NHL STENDEN
computer vision & data science

Region-based Anomaly Detection for Cyclist Behavior

NHL Stenden Lectoraat in Computer Vision & Data Science

Jarno Korstanje

Supervisor: Maya Aghaei Gavari

Abstract— Assessing behavior is crucial in bicycle traffic safety research, as it helps improve infrastructure to promote the use of bicycles as a safe mode of transport. However, current manual methods for analyzing behavior in video are time-consuming and require significant expertise. Therefore, there is a need for efficient automated tools. Our goal is to automatically extract the most interesting parts of long videos through video anomaly detection, enabling researchers to focus on detailed analysis on shorter extracted videos. We propose a novel future frame-based video anomaly detection strategy using the Dynamic Multi-scale Voxel Flow Network (DMVFN) frame prediction. DMVFN is a state-of-the-art frame prediction model that learns to synthesize video frames by flowing pixel values from previous frames, called voxel flow. Additionally, handcrafted segmentation masks are applied to focus on the relevant regions in traffic video, such as bike lanes. We evaluate our methods against the previous U-Net frame predictor on two proprietary datasets, Trafficcam and Railcam, focusing on cyclist behaviour in traffic video. The segmentation masks significantly improved anomaly detection results, with an overall increase in AUC score across our datasets. Our novel approach using DMVFN as the frame predictor achieved mixed results compared against the previous U-Net frame predictor. Furthermore, qualitative results indicated that all models effectively distinguished between normal and abnormal movements on heatmaps, showing the explainability capabilities of this approach.

Index Terms—Cyclist Safety, Traffic Behavior, Video Anomaly Detection, Future Frame Prediction, Region-based, Voxel Flow

1 Introduction

The decline in fatal bicycle accidents has slowed in the Netherlands, with an alarming increase in victims among those aged 75 and over [1, 2, 3, 4]. In recent years, a rise in electric modes of transport has made traffic more complicated, particularly with the growing popularity of e-bikes. This increase has contributed to greater complexity in cyclists' behavior [5].

Behavioral psychology plays a pivotal role in bicycle traffic safety research. Road safety researchers analyze complex cyclist behavior, enabling them to propose recommendations to policymakers regarding necessary infrastructure adjustments. Unfortunately, current methods require a high degree of expertise, are time-consuming, and present challenges in scalability. Therefore, there is an urgent demand for efficient software tools that can facilitate the analysis of cycling behavior by road safety researchers.

Current automated solutions for analyzing road safety are often limited in their ability to assess behavior. They mainly focus on counting and tracking road users, without analyzing the complexity of bicycle traffic. Since current automation tools are often inadequate, a significant portion of behavioral analysis remains manual. We argue that part of the solution lies in automatically extracting the most potentially interesting sections from a video. This would enable researchers to skip the manual video scanning step and focus more quickly on analyzing a smaller part of the video.

Therefore, we propose the use of video anomaly detection to extract the abnormal and potentially interesting sections of traffic video. We introduce a novel strategy that combines previous work that uses future frame prediction for this task and modify it with a recent state-of-the-art frame predictor, which has not yet been applied to this task. Additionally, our strategy applies segmentation masks to focus on anomalous movement in relevant image regions, such as bike lanes. Furthermore, we conduct experiments on traffic video, showcasing the practical use of the solution.

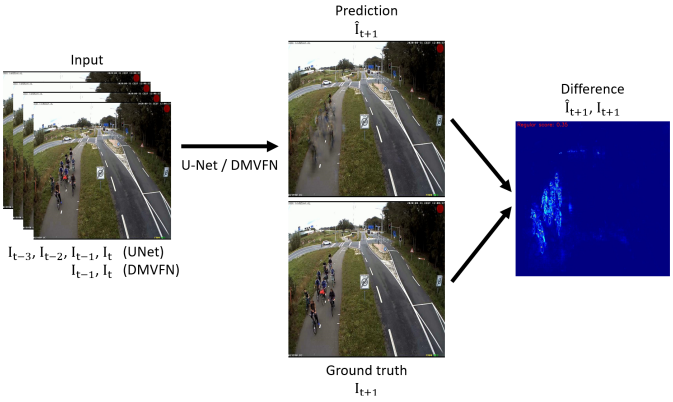

Fig. 1. Overview of the general approach following [6]. Based on movement in previous images, we predict image t+1 and compare it to the ground truth. Movement that the model has not learned will show a larger difference between the prediction and the ground truth, indicating potential abnormal movement. Our novel approach replaces the original U-Net frame predictor with the current state-of-the-art frame prediction method, the Dynamic Multi-Scale Voxel Flow Network (DMVFN).

Our research contributes to the advancement of traffic behavior understanding, offering a promising solution for traffic researchers to propose effective infrastructure adjustments. Ultimately, this can stimulate the use of bicycles as a safe and environmentally friendly alternative to motorized transport.

