NHL STENDEN
computer vision & data science

Interpretable Video Anomaly Detection for Enhancing Cyclist Safety

NHL Stenden Lectoraat in Computer Vision & Data Science

Çağkan Gürsoy

Supervisor: Maya Aghaei Gavari

Abstract— Cyclist safety is a critical concern in urban environments, involving interactions with other road users such as cyclists, vehicles, pedestrians, and objects on the road. Monitoring cyclists' behavior in such a setting is essential to identify risks and requires innovative approaches to mitigate them in real time. This research investigates the application of video anomaly detection (VAD) algorithms to identify unsafe cyclist behaviors and interactions in urban settings. Current manual methods for analyzing cyclist behavior in video footage require significant expertise and are not scalable for large datasets. To address these challenges, we propose an automated system that leverages VAD to efficiently extract the most relevant and anomalous segments or objects from long video recordings, enabling researchers and traffic safety professionals to focus on detailed analysis of shorter, high-priority clips. This study builds on the methodology of Reiss and Hoshen's work on attribute-based representations, which demonstrated the effectiveness of combining interpretable features with deep embeddings to enhance anomaly detection in video data. Our approach integrates velocity as an interpretable feature alongside deep representations to achieve robust and interpretable anomaly detection. Object detection using an off-the-shelf detector and optical flow estimation were employed to extract these features, which were then analyzed using density-based methods. Unique models were trained for each feature type: a Gaussian Mixture Model (GMM) for velocity features and k Nearest Neighbors (kNN) for deep representations. The anomaly scores from these models were aggregated to produce a final anomaly score, enabling detection at both frame and object levels. We have retrained and evaluated this method on a real-world dataset containing footage of the cyclist path and the road, annotated with cyclist safety anomalous labels. Notably, the proposed method achieved an AUC-ROC of 94.6% in frame-level evaluations. However, object-level evaluations yielded an AUC-ROC of 55.5%. This decline can be primarily attributed to the use of an off-the-shelf object detector, which introduces excessive noise and inaccuracies in feature extraction. Addressing these limitations remains a priority for future research.

Index Terms—Cyclist Safety, Video Anomaly Detection, Deep Learning, Density Estimation, Accurate and Interpretable Features.

1 Introduction

The safety of cyclists is a critical global concern, and various influential factors are involved, such as cultural attitudes and infrastructural resource availability. The interaction between cyclists and other parties, such as the environment, vehicles, and other cyclists, presents significant risks [1]. To approach the cyclist safety issue comprehensively, aspects involving legislation, enforcement, behavioral measures, and the design of the built environment should be analyzed thoroughly [2].

Previous research has shown that the quality of the infrastructure plays a crucial role in the safety of cyclists. Studies indicate that having separate cycling tracks presents the lowest risk of accidents compared to other types of roads [3]. However, there is also a psychological aspect: the behavior of cyclists significantly impacts safety outcomes. Risky behaviors and traffic rule violations are major factors contributing to accidents [1].

The focus area, the Netherlands, offers crucial insights into both success cases and challenges faced in promoting cyclist safety. Despite having extensive infrastructural availability and a strong cultural emphasis on cycling safety, the country still faces several safety issues. In 2022, nearly 40% of road fatalities were cyclists, which translates to approximately 300 deaths per year [4]. To further enhance cyclist safety, continuous efforts are necessary to address ongoing and new challenges such as infrastructure issues, increasing rates of fatalities among older cyclists, and adapting intelligent transportation systems (ITS) for bikes [1, 3, 5].

Çağkan Gürsoy is a Computer Vision Data Science student at the NHL Stenden University of Applied Sciences, E-mail: cagkan.gursoy@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.

