Track Intrusion Detection and Track Integrity Evaluation

Yu Qian
Associate Professor
Department of Civil and Environmental Engineering
University of South Carolina

Youzhi Tang
Ph.D. Candidate
Department of Civil and Environmental Engineering
University of South Carolina

Dimitris Rizos
Professor
Department of Civil and Environmental Engineering
University of South Carolina

Nikolaos Vitzilaios
Associate Professor
Department of Mechanical Engineering
University of South Carolina

A Report on Research Sponsored by

University Transportation Center for Railway Safety (UTCRS)

Molinaroli College of Engineering and Computing
University of South Carolina

September 2024

Technical Report Documentation Page

1. Report No. UTCRS-USC-O5CY23	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Track Intrusion Detection and Track Integrity Evaluation		5. Report Date September 30, 2024
		6. Performing Organization Code UTCRS-USC
7. Author(s) Yu Qian, Youzhi Tang, Dimitris Rizos, and Nikolaos Vitzilaios		8. Performing Organization Report No. UTCRS-USC-O5CY23
9. Performing Organization Name and Address University Transportation Center for Railway Safety (UTCRS) University of South Carolina (USC) Columbia, SC 29208		10. Work Unit No. (TRAIS)
		11. Contract or Grant No. 69A3552348340
12. Sponsoring Agency Name and Address U.S. Department of Transportation (USDOT) University Transportation Centers Program 1200 New Jersey Ave. SE Washington, DC, 20590		13. Type of Report and Period Covered Project Report June 1, 2023 – August 31, 2024
		14. Sponsoring Agency Code USDOT UTC Program
15. Supplementary Notes		

16. Abstract

Track intrusion, particularly trespassing within railroad rights-of-way, poses a major safety risk, leading to more fatalities than train-vehicle collisions. This research introduces a Hybrid Region-based Convolutional Network (Hybrid-RCNN) that integrates foreground segmentation and object detection to enhance surveillance at rail crossings. The model performs multiple tasks, including object detection, classification, and tracking, to efficiently monitor unauthorized activities. Testing shows the Hybrid-RCNN's superior accuracy compared to models like Mask-RCNN, highlighting its potential to improve railway safety by identifying and mitigating hazards more effectively.

17. Key Words Railroad Tracks, Trespassers, Artificial Intelligence, Neural Networks, Computer Vision	18. Distribution Statement This report is available for download from https://www.utrgv.edu/railwaysafety/research/operations/index.htm		
19. Security Classification (of this report) None	20. Security Classification (of this page) None	21. No. of Pages 15	22. Price

Table of Contents

List of Figures

List of Abbreviations

AI	Artificial Intelligence
CNNs	Convolutional Neural Networks
COCO	Common Objects in Context
GTR	Global Tracking Transformer
LVIS	Large Vocabulary Instance Segmentation
RCNN	Region-based Convolutional Network
TAO	Tracking Any Object
USDOT	U.S. Department of Transportation
UTC	University Transportation Center

Disclaimer

Acknowledgements

The authors wish to acknowledge the University Transportation Center for Railway Safety (UTCRS) for funding this project under the USDOT UTC Program Grant No 69A3552348340.

1. SUMMARY

Track intrusion, especially trespassing, which encompasses unauthorized entry and lingering within the railroad right-of-way, is a significant safety concern. It has been associated with a higher number of fatalities compared to incidents involving collisions between vehicles and trains. This stark statistic underscores the urgent need for advanced surveillance and detection systems at rail crossings to ensure track integrity and enhance overall railway safety. This research aims to develop a novel Hybrid Region-based Convolutional Network (Hybrid-RCNN) that innovatively merges foreground segmentation and object detection methodologies to form a comprehensive railroad crossing surveillance system, as shown in **Figure 1**.