The main contributions are summarized as follows:

- We propose a novel video anomaly detection approach that combines the future frame-based method and a recent state-of-the-art frame prediction model, the Dynamic Multi-Scale Voxel Flow Network (DMVFN).
- Our approach introduces handcrafted segmentation masks to focus on behavioral anomalies in the relevant regions of traffic video to improve prediction performance.

- *Jarno Korstanje is a Computer Vision & Data Science student at NHL Stenden University of Applied Sciences, E-mail: jarno.korstanje@student.nhlstenden.com.*
- *Maya Aghaei Gavari is a teacher researcher at the NHL Stenden Lectoraat in Computer Vision & Data Science, E-mail: maya.aghaei.gavari@nhlstenden.com.*

1.1 Related work

In this research, we explore the application of computer vision techniques to assess the safety of cyclist behavior. In this section, we discuss related work on the safety factors regarding cyclist behavior and the application of computer vision techniques in this context. Furthermore, we explore various approaches for video anomaly detection. Since we propose to apply a recent frame prediction network for anomaly detection purposes, we explore related work on video frame prediction.

1.1.1 Bike safety factors

Cycling accidents often result from a combination of infrastructural and behavioral factors [3]. Infrastructural safety factors include obstacles, bike path width, degree of bike path occupancy, and lighting conditions [3]. Behavioral safety factors include cycling speed, relative speed to other road users, lateral position relative to the center and edge of the bike path, junction crossing, left turns, failure to yield right of way, use of bike lighting, disregarding red traffic lights, cycling against traffic, and phone usage [3].

1.1.2 Bike safety monitoring using computer vision

Previous research has been done on assessing bike safety risks using computer vision. In [7], a method is presented to identify objects that pose potential cycling risks from Google Street View images in London, United Kingdom. For scene understanding, a YOLOv5 model was used to detect objects such as cars, buses, and trucks as potential safety hazards for cyclists. A PSPNet101 model was used to segment structures such as roads, sidewalks, bike lanes. Both models were used to gather statistics and assess the cycling infrastructure's safety across locations in London. Approximately the same method was applied in another study by the same authors in Lisbon, Portugal [8]. In this study, the PSPNet101 model was replaced by a segmentation model from NVIDIA [9]. Additionally, a Multilayer Perceptron model was used to assign a safety score to each image using the outputs from the prior models. Therewith, they aim to provide risk information for cyclists, decision-makers and other stakeholders. A current limitation of their research is that static Google Streetview images cannot capture additional features such as pedestrian, cyclist and vehicle movement.

To correctly assess bicycle infrastructure we argue that one must not solely inspect the infrastructure itself, but also monitor the events happening. Hence, our goal is to assist traffic safety researchers by developing a method to automatically identify irregular movements, which include swerving and cycling on the opposite side of the bike lane. This will help traffic safety researchers monitor a situation based on actual cycling behavior. Additionally, our approach involves the application of handcrafted segmentation masks to focus on relevant regions within images, such as a bike lane. This opens a future possibility to replace handcrafted masks with segmentation techniques. These technical concepts are further explored in the next subsections.

1.1.3 Semantic segmentation of traffic scenes

Semantic segmentation is a computer vision task of classifying each pixel in an image into a specific class or category [10]. It provides a more detailed understanding of the image's content at the pixel level compared to bounding-box based object detection. Semantic segmentation can aid in higher-level reasoning tasks, such as scene understanding, action recognition and event detection.

Mask2Former [11] is a segmentation model with a unified architecture for all image segmentation tasks (semantic, instance or panoptic segmentation). It uses masked attention, which constrains cross-attention within predicted mask regions to extract features. Mask2Former achieves a 84.3 mIoU on Cityscapes val and 64.7 mIoU on Mapillary val using an Swin-L [12] backbone. Although Mask2Former can address any segmentation task, it still needs to be trained separately on those different tasks. Oneformer [13] addresses this limitation and provides a multi-task train-once architecture for all segmentation tasks. It achieves a 84.6 mIoU on Cityscapes val using an ConvNeXt-XL [14] backbone and 64.9 mIoU on Mapillary val using an DiNAT-L [15] backbone. Both the Mapillary and Cityscapes datasets contain traffic scenes, which shows potential to replace handcrafted masks with a segmentation model in a later stadium.

1.1.4 Video anomaly detection

The detection of anomalies in videos refers to recognizing events that deviate from expected behavior. Traditionally, the most common deep learning approaches have been reconstruction-based methods. Many of these approaches utilize an autoencoder [16, 17, 18], which is trained to encode an image into a feature representation and reconstruct the feature representation into the original image. Since it is trained on normal instances, it is assumed that the reconstruction error would be higher for abnormal instances, enabling to distinguish between normality and abnormality. However, reconstruction errors for abnormal events do not necessarily occur, as abnormal events can be identified not only by appearance but also by motion.