An effective approach to improve cyclist safety and promote cycling as a healthy, sustainable mode of transportation can be the incorporation of computer vision systems and video anomaly detection. This technology can be used with footage obtained from city surveillance systems to create safer cycling environments by complementing traditional safety measures. Such systems can be deployed to detect and track cyclists automatically and monitor their interactions with all surrounding elements, including vehicles, pedestrians, infrastructure, and the environment. By doing so, they can generate valuable data that traffic safety professionals can use [6]. This data includes unsafe behaviors, potential conflict points, and indicators of impending incidents that may not be easily detected by traditional methods. Video anomaly detection algorithms identify unusual or unexpected activities or situations that deviate from normal patterns, which could signal unsafe conditions or risks requiring attention. Integrating video anomaly detection into urban planning and traffic management systems can inform authorities about approaching risks or recurring safety issues [7]. Moreover, continuous monitoring of cyclists' actions can be useful to evaluate the effectiveness of the implemented safety measures, which makes room for an iterative strategy to improve cyclist safety. However, existing video anomaly detection methods often face challenges such as scalability and interpretability [8]. This study explores how the integration of a video anomaly detection algorithm can optimize the identification of anomalous behavior in cyclist safety. Specifically, it investigates:

1. How effectively can video anomaly detection algorithms identify potentially unsafe cyclist interactions and behaviors in urban environments?

2. How can the results of video anomaly detection (VAD), with enhanced interpretability, assist professionals in making informed decisions for cyclist safety?

2 State of the Art

In this research, we investigate how computer vision techniques can be applied to evaluate cyclist safety. This section reviews existing studies on factors influencing cyclist behavior and examines the use of computer vision in this domain. Additionally, we delve into various methods for video anomaly detection. Given our intention to utilize a state-of-the-art frame prediction network for detecting anomalies, we also investigate relevant literature on video frame prediction.

2.1 Cyclist Safety

There are multiple studies about cyclist safety, which are briefly discussed in this section. The UDRIVE (eUropean naturalistic Driving and Riding for Infrastructure & Vehicle safety and Environment) project [9] represents a key point in research on naturalistic driving, providing in-depth insights into real-world driver behavior, safety, and environmental impacts. Using advanced sensors and cameras across 240 vehicles, UDRIVE captured millions of kilometers of driving data in diverse European conditions. This data has allowed researchers to analyze Critical Safety Events (CSEs), such as driver distractions and near-accidents, offering key insights for improving road safety[10].

Another significant research utilized smartphone GPS data from cyclists in Montreal to evaluate speed, travel time, and delays at intersections and road segments, with a particular focus on cyclist safety. By capturing GPS data, they attempted to identify CSEs to come up with promising results [11].

The BIKEALYZE project [12] integrates real-time GPS tracking and sensor technologies to monitor and enhance urban cycling environments. By harnessing data from cyclists' smartphones and public infrastructure, BIKEALYZE provides a comprehensive analysis of CSEs. Unlike traditional cycling safety studies, which often rely on historical data, this project uses real-time information to offer dynamic insights, ensuring that urban planners and policymakers can make informed, timely decisions [13].

While this research contributes to understanding cyclist safety by focusing on identifying safety-critical events (CSEs) through manual review of video data and then analyzing them with statistical methods, recent studies include deep learning techniques to address cyclist safety more effectively [14].

2.2 Video Anomaly Detection

Video anomaly detection using computer vision and deep learning is studied on various approaches, including:

- **One-Class Classification**: The model is trained exclusively on normal video data and learns to identify anomalies based on deviations from the normal patterns [15].
- **Unsupervised Learning:** Anomalies are present in the training data but are not explicitly annotated. [16].
- **Supervised or Weakly-Supervised:** The training set includes sufficient or limited anomalous data, and both normal and anomalous data are labeled [17].

VAD methods can be broadly classified into:

- **Object-centric Methods**: These methods focus on detecting anomalies in individual objects within the video. This process often depends on accurate object detection [18].
- **Frame-centric Methods** These methods consider the entire video frame without focusing solely on objects [19].