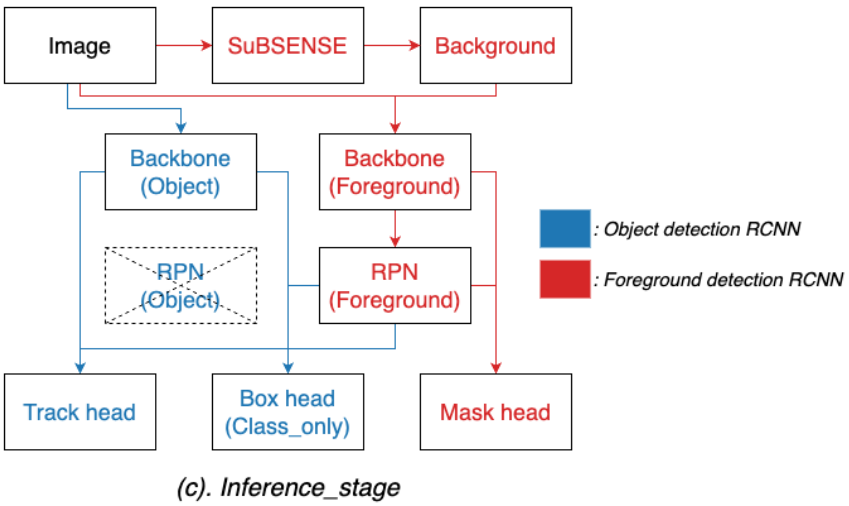

Figure 1: Hybrid Region-based Convolutional Network (Hybrid-RCNN) architecture

The design of the proposed Hybrid-RCNN model is strategically tailored to perform multiple functions simultaneously: foreground detection, segmentation, classification, and tracking of objects. This multifaceted approach allows for a more precise and efficient monitoring system adept at identifying and responding to any non-compliant objects or unauthorized activities within the railroad area. The integrated system not only detects but also classifies various types of objects, making it possible to differentiate between harmless elements and potential hazards.

Our evaluation and testing phase highlights the effectiveness of the Hybrid-RCNN model. The findings from these experiments validate the robustness of the network, demonstrating its

superior capability to accurately identify and track unauthorized or non-compliant objects at railroad crossings. This enhanced detection is critical as it significantly surpasses the performance of traditional object detection models such as Mask-RCNN. Through a series of comparative analyses, the Hybrid-RCNN consistently outperformed existing models, proving its potential as a pivotal technology in railway safety systems.

2. BACKGROUND

Accidental intrusions at rail crossings represent a major threat to railroad safety. Over the past decades, numerous enhancements have been implemented to improve safety at rail crossings, including armed gates, upgraded road signs, barriers, traffic warning lights, and surveillance cameras [1]. Despite these advancements, many incidents still occur due to unpredictable trespassing, track fouling, and unintentional obstructions. Traditional surveillance methods are inadequate as they fail to detect obstacles at rail crossings and do not provide real-time alerts to both incoming trains and road traffic, encompassing both vehicles and pedestrians.

In recent years, the rapid advancement of deep learning and artificial intelligence (AI) has led to significant success for Convolutional Neural Networks (CNNs) in various computer vision applications, including image classification, object detection, and semantic segmentation. The application of CNNs in enhancing railroad safety and track resilience has become increasingly popular. CNN-based models have significantly improved detection efficiency and accuracy, reducing human errors and supporting auxiliary decision-making. For instance, Zaman et al. utilized Mask R-CNN to detect intrusion events on the railroad [2]. Additionally, Guo et al. developed an automated video analysis detection and tracking system to assess traffic conditions at rail crossings [3]. Among these deep learning approaches, object detection and foreground segmentation are crucial for video surveillance, playing an indispensable role in ensuring safety at railroad crossings.

Object detection involves identifying and pinpointing instances of specific object classes within an image or a video frame [4]. This technique primarily employs deep learning algorithms and convolutional neural networks to recognize and classify various objects such as cars, pedestrians, or animals, and to determine their boundaries or locations within the scene. By accurately detecting these objects, this method provides crucial information about their presence, position, and characteristics. Such data is essential for enabling further analyses such as tracking

or behavioral assessment [4, 5].

On the other hand, foreground segmentation, also known as change detection or background subtraction, is a technique designed to differentiate moving elements, known as the foreground, from the static scene, termed the background [6]. This differentiation is achieved by analyzing the differences between consecutive frames in a video sequence and identifying regions with significant changes, which are subsequently classified as the foreground. Foreground segmentation is especially valuable in applications where detecting and monitoring the movement of objects within an environment is crucial, such as in traffic monitoring or intrusion detection systems.

While foreground segmentation effectively identifies all moving elements within a scene as foreground objects, it lacks the ability to classify or recognize these objects. This means that while it can detect motion, it cannot offer specific insights into the nature or type of the moving objects. On the other hand, object detection is capable of recognizing and classifying common objects like pedestrians and cars, thereby providing a more comprehensive understanding of the scene. However, object detection can struggle with uncommon or irregular objects, potentially leading to false-negative errors (where an object is not detected) or false-positive errors (where background objects are mistakenly identified as foreground). These limitations highlight the need for integrating multiple detection and segmentation approaches to enhance overall accuracy and reliability in object detection systems.