Future frame-based methods aim to address the challenge of detecting anomalous movement. This method involves training a frame predictor on only normal movement. Given previous frames, the future frame is predicted and compared against the actual frame. Since it is trained on normal movement, abnormal movements are predicted worse and give a higher error when compared to the actual frame. Therefore, a high frame error can be used to detect anomalous movement on frame level. To predict the next frame, Generative Adversarial Networks (GANs) [19] are frequently used. It is a generative model that consists of a generator and a discriminator, which are trained simultaneously in a competitive manner. The generator learns to generate new frames while the discriminator learns to classify the generator's fake data from real data. The generator is trained to produce samples that the discriminator is more likely to classify as real while the discriminator gets better at distinguishing between real and fake samples.

The method in [6] utilized the U-Net architecture [20] as the generator. For the discriminator, the PatchGAN architecture [21] is utilized, which classifies each N×N patch and enables faster processing. Alongside the discriminator loss, an intensity loss and gradient loss are incorporated as appearance constraints to produce high-quality frames. However, appearance constraints do not ensure accurate motion prediction. To address this, the optical flow is calculated using FlowNet2 [22] for both the current and predicted frames, as well as for the current and actual frames. The difference between the two outputs is introduced as the optical flow loss. Finally, the difference between the predicted and actual frames determines whether the given frame is normal or abnormal. The approach was evaluated on the UCSD Ped2 [23], CUHK Avenue [24] and ShanghaiTech Campus [6] datasets, showing the ability to detect abnormal movement such as people throwing objects or cyclists and golf carts in a pedestrian area.

1.1.5 Video frame prediction

In this research, we propose to improve the existing video anomaly detection algorithms by replacing the existing frame predictor with a state-of-the-art frame prediction network. This section explores recent advancements in video frame prediction.

The conventional method for video frame prediction involves estimating the optical flow between frames and then extrapolating along these flow vectors. This technique performs well when the estimated flow is precise, but it can produce artifacts when the estimated flow is inaccurate. Later approaches used generative convolutional neural networks (CNNs) to directly predict the future frame from previous frames [25]. However, this approach is less successful than methods utilizing optical flow, as it often produces blurry results.

The work of [26] proposed Deep Voxel Flow, which combines the strengths of both approaches. It employs a convolutional encoder-decoder with a so-called voxel flow layer, a 3D flow field across the two input frames. The future frame is predicted by extrapolating pixel values from previous frames, similarly to optical

flow approaches. Since the voxel flow layer is an intermediate layer, it is not evaluated directly. This method was an improvement upon both optical flow and CNN approaches for video frame prediction.

The Dynamic Multi-Scale Voxel Flow Network (DMVFN) [27] introduces a routing module that activates different sub-networks based on the input frames, which improves both prediction accuracy and inference speed. DMVFN outperforms previous state-of-the-art frame prediction methods at lower computational costs and achieves speeds an order of magnitude faster than Deep Voxel Flow.

2 Materials and Methods

2.1 Anomaly detection approach

As shown in Figure 1, our approach closely follows the methodology described by Liu et al. [6], where predicted frames are compared against ground-truth frames to detect anomalies. The existing implementation employs a Generative Adversarial Network (GAN) with a U-Net model as the generator. We test this architecture at a resolution of 960x960, an increase from the 256x256 resolution used in the original paper.

Since we evaluate the model on a larger resolution, we argue that the U-Net model could benefit from a larger capacity. We propose to improve the existing U-Net architecture by enlarging its capacity through the addition of more layers. Furthermore, we introduce a novel approach that completely replaces the U-Net frame predictor with the latest state-of-the-art frame prediction model, the Dynamic Multi-Scale Voxel Flow Network (DMVFN). We experiment with three different frame prediction models:

- U-Net - 6 layers (Original)
- U-Net - 8 layers
- DMVFN

2.1.1 U-Net

We experiment on a 6-layer and 8-layer variant of the U-Net model, which are further described in Appendix A. Both variants take four resized 960x960 input images that are stacked, resulting in 12 channels. For both variants, the U-Net frame generator is trained in an adversarial manner with a PatchGAN [21] discriminator. The discriminator learns to classify between real and generated images. During the training of the discriminator, we keep the generator's weights fixed and use a Mean Square Error (MSE) loss. During the training of the generator, we keep the discriminator's weights fixed.

The U-Net generator is trained using an objective function that combines four different losses. Firstly, the adversarial loss, derived from the output of a discriminator network, aims to generate realistic data samples that are indistinguishable from real samples. Secondly, optical flow estimation is used to improve the motion prediction. A pre-trained FlowNet2-SD [22] model calculates the optical flow between the current frame and the predicted frame, as well as between the current frame and the ground-truth frame. This results in the flow loss, which measures the mean absolute difference between the two flow fields. To make the generated frame close to its ground truth, we use L2 distance as the intensity loss, which enforces the similarity of pixels in RGB space. To encourage sharpness in the generated frames, a gradient loss penalizes differences in vertical and horizontal gradients between the predicted and ground truth frames.