The choice between the two paradigms is task-specific, often constrained by computational power and the desired localization accuracy. Object-centric methods depend heavily on precise annotations, which may be impractical for large datasets, whereas frame-centric methods may lack the granularity required for identifying specific safety-critical behaviors. These challenges can arise a need of for hybrid approaches that combine object-specific insights with broader frame-level analysis.

Recent research in video anomaly detection has made considerable progress, establishing the state-of-the-art methodologies in the domain. Some of the example studies include:

Normalizing Flows for Human Pose Anomaly Detection (STG-NF) [20] represents a significant advancement in video anomaly detection. This approach focuses solely on human pose, reducing the influence of outer parameters like appearance and background. The model works directly on human pose graph sequences. By leveraging a highly compact pose representation within a normalizing flows framework, the method achieves state-of-the-art results on two major benchmarks: the unsupervised ShanghaiTech dataset and the supervised UBnormal dataset [21, 22]. The algorithm's versatility allows it to handle both unsupervised scenarios with only normal examples and supervised settings with labeled normal and abnormal data. This approach not only improves accuracy but also minimizes bias associated with appearance-based methods [20]

The Attribute-based Representations for Accurate and Interpretable Video Anomaly Detection (AccI-VAD) [23] method outperforms previous approaches with significant improvements. This model utilizes attribute-based representations to detect anomalies in videos, offering both accuracy and interpretability. The approach extracts high-level semantic attributes from video frames using pre-trained models and uses these attributes to characterize normal and anomalous events. By focusing on interpretable attributes, the method provides insights into the nature of detected anomalies, making it easier for human operators to understand and validate the results. The attribute-based approach demonstrates robustness across various scenarios and achieves state-of-the-art performance on several popular video anomaly detection benchmarks, including Ped2, Avenue, ShanghaiTech, and UCFCrime [24, 25, 21, 26]. This method shows versatility in handling both object-centric and frame-centric anomalies, making it a powerful and flexible tool for video surveillance and security applications [23].

3 Materials and Methods

3.1 Video Anomaly Detection Strategy

Our approach follows the methodology of AccI-VAD proposed by Reiss & Hoshen [23], which introduces a highly interpretable framework. It combines explicit, semantic attributes such as velocity and pose with implicit deep representations, leveraging the strengths of both paradigms to deliver state-of-the-art performance in video anomaly detection.

The key innovation of AccI-VAD lies in its dual representation:

- **Explicit attributes (Velocity and Pose)**: These are intuitive and human-interpretable, enabling users to understand the rationale behind anomaly detection results.
- **Residual features (Deep Representations)**: These capture more abstract patterns that explicit attributes might miss, ensuring robustness across complex scenarios.

While pose features are part of AccI-VAD, they were not utilized in our approach primarily because the cyclists in our dataset are not large enough in the video frames to allow meaningful pose representations to be captured.

This dual approach ensures both accuracy and transparency, making AccI-VAD particularly valuable for applications requiring user validation, such as cyclist safety monitoring. In urban environments, ensuring cyclist safety demands not only detecting unusual events but also understanding their nature for timely intervention. For instance, detecting anomalies like sudden changes in cyclist velocity, erratic movement patterns, or interactions with other vehicles requires a system that is both accurate and interpretable.

