622–637.

[24] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," *International Conference on Information Processing in Medical Imaging*, pp. 146–157, 2017.

[25] H. Zenati, C. S. Foo, B. Lecouat, G. Manek, and V. R. Chandrasekhar, "Adversarially learned anomaly detection," *IEEE International Conference on Data Mining*, pp. 727–736, 2018.

[26] I. Golan and R. El-Yaniv, "Deep anomaly detection using geometric transformations," *Advances in Neural Information Processing Systems*, vol. 31, pp. 9758–9769, 2018.

[27] K. Lis, K. K. Nakka, P. Fua, and M. Salzmann, "Detecting the unexpected via image resynthesis," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2152–2161, 2019.

[28] Y. Shi, J. Yang, and Z. Qi, "Unsupervised anomaly segmentation via deep feature reconstruction," *Neurocomputing*, vol. 424, pp. 9–22, 2021.

[29] N. Cohen and Y. Hoshen, "Sub-Image Anomaly Detection with Deep Pyramid Correspondences," *arXiv 2005.02357*, 2020.

[30] L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich, and K.-R. Muller, "A Unifying Review of Deep and Shallow Anomaly Detection," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 756–795, 2021.

[31] D. Kwon, K. Natarajan, S. C. Suh, H. Kim, and J. Kim, "An Empirical Study on Network Anomaly Detection Using Convolutional Neural Networks," in *IEEE International Conference on Distributed Computing Systems*, 2018, pp. 1595–1598.

[32] D. Ristić-Durrant, M. Franke, and K. Michels, "A Review of Vision-Based On-Board Obstacle Detection and Distance Estimation in Railways," *MDPI Sensors*, vol. 21, no. 10, p. 3452, 2021.

[33] M. Rüder, N. Möhler, and F. Ahmed, "An obstacle detection system for automated trains," *IEEE Intelligent Vehicles Symposium, Proceedings*, pp. 180–185, 2003.

[34] H. Mukojima, D. Deguchi, Y. Kawanishi, I. Ide, H. Murase, M. Ukai, N. Nagamine, and R. Nakasone, "Moving camera background-subtraction for obstacle detection on railway tracks," in *IEEE International Conference on Image Processing*, 2016, pp. 3967–3971.

[35] R. Nakasone, N. Nagamine, M. Ukai, H. Mukojima, D. Deguchi, and H. Murase, "Frontal Obstacle Detection Using Background Subtraction and Frame Registration," *Quarterly Report of RTRI*, vol. 58, no. 4, pp. 298–302, 2017.

[36] L. A. Rodriguez, J. A. Uribe, and J. F. Bonilla, "Obstacle detection over rails using hough transform," *17th Symposium of Image, Signal Processing, and Artificial Vision (STSIVA)*, pp. 317–322, 2012.

[37] J. A. Uribe, L. Fonseca, and J. F. Vargas, "Video based system for railroad collision warning," in *IEEE International Carnahan Conference on Security Technology*, 2012, pp. 280–285.

[38] M. Franke, V. Gopinath, D. Ristić-Durrant, and K. Michels, "Object-level data augmentation for deep learning-based obstacle detection in railways," *Applied Sciences*, vol. 12, no. 20, p. 10625, Oct 2022.

[39] A. Mahtani, W. Ben-Messaoud, A. Taleb-Ahmed, S. Niar, and C. Strauss, "Pedestrian Detection and Classification for Autonomous Train," in *IEEE International Conference on Image Processing, Applications and Systems*, 2020, pp. 52–57.

[40] S. Gupta, N. Mohan, P. Nayak, K. C. Nagaraju, and M. Karanam, "Deep vision-based surveillance system to prevent train–elephant collisions," *Soft Computing*, vol. 26, no. 8, pp. 4005–4018, 2022.

[41] T. Ye, X. Zhang, Y. Zhang, and J. Liu, "Railway Traffic Object Detection Using Differential Feature Fusion Convolution Neural Network," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1375–1387, 2021.

[42] R. Gasparini, A. D'Eusanio, G. Borghi, S. Pini, G. Scaglione, S. Calderara, E. Fedeli, and R. Cucchiara, "Anomaly detection, localization and classification for railway inspection," *International Conference on Pattern Recognition*, pp. 3419–3426, 2020.

[43] Y. Wang, Z. Yu, and L. Zhu, "Intrusion detection for high-speed railways based on unsupervised anomaly detection models," *Applied Intelligence*, July 2022.

[44] J. Gast and S. Roth, "Lightweight Probabilistic Deep Networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3369–3378.

[45] Z. Hu and K. Uchimura, "Fusion of vision, GPS and 3D gyro data in solving camera registration problem for direct visual navigation," *International Journal of ITS Research*, vol. 4, no. 1, p. 10, 2006.

[46] F. Tschopp, T. Schneider, A. W. Palmer, N. Nourani-Vatani, C. Cadena, R. Siegwart, and J. Nieto, "Experimental Comparison of Visual-Aided Odometry Methods for Rail Vehicles," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1815–1822, 2019.

[47] J. Otegui, A. Bahillo, I. Lopetegi, and L. E. Diez, "A Survey of Train Positioning Solutions," *IEEE Sensors Journal*, vol. 17, no. 20, pp. 6788–6797, 2017.

[48] 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*, vol. 13, no. 4, pp. 600–612, 2004.

[49] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[50] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[51] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1125–1134, 2017.

[52] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," *arXiv 1411.1784*, 2014.

[53] M. Avi-Aharon, A. Arbelle, and T. R. Raviv, "DeepHist: Differentiable Joint and Color Histogram Layers for Image-to-Image Translation," *arXiv 2005.03995*, 2020.

[54] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[55] M. Everingham, S. M. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes Challenge: A Retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2015.

[56] B. Guo, G. Geng, L. Zhu, H. Shi, and Z. Yu, "High-speed railway intruding object image generating with generative adversarial networks," *MDPI Sensors*, vol. 19, no. 14, 2019.

[57] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.