In the objective function of the U-Net generator, the four losses are represented with corresponding coefficients as follows:

- Adversarial loss: 0.05
- Flow loss: 2
- Intensity loss: 1
- Gradient loss: 1

An Adam [28] optimizer is used, where the learning rates are set to 0.0002 for the U-Net generator and 0.00002 for the PatchGAN discriminator. In the experiments, we use a PyTorch implementation with automatic mixed precision (AMP) in addition to the original implementation.

2.1.2 Dynamic Multi-Scale Voxel Flow Network

In contrast to U-Net, DMVFN takes only two resized 960x960 input images. The core of the DMVFN architecture are the Multi-scale Voxel Flow Blocks (MVFB). Each MVFB learns to estimate the target voxel flow by taking two consecutive frames, the synthesized target frame and the voxel flow estimated by previous blocks as inputs.

Each MVFB has a scaling factor S, which means that the motion path is performed on images whose sizes are 1/S of the original. DMVFN contains 9 MVFB blocks with decreasing scaling factors [4, 4, 4, 2, 2, 2, 1, 1, 1]. Since each MVFB refines the current voxel flow estimation to a new one, some MVFBs can be skipped. In order to choose which MVFB blocks are activated, a Routing Module generates a routing vector. The Routing Module takes two frames as input into several neural layers and Differentiable Bernoulli sampling.

During training, Laplacian pyramid representations are extracted from the output and target image for every MVFB block. An ℓ_1 loss is calculated from the extracted representations as the reconstruction loss. A total ℓ_1 loss is calculated by summing the reconstruction losses of each block. The final training loss is a combination of the total ℓ_1 loss and VGG loss. DMVFN uses an AdamW [29] optimizer with a weight decay of 10^{-4}. Additionally, a cosine annealing strategy is applied to reduce the learning rate from 10^{-4} to 10^{-5} throughout the training process. In the experiments, we use a PyTorch implementation with automatic mixed precision (AMP) in addition to the original implementation.

2.2 Datasets

The performance of the models is assessed on Trafficcam and Railcam, two proprietary datasets. Example images of both datasets are shown in Figure 2. Both datasets are divided into training, validation, and testing splits, with each subset containing multiple videos.

2.2.1 Trafficcam

The Trafficcam dataset is a proprietary dataset containing video of a traffic intersection from a CCTV camera nearby Leusden, the Netherlands. It was specifically created in September 2020 for a pilot research project into traffic safety indicators. The dataset contains 3098 frames with a resolution of 800x608 and has a frame rate of 5 fps. The selected abnormal instances are relatively abnormal to the rest of the dataset, with examples being large groups of cyclists and a fast scooters. It is split into 1818 normal frames for training, 620 frames for validation and 660 frames for testing. The validation and testing frames are annotated on frame level. Validation has 436 normal frames and 184 abnormal frames, while testing has 443 normal frames and 217 abnormal frames.

2.2.2 Railcam

The Railcam dataset is another proprietary dataset, containing livestream video of a traffic intersection near the railway crossing in Mierlo-Hout, the Netherlands [30]. The dataset contains 6603 frames with a resolution of 1920x1080 and has a frame rate of 25 fps, which results in a clearer movement compared to the Trafficcam dataset. The video was captured with a Milesight H.265 Mini PTZ Bullet network camera. The dataset focuses on the abnormality of cyclists going into the opposite direction, which can potentially result in dangerous situations. It is split into 3690 normal frames for training, 1174 frames for validation and 1739 frames for testing. The validation and testing frames are annotated on frame level. Validation has 545 normal frames and 629 abnormal frames, while testing has 1221 normal frames and 518 abnormal frames.

2.3 Masking approach

In this work, our focus is to recognize anomalous behavior in the bike lane. Therefore, we anticipate that car movements may negatively impact the results by causing false positives. To counter that, our approach focuses solely on the relevant areas through a segmentation mask. As a feasibility study, we initially experiment with a