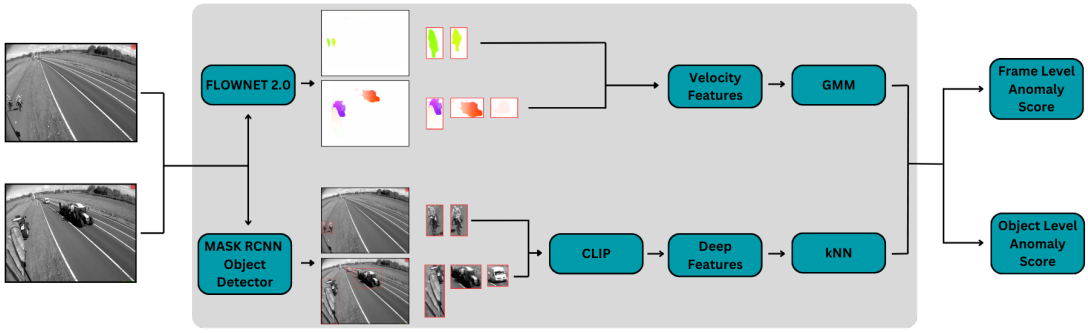

Fig. 1: An overview of the customized Attribute-based Representations for Accurate and Interpretable Video Anomaly Detection method. The process begins with extracting bounding boxes using an object detector and generating flow maps with FlowNet. The flow maps are then cropped to align with the detected bounding boxes, enabling the extraction of velocity features. Simultaneously, deep features are obtained using CLIP. Finally, density estimators are employed to compute frame-level and object-level anomaly scores, providing a comprehensive anomaly detection framework [27].

The explicit (velocity) and implicit (deep features) attributes of AccI-VAD are highly relevant to cycling safety:

- **Velocity**: Sudden acceleration, deceleration, or irregular speed changes can signal potential hazards, such as a cyclist avoiding an obstacle or an impending collision.
- **Deep Features**: Unusual visual patterns or object appearances can signal potential anomalies, such as the presence of unexpected objects on the cyclist's path, unfamiliar vehicle types, or atypical interactions between cyclists and their surroundings. These features capture high-level abstractions, enabling the detection of subtle or rare anomalies that may not be evident from explicit attributes.

3.1.1 Pre-processing

The preprocessing stage ensures that raw video data is systematically prepared for feature extraction through object detection and flow estimation. These steps lay the foundation for effective anomaly detection by accurately identifying and tracking objects in the video frames.

- **Object Detection:** Objects in each video frame are detected using ResNet50 Mask-RCNN [28], a deep learning-based object detection model pretrained on MS-COCO [29] dataset. This model identifies objects by generating bounding boxes and assigning class labels such as cyclists, pedestrians, and vehicles. The detected bounding boxes are filtered based on confidence thresholds specific to the dataset. By ensuring high detection accuracy, this step minimizes noise in subsequent feature extraction.
- **Optical Flow Estimation:** Flow patterns are captured by computing optical flow between consecutive frames using FlowNet2.0 [30]. These optical flow maps provide a pixel-wise representation of motion, which is cropped and aligned to each detected object's bounding box. The direction and magnitude of object movements within the bounding boxes are then extracted and standardized to a fixed resolution, ensuring consistency across frames and objects. This step is particularly critical for detecting anomalous or unexpected motion patterns.

3.1.2 Feature Extraction

The feature extraction stage focuses on representing objects within video frames using three complementary features, each leveraging a specific pre-trained model:

- **Velocity Features:** Velocity features are extracted from optical flow maps generated using FlowNet2.0 [30]. For each detected object, the optical flow within its bounding box is cropped, aligned, and rescaled to a fixed size. To summarize motion information, flow directions are quantized into discrete bins, and the average magnitude of flow vectors within each bin is calculated. This representation effectively captures speed and movement orientation, enabling the detection of anomalies.
- **Deep Features:** Deep features are extracted using the CLIP image encoder (ViT-B/16) [31], a pre-trained vision transformer. They represent high-dimensional, abstract characteristics of objects within video frames, capturing patterns that explicit attributes may not address. These features, extracted from pre-trained deep learning models, encode residual information such as texture, shape, and contextual relationships. For example, they can identify unusual object appearances, atypical environmental interactions, or rare occurrences that deviate from learned normal patterns. This abstraction allows deep features to generalize across complex scenarios and detect anomalies that may not have explicit, predefined descriptors.