To address the limitations of both foreground detection and object detection in video surveillance, this study introduces Hybrid-RCNN. This innovative approach synergistically combines the strengths of both foreground segmentation and object detection techniques. By integrating these methodologies, the Hybrid-RCNN is designed to concurrently perform foreground detection, segmentation, classification, and tracking. This integration results in a video surveillance system that is both more accurate and efficient, offering enhanced capabilities for identifying and monitoring various objects within a dynamic environment.

3. METHODOLOGY

3.1 Foreground Detection RCNN

Input head

Instead of object detection and segmentation, the foreground detection RCNN of Hybrid-RCNN performs foreground detection and segmentation. To achieve this goal, a simple modification is required, adding an input head to the existing Faster-RCNN architecture. This setup allows the foreground detection RCNN to accept both the current frame and background images as inputs. The model then executes foreground detection by comparing the differences between these two inputs. Worth noting is that the background images are generated and continually updated via the aforementioned SuBSENSE [8] algorithm.

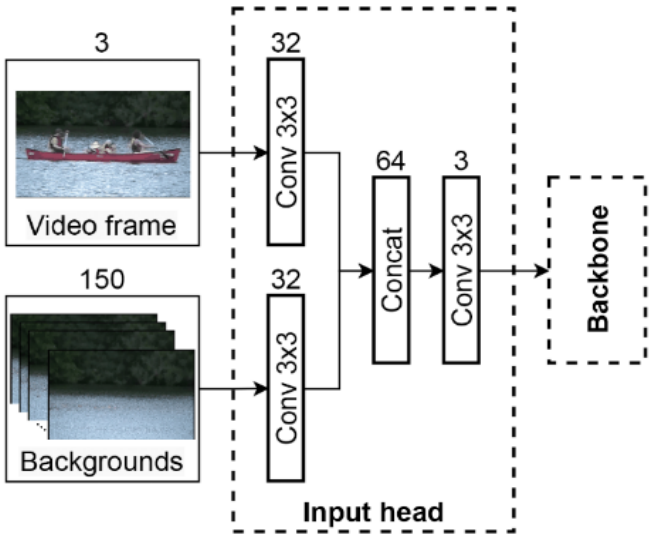

Figure 2: Input head structure of foreground detection RCNN

Figure 2 shows the structure of the input head whose main function is to balance the weights of the video frame and background images. The video frame, an RGB image, comprises three channels, while the backgrounds generated by SuBSENSE consist of 50 RGB background images and 150 channels as a default. Direct concatenation of the video frame and background images as the network input could lead to an imbalance in the network's focus, as it would observe the background images significantly more than the video frames, thus making it challenging to

identify their differences.

To address this, the input head balances the weight between the video frame and background images by converting them into feature maps of the same size. The input head has two branches, with each branch only containing one convolutional layer. The first branch is dedicated to the input frame, and the second is allocated for the 50 background images. Consequently, both the input frame and background images are converted into a set of feature maps with 32 channels. By concatenating these two sets, a 64-channel feature map is created. To avoid any modification on the backbone, a final convolutional layer compresses the 64-channel feature maps to a 3-channel output, which then serves as the backbone input.

No box head strategy

Foreground detection RCNN is devoid of a box head. For foreground detection, the box head can be deleted due to its single-class detection task. In the inference stage, the results of foreground detection can be garnered by filtering the bounding boxes proposed by the RPN. As a result, these foreground boxes serve as the final boxes predicted by the Hybrid-RCNN.

3.2 Object Detection RCNN

In the training phase, the object detection RCNN doesn't require any modification and follows the standard training methods commonly employed by other object detectors.

However, the inference stage presents a different scenario. Here, the object detection RCNN does not retain all network components. Specifically, the RPN, which generates proposal boxes, is set aside and not employed during the inference stage. Instead, only the backbone and RoI heads are kept in operation.

The backbone serves as the key component of the network and is responsible for extracting features from the input images. On the other hand, the RoI heads are involved in classifying and tracking the foreground boxes. They achieve this by cropping the feature maps produced by the backbone using the foreground boxes predicted by the foreground detection RCNN.