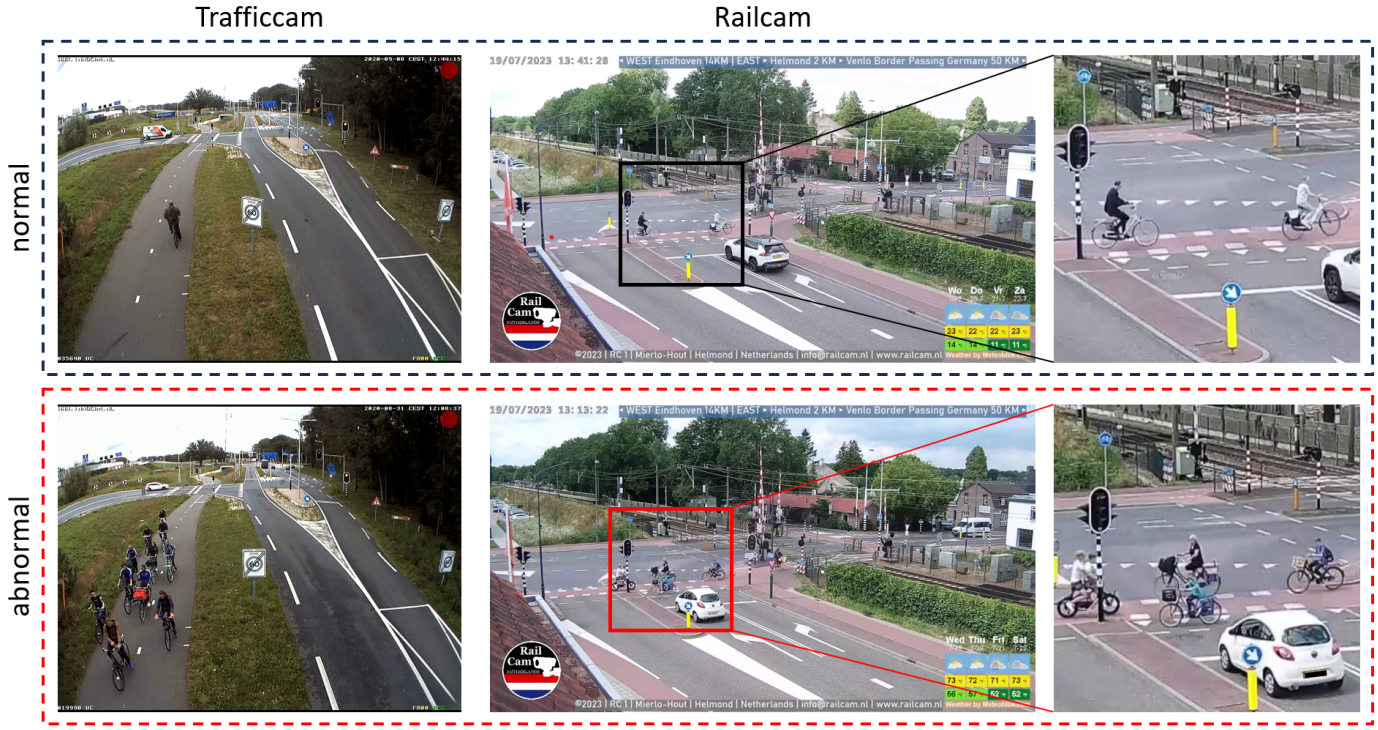

Fig. 2. Example images of the Trafficcam and Railcam datasets, showing both a normal and abnormal example.

handcrafted mask to determine if a semantic segmentation model could potentially improve anomaly detection.

We create a handcrafted mask that covers the bike lane and its surrounding area. During the masking process, we zero out the pixels outside the mask, preserving only the relevant area. This produces two additional masked versions of the Trafficcam and Railcam datasets, as shown in Figure 3. In our experiments, we compare the masked versions with the original counterparts to evaluate the potential benefits of semantic segmentation for anomaly detection.

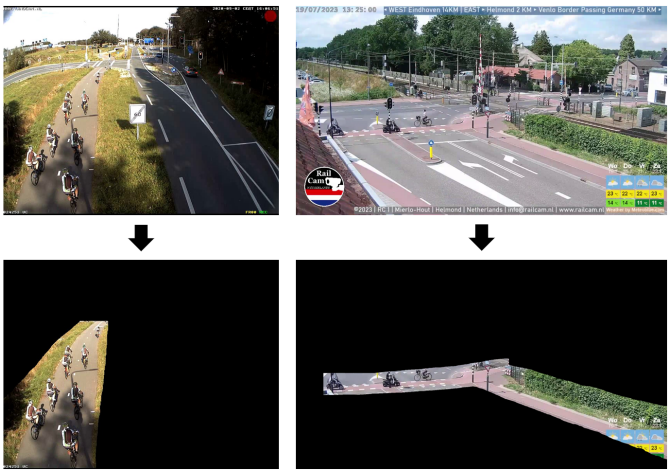

Fig. 3. Original and masked versions of the Trafficcam and Railcam datasets. Both masked versions are produced using a handcrafted static mask.

2.4 Metrics

Since the frame predictor is trained on normal movements, it will perform well in predicting normal movement. As [6] has shown, abnormal movement will appear unclear in the generated frames. Therefore, the image quality of the predicted frame is compared against the actual frame for anomaly detection. When predicting a frame, the corresponding image quality is simultaneously calculated using the Peak Signal to Noise Ratio (PSNR):

$$PSNR(I,\hat{I}) = 10\log_{10}\frac{[\max_{\hat{I}}]^2}{\frac{1}{N}\sum_{i=0}^{N}(I_i-\hat{I}_i)^2}$$

Here, I and $\hat{I}$ are the original and predicted frames. N and represents the total number of pixels, where i is the index running over the pixels.