3.1.3 Density Estimation

Density estimation is used to compute anomaly scores based on the extracted features. Separate models are trained for each feature type to handle their unique characteristics:

- **Velocity Features:** For velocity features, which are lower-dimensional, a Gaussian Mixture Model (GMM) is used. The GMM learns a probabilistic density function from the velocity features of normal training data. During inference, objects with low estimated densities are flagged as anomalous.
- **Deep Features:** Deep features are high-dimensional and do not conform to specific parametric assumptions[23]. To estimate their density, a k-Nearest Neighbors (kNN) algorithm is employed. This involves calculating the L2 distance between the deep feature of a target object and its k nearest neighbors in the training feature set. Larger distances indicate greater deviations from normal patterns, identifying potential anomalies.

The density estimation models are trained exclusively on normal training data, ensuring that anomalies can be identified as deviations from the learned normal patterns. By employing different approaches tailored to the characteristics of the features, the system effectively models feature distributions with flexibility.

3.1.4 Score Calibration

Score calibration ensures that anomaly scores from different feature types (velocity and deep features) are normalized and combined effectively for robust anomaly detection.

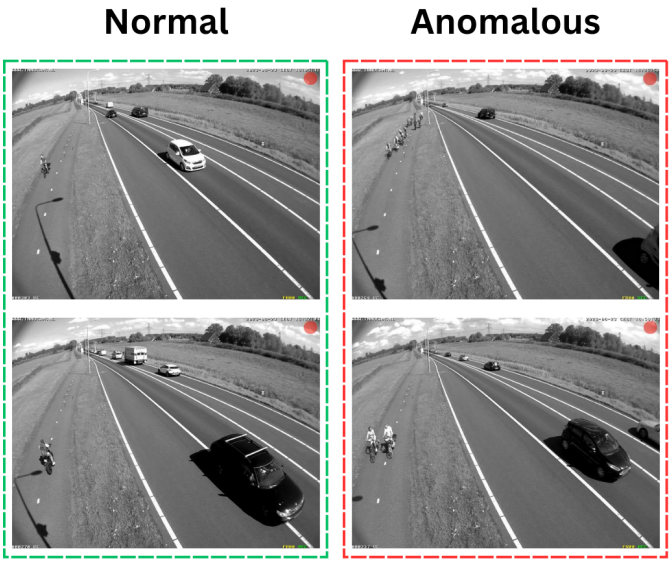

Fig. 2: Examples frames from Vinkeveen Dataset, left column shows normal frame examples, right shows anomalous.

- **Feature-Specific Score Normalization:** Each feature type produces an anomaly score measuring the deviation of the object from normal patterns learned during training. To ensure consistency across features, scores are normalized using min-max scaling:

$$s_{\text{norm}}(x) = \frac{s(x) - \nu}{\mu - \nu}, \tag{1}$$

where $s(x)$ is the raw anomaly score, ν is the minimum score, and μ is the maximum score observed during training. This normalization maps all scores into the range $[0, 1]$, ensuring they are comparable across feature types. By scaling the scores uniformly, the system prevents dominance by features with inherently larger ranges.

- **Clip-wise kNN Adjustment for Deep Features:** For high-dimensional deep features, emphasis is placed on avoiding biases caused by repetitive patterns within individual video clips during kNN-based scoring.
  - During training, the nearest neighbors for a given object are restricted to originate from different video clips. This prevents artificially low distances that may arise from consecutive frames showing the same object with minimal variation.
  - Sampling neighbors from different clips ensures that anomaly scores capture general patterns of normal behavior, rather than redundancies within the same clip. This adjustment is critical for effectively detecting anomalies in high-dimensional feature spaces.

3.1.5 Inference

- **Object-Level Anomaly Score:** The anomaly score for each object is computed as the **sum** of its normalized scores across all feature types:

$$t_{\text{object}}(o_k) = s_{\text{norm, velocity}}(o_k) + s_{\text{norm, deep}}(o_k), \tag{2}$$

where $s_{\text{norm, feature}}(o_k)$ is the normalized anomaly scores for the velocity or deep features of object o_k. This approach ensures that each object's behavior is analyzed comprehensively by integrating all feature types while considering their individual contributions.