Box head

During the inference stage, one role of the object detection RCNN is to classify the foreground boxes predicted by the foreground detection RCNN. As these foreground boxes represent the final boxes, there is no need for further refinement.

Achieving this objective is straightforward. The box head of an RCNN consists of two independent branches: one for box refinement and another for box classification. When the boxes are input into the box head, it generates two outputs: the classification and adjustment parameters for the input boxes. To disable the box refinement function of the box head, it is merely necessary to discard the box adjustment parameters.

Track head

The study utilizes the Global Tracking Transformer (GTR) [9] for foreground tracking. Besides leveraging the cross-attention weights of current and prior objects, GTR also incorporates a box-trajectory IoU tracker. This IoU tracker can boost tracking accuracy, particularly at high frame rates. However, to demonstrate the standalone efficiency of GTR in foreground box tracking, we disable the IoU tracker in the inference stage.

4. EXPERIMENTS

4.1 Datasets

Combined dataset of LVIS and COCO

Both LVIS [10] and COCO [11] datasets are widely used in the field of object detection. The COCO 2017 training and validation sets contain roughly 118,000 and 5,000 images, respectively, across 80 classes. Conversely, LVIS adopts the same image set as COCO 2017 but significantly expands the range of object classes available in COCO by incorporating over 1,203 object classes. Another difference between these two datasets lies in their class distribution. COCO has a relatively balanced instance count across its 80 categories, while LVIS exhibits a "long-tailed" distribution in which a few classes have many instances and many classes have few instances.

The combined dataset of LVIS and COCO was first introduced in the paper of TAO. The authors added COCO annotations to every image in the LVIS dataset. To prevent redundancy, they eliminated COCO annotations that had an IoU greater than 0.7 with an LVIS annotation. They discovered that training on a combination of COCO and LVIS annotations yielded a substantial enhancement in detection quality. This improvement was especially pronounced for the class of people in comparison to training solely on the LVIS dataset.

TAO

TAO [12] is a large-scale benchmark for object tracking in video data. Its training, validation, and testing sets have approximately 500, 1,000, and 1,500 videos, respectively. These videos are captured across a wide range of environments, enhancing the diversity of this benchmark. Each video incorporates 40 annotated frames at the rate of 1 FPS. TAO adopts 488 classes out of the 1,203 available from LVIS and also in the "long-tailed" distribution setting. The diversity of object classes within TAO allows the benchmark to cover a comprehensive range of object tracking scenarios.

CDnet 2014

CDnet 2014 [7] is the largest dataset available for foreground segmentation. It contains 53 videos across 11 distinct categories, each presenting unique challenges such as bad weather, dynamic backgrounds, and night videos. The diversity of CDnet 2014 makes it an ideal choice for algorithm validation. The dataset has 159,278 frames, of which 91,435 have been provided with ground-truth foreground masks. Prior to training, we employed SuBSENSE to preprocess each video, generating background frames for each image.

However, CDnet 2014 wasn't specifically designed for deep learning research. Consequently, it doesn't separate the data into training and setting subsets. Therefore, we randomly divided the labeled images into training and testing sets using a 7:3 ratio. One noteworthy exception is the Pan-Tilt-Zoom (PTZ) category, which we excluded from the training set. This decision was made because, much like most foreground detection models, the Hybrid-RCNN model isn't able to handle panning cameras.

4.2 Training and Evaluation

The training process for the Hybrid-RCNN model is composed of three stages. In the initial two stages, the focus is on training the object detection RCNN. Following this, the foreground detection RCNN will be trained in the third stage.

The training of the object detection RCNN closely follows the methods presented in the paper of the GTR. During the first stage, the object detection RCNN is pretrained exclusively in detection mode utilizing a combined dataset of LVISv1 and COCO.

In the second stage, the track head, also known as the GTR, started participating in the training. Adhering to the training strategy outlined in GTR, static image datasets are employed, augmented

with artificial video clips generated by data augmentation techniques. Specifically, zooming and translation are applied to an image, which then serve as the start and end frames of a video. Through linear interpolation of these images and their respective annotations, a smooth video clip is synthesized. The combined dataset of LVISv1 and COCO continues to be utilized for network training throughout this stage.