After all frames and their PSNR values in the testing set are predicted, we calculate a regular score is by normalizing the PSNR values between 0 to 1 for the testing set. This is different to the original approach of [6], which normalized the PSNR values over a single video. The reason for this is that our testing datasets also contain videos with only normal movement. By normalizing over the whole testing set, we prevent a low regular score for a normal video. We predict whether a frame is normal or abnormal based on a regular score using a threshold. The regular score for a single frame is calculated with the following equation:

$$S(t) = \frac{PSNR(I_t,\hat{I}_t) - \min_t PSNR(I_t,\hat{I}_t)}{\max_t PSNR(I_t,\hat{I}_t) - \min_t PSNR(I_t,\hat{I}_t)}$$

Here, t is the index of the frame.

The regular score is thresholded to differentiate normal from abnormal frames. We calculate the Receiver Operating Characteristic (ROC) curve, which evaluates the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) for different regular score thresholds. To summarize the overall performance of the model, we use the Area Under the Curve (AUC) metric. A higher AUC means better performance, with 1 being a perfect classifier and 0.5 being random guessing.

3 Experiments & Results

In this section, we evaluate our novel approach using DMVFN as a frame predictor on two proprietary anomaly detection datasets, Trafficcam and Railcam. We compare our method to the original

U-Net frame predictor, both in its default 6-layer configuration and an expanded 8-layer variant. Additionally, we test our proposed masking technique, which focuses on movement in the relevant region of the traffic scene using a handcrafted mask. For the masked experiments, the models were both trained and tested on the masked versions.

3.1 Trafficcam

All model versions are trained on a load-balanced NVIDIA A-40 GPU with 16 GB of allocated memory, using a batch size of 2 for 90,000 iterations. Training took approximately 24 hours for U-Net and 48 hours for DMVFN. The best result is saved based on the AUC score on the validation set.

Quantitative results. The quantitative results are reported in Table 1. Masking the frames led to an average AUC score increase of 22.9%. Among the U-Net frame predictors, the original 6-layer variant showed the greatest improvement with the masking applied. Additionally, DMVFN outperformed both U-Net models in the original variant as well as the masked variant of the Trafficcam dataset.

Model	Masked	AUC (Trafficcam)	AUC (Railcam)
U-Net (6 layers)	No	0.482	0.857
	Yes	0.616	0.908
U-Net (8 layers)	No	0.513	0.859
	Yes	0.590	**0.915**
DMVFN	No	0.525	0.840
	Yes	**0.662**	0.820

Table 1. Results of the U-Net (6 and 8 layers) and DMVFN frame predictors on the Trafficcam and Railcam anomaly detection datasets, with and without the handcrafted mask applied.

Qualitative results. The qualitative results for the masked variant of Trafficcam are reported in Figure 4. There are minimal visible differences among the three models. All models accurately predict normal movement, indicated by low intensity on the heatmaps. As expected, they correctly struggle to predict the fast scooter and the group of cyclists, as these are anomalous behaviors, resulting in higher intensity on the heatmaps. However, all models suffer from false negatives for anomalous movement in the background and false positives for normal behavior in the foreground.

3.2 Railcam

All model versions are trained on a load-balanced NVIDIA A-40 GPU with 16 GB of allocated memory, using a batch size of 2 for 70,000 iterations. Training took approximately 22 hours for U-Net and 32 hours for DMVFN. The best result is saved based on the AUC score on the validation set.

Quantitative results. The quantitative results are reported in Table 1. Overall, the masked version of the dataset shows mixed results compared to the original, with an average AUC score increase of 3.4%. Notably, DMVFN performs worse than both U-Net models. Additionally, the 8-layer U-Net model offers a slight improvement over the 6-layer version.

Qualitative results. The qualitative results for the masked variant of Railcam are reported in Figure 5. Both U-Net models accurately predicted the frames for a scene with only normal behavior. However, DMVFN struggled more with predicting normal behavior, displaying higher intensity on the heatmap. In scenes with abnormal behavior, all models had difficulty predicting the movement, which helps classify the scene as abnormal. In scenes containing both normal and abnormal behavior, all models successfully differentiated between the behaviors, aiding in identifying the contributing factors.

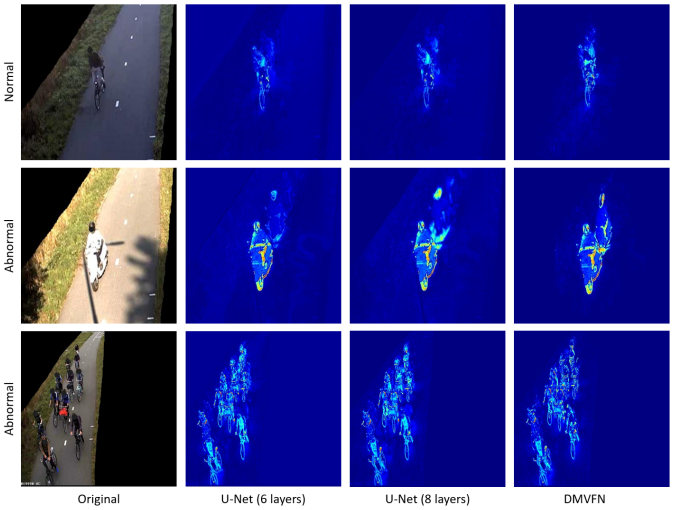

Fig. 4. Qualitative results for masked version of the Trafficcam dataset, showing the original frame and the difference between the predicted frame and the original frame for U-Net (6 and 8 layers) and DMVFN. The top row shows a normal scene (single cyclist) and two abnormal scenes in the middle and bottom row (a fast scooter and a large group of cyclists).