- **Frame-Level Anomaly Score:** The final anomaly score for each frame is derived as the sum of the highest normalized scores across all objects in the frame for each feature type:

$$t_{\text{frame}}(f) = \max_k s_{\text{norm, velocity}}(o_k) + \max_k s_{\text{norm, deep}}(o_k) \tag{3}$$

where o_k represents the k-th object in the frame. This aggregation highlights the most anomalous objects for each feature type, ensuring the frame-level score captures significant deviations.

3.2 Datasets

For this study, we utilize video data sourced from innocam.nl, a platform focused on smart camera technology. The dataset includes footage captured in Vinkeveen, Netherlands. The videos are recorded in an urban environment, offering a diverse range of cyclist interactions with vehicles, pedestrians, and infrastructure. The dataset is particularly suited for video anomaly detection tasks, as it contains real-world scenarios that present potential safety risks to cyclists.

3.2.1 Vinkeveen Dataset

The dataset consists of the following details:

- **Number of Clips:** 327 video clips comprising 66,819 frames, each with a resolution of 800×600.
- **Annotations:**
  - Bounding boxes for detected objects, including cyclists and pedestrians.
  - Frame-level labels indicating anomalous events.
  - Detailed annotations capturing various risky behaviors, including:
    - Lateral position
    - Spacing
    - Head orientation
    - Phone usage

- Peloton (group cycling dynamics)
- Swerving
- Hands off the steering wheel
- Speed

Example frames from the Vinkeveen Dataset are shown in Figure 2.

	Clips			Frames			Anomaly Types
	Train	Test	Total	Train	Test	Total	
Normal	153	71	224	46,207	17,823	64,030	-
Anomalous	-	103	103	-	2,789	2,789	8

Table 1: Statistics of the Innocam Vinkeveen Dataset, showing clip and frame distributions for normal and anomalous data, along with anomaly types.

3.3 Metrics

The performance of the method is evaluated with the AUC-ROC (Area Under the Receiver Operating Characteristic Curve) metric, which is commonly used for video anomaly detection evaluation. We compute both Micro AUC-ROC and Macro AUC-ROC metrics at the frame and object levels. These metrics assess the model's ability to distinguish between normal and anomalous frames and objects based on anomaly scores.

3.3.1 Frame-Level AUC-ROC

- **Micro AUC-ROC:** At the frame level, Micro AUC-ROC evaluates anomaly scores across all frames collectively. The process involves:

  1. Computing a single anomaly score for each frame f by selecting the maximum score among all objects within the frame:
  $$t(f) = \max_k s(o_k),$$
  where o_k denotes the k-th object in the frame, and $s(o_k)$ represents its corresponding anomaly score.
  2. Aggregating all frame-level anomaly scores and ground truth labels from the entire dataset to construct a single Receiver Operating Characteristic (ROC) curve.
  3. Calculating the area under this ROC curve to determine the Micro AUC-ROC.

- **Macro AUC-ROC:** For Macro AUC-ROC, the evaluation is performed at the clip level. This involves:

  1. Calculating the frame-level AUC-ROC for each clip c using the anomaly scores and ground truth labels specific to that clip.
  2. Computing the Macro AUC-ROC as the average of the frame-level AUC-ROC values across all clips:
  $$\text{Macro AUC-ROC} = \frac{1}{N}\sum_{c=1}^{N} \text{AUC-ROC}_{c_{\text{frame-level}}}$$
  where N represents the total number of clips, and $\text{AUC-ROC}_{c_{\text{frame-level}}}$ is the frame-level AUC-ROC for the c-th clip.

3.3.2 Object-Level AUC-ROC

- **Micro AUC-ROC:** At the object level, Micro AUC-ROC assesses anomaly scores for all detected objects collectively:

  1. An anomaly score $s(o_k)$ is calculated for each object o_k in every frame, using features such as velocity and deep embeddings.
  2. All object-level anomaly scores and their corresponding ground