The third stage is devoted to training the foreground detection RCNN. Prior to training, a duplicate of the object detection RCNN's backbone is made to serve as the backbone of the foreground detection RCNN. Given that the RPN of the object detection RCNN will be discarded during the inference stage, we directly allocate the RPN of the object detection RCNN to the foreground detection RCNN. Training and evaluation of the foreground detection RCNN is conducted using the CDnet 2014 dataset.

After training, the object detection RCNN yields 34.5 mAP on the LVISv1 validation set and 20.3 mAP on the TAO validation set. On the other hand, foreground detection RCNN achieves 0.90 F-measure on the CDnet 2014 dataset.

4.3 Experiments on the Railroad Crossing Dataset

To demonstrate that the proposed Hybrid-RCNN is better than the object detection types of networks in railroad crossing monitoring, the popular object detection network Mask-RCNN [4], which is well-known for its superior performance, is selected to compare with the proposed Hybrid-RCNN. The Mask-RCNN [4] is trained by the COCO dataset, which has 80 classes.

Figure 3 shows the example detection results of both Hybrid-RCNN and Mask-RCNN. While Mask-RCNN only detects some ordinary objects such as pedestrians and cars, it unfortunately failed to detect a cone and wheelbarrow, resulting in false-negative errors. Also, Mask-RCNN detects many background objects, causing serious false-positive errors. On the contrary, the proposed weakly supervised foreground segmentation network Hybrid-RCNN successfully detects all the objects that do not belong to the railroad crossing as foreground and achieves high precision and recall. Therefore, the proposed Hybrid-RCNN outperforms Mask-RCNN in the railroad crossing monitoring task.

Figure 3: Example detection results on the railroad crossing dataset (from top to bottom: Hybrid-RCNN, Mask-RCNN

5. CONCLUSIONS

In this research, we introduced a novel Hybrid-RCNN model, an innovative approach that integrates the advantages of both foreground segmentation and object detection methodologies, specifically for monitoring railroad crossings. Our design leverages the distinct capabilities of these two techniques to create a more effective and reliable surveillance system. The experimental validation conducted on a dedicated railroad crossing dataset highlights the robust performance of the proposed Hybrid-RCNN network in various conditions. This model is particularly adept at identifying any unauthorized or unexpected objects within the railroad crossing area. Notably, the Hybrid-RCNN model demonstrates superior performance compared to traditional object detection-based models such as Mask-RCNN. The effectiveness of Hybrid-RCNN in detecting anomalies and ensuring safety at railroad crossings is evident from its ability to accurately identify and segment objects that are not typically part of the railroad environment. This enhancement in detection capability is a significant step forward in the application of artificial intelligence in public safety and infrastructure monitoring.

6. REFERENCES

[1] Sikora, P., L. Malina, M. Kiac, Z. Martinasek, K. Riha, J. Prinosil, L. Jirik, and G. Srivastava, Artificial intelligence-based surveillance system for railway crossing traffic. IEEE Sensors Journal, Vol. 21, No. 14, 2020, pp. 15515–15526.

[2] Zaman, A., B. Ren, and X. Liu, Artificial intelligence-aided automated detection of railroad trespassing. Transportation research record, Vol. 2673, No. 7, 2019, pp. 25–37.

[3] Guo, F., Y. Wang, and Y. Qian, Computer vision-based approach for smart traffic condition assessment at the railroad grade crossing. Advanced Engineering Informatics, Vol. 51, 2022, p. 101456.

[4] He, Kaiming, et al. "Mask r-cnn." Proceedings of the IEEE international conference on computer vision. 2017.

[5] Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

[6] Varadarajan, Sriram, Paul Miller, and Huiyu Zhou. "Region-based mixture of gaussians modelling for foreground detection in dynamic scenes." Pattern Recognition 48.11 (2015): 3488-3503.

[7] Wang, Yi, et al. "CDnet 2014: An expanded change detection benchmark dataset." Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2014.

[8] St-Charles, P.-L., G.-A. Bilodeau, and R. Bergevin, SuBSENSE: A universal change detection method with local adaptive sensitivity. IEEE Transactions on Image Processing, Vol. 24, No. 1, 2014, pp. 359–373.

[9] Zhou, Xingyi, et al. "Global tracking transformers." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[10] Gupta, Agrim, Piotr Dollar, and Ross Girshick. "Lvis: A dataset for large vocabulary instance segmentation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[11] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014.

[12] Dave, Achal, et al. "Tao: A large-scale benchmark for tracking any object." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16. Springer International Publishing, 2020.