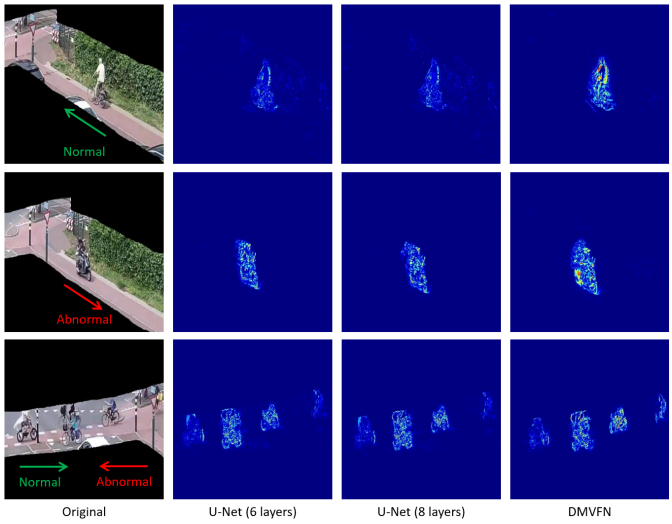

Fig. 5. Qualitative results for masked version of the Railcam dataset, showing the original frame and the difference between the predicted frame and the original frame for U-Net (6 and 8 layers) and DMVFN. It shows a normal scene (top row), abnormal scene (middle row) and a scene containing both normal and abnormal behavior (bottom row).

4 Discussion, Conclusion & Future Work

4.1 Discussion

In this work, we proposed the use of the Dynamic Multi-scale Voxel Flow Network (DMVFN) for future frame-based video anomaly detection of cyclist behavior. We compared its performance against the previous U-Net frame predictor, both as the 6-layer original and 8-layer expanded variant. Additionally, we applied handcrafted segmentation masks to focus on the relevant regions in traffic video.

The results on the Trafficcam dataset were significantly worse compared to the Railcam dataset across all models. This discrepancy may be due to the less distinct boundaries between normal behavior (single cyclists) and abnormal behavior (groups of cyclists and fast scooters) in the Trafficcam footage. Additionally, anomalies occurring in the background are not easily detected, while normal

behavior in the foreground is often incorrectly flagged as anomalous. In contrast, the Railcam dataset focuses on detecting cyclists riding in the opposite direction, which deviates more from the learned behavior. Furthermore, all movements in the Railcam dataset occur at approximately the same depth within the scene, which may also contribute to an overall higher detection accuracy. There was also a missed opportunity to use the Railcam dataset in its full resolution instead of the reduced 960x960 resolution.

Expanding the original U-Net model with two additional layers resulted in a slight improvement and are especially minimal on the Railcam dataset. Considering the experiments were conducted only once, they might be considered negligible. Our proposed method, which replaces U-Net with the DMVFN frame predictor, showed mixed results. It outperformed the U-Net models on the Trafficcam dataset but underperformed on the Railcam dataset. To draw a definitive conclusion, further experiments using real-world data are necessary.

The application of handcrafted masks enhanced performance across both datasets. However, it is notable that masking did not lead to improved results for DMVFN on the Railcam dataset, though the reason for this remains unclear. Nonetheless, given the success of the handcrafted masks, there is potential in employing semantic segmentation models for further automation.

4.2 Conclusion

On the Trafficcam dataset, DMVFN showed slight improvements over both U-Net variants. However, on the Railcam dataset, DMVFN performed slightly worse compared to the U-Net models. Qualitative results demonstrated that all models effectively distinguished between normal and abnormal movements on the heatmap, highlighting the models' explainability. To draw definitive conclusions, further experiments with real-world data are necessary. Handcrafted segmentation masks significantly improved anomaly detection results, with an average AUC score increase of 22.9% on the Trafficcam dataset and 3.4% on the Railcam dataset. This suggests further opportunities for using semantic segmentation models to automate the process. Overall, our approach to detecting anomalous cyclist behavior shows promise in automatically identifying interesting sections of long traffic videos. Ultimately, this can help traffic researchers to focus on smaller videos for further analysis to make necessary infrastructure adjustments to improve safety for cyclists.

4.3 Future Work

Building on the findings of this study, future research could focus on several areas. Firstly, the scope of evaluation could expand by testing our proposed method on public datasets. While initial experiments used proprietary datasets, validating the effectiveness of our method with DMVFN on publicly available datasets would demonstrate its broader applicability.

To eliminate the need for handcrafted segmentation masks, further experiments could utilize a semantic segmentation model. As discussed in Section 1.1, one such model is Mask2Former. It comes with pre-trained weights for the Mapillary Vistas dataset, making it capable of segmenting bike lanes.

There is additional potential in customizing the DMVFN architecture for anomaly detection purposes. Although DMVFN was initially designed to perform well on datasets containing large-scale movement in every part of a frame, our videos are captured by a static camera, resulting in motion occurring in relatively small areas of the frame. A known trade-off in DMVFN's motion scaling factors is that a higher scaling factor provides a larger receptive field but with lower spatial resolution, while a lower scaling factor offers a smaller receptive field but with higher spatial resolution [27]. Therefore, we recommend conducting further experiments with lower motion scaling factors, as these should be more effective in capturing smaller movements.

Acknowledgements

- This project is financially supported by Regieorgaan SIA (part of NWO) under the RAAK Publiek project BikeSafeAI.
- We sincerely thank Railcam Netherlands for the video footage.





References

[1] Centraal Bureau voor de Statistiek (CBS). More traffic deaths in 2022, particularly among cyclists over 75, 2023.

[2] J Oude Mulders, L.T. Aarts, R.J. Decae, N.M. Bos, Ch. Goldenbeld, G. Schermers, C. Mons, A.T.G. Hoekstra, and F.D Bijleveld. Backgrounds of the state of road safety 2023; the annual monitor. 2023.

[3] J.P. Schepers, W.A.M. Weijermars, M.J. Boele, A. Dijkstra, and N.M. Bos. Older cyclists; crashes involving older cyclists and contributory factors. 2020.

[4] Centraal Bureau voor de Statistiek (CBS). 684 road traffic deaths in 2023, 2024.

[5] Tibor Petzoldt, Katja Schleinitz, Sarah Heilmann, and Tina Gehlert. Traffic conflicts and their contextual factors when riding conventional vs. electric bicycles. *Transportation research part F: traffic psychology and behaviour*, 46:477–490, 2017.

[6] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6536–6545, 2018.

[7] Luís Rita, Miguel Peliteiro, Tudor-Codrin Bostan, Tiago Tamagusko, and Adelino Ferreira. Using deep learning and google street view imagery to assess and improve cyclist safety in london. *Sustainability*, 15(13):10270, 2023.

[8] Tiago Tamagusko, Matheus Gomes Correia, Luís Rita, Tudor-Codrin Bostan, Miguel Peliteiro, Rodrigo Martins, Luísa Santos, and Adelino Ferreira. Data-driven approach for urban micromobility enhancement through safety mapping and intelligent route planning. *Smart Cities*, 6(4):2035–2056, 2023.

[9] Andrew Tao, Karan Sapra, and Bryan Catanzaro. Hierarchical multi-scale attention for semantic segmentation. *arXiv preprint arXiv:2005.10821*, 2020.

[10] Yanming Guo, Yu Liu, Theodoros Georgiou, and Michael S Lew. A review of semantic segmentation using deep neural networks. *International journal of multimedia information retrieval*, 7:87–93, 2018.

[11] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022.

[12] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[13] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2989–2998, 2023.

[14] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[15] Ali Hassani and Humphrey Shi. Dilated neighborhood attention transformer. *arXiv preprint arXiv:2209.15001*, 2022.

[16] Jinwon An and Sungzoon Cho. Variational autoencoder based anomaly detection using reconstruction probability. *Special lecture on IE*, 2(1):1–18, 2015.

[17] Chong Zhou and Randy C Paffenroth. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 665–674, 2017.

[18] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1705–1714, 2019.

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[22] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.

[23] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):18–32, 2013.

[24] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013.

[25] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.

[26] Ziwei Liu, Raymond A Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *Proceedings of the IEEE international conference on computer vision*, pages 4463–4471, 2017.

[27] Xiaotao Hu, Zhewei Huang, Ailin Huang, Jun Xu, and Shuchang Zhou. A dynamic multi-scale voxel flow network for video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[30] Railcam Netherlands. Railcam netherlands. `https://railcam.nl/`, 2024. Accessed: 2024-06-05.

A U-NET ARCHITECTURE VARIANTS

A.1 6-layer U-Net variant

- Convolution kernel size: 3 x 3
- Max pooling kernel size: 2 x 2
- Encoder layers:
    1. Downsample 64 channels to 128
    2. Downsample 128 channels to 256
    3. Downsample 256 channels to 512
- Decoder layers:
    1. Upsample 512 channels to 256
    2. Upsample 256 channels to 128
    3. Upsample 128 channels to 64

A.2 8-layer U-Net variant

- Convolution kernel size: 3 x 3
- Max pooling kernel size: 2 x 2
- Encoder layers:
    1. Downsample 64 channels to 128
    2. Downsample 128 channels to 256
    3. Downsample 256 channels to 512
    4. Downsample 512 channels to 1024
- Decoder layers:
    1. Upsample 1024 channels to 512
    2. Upsample 512 channels to 256
    3. Upsample 256 channels to 128
    4. Upsample 128 channels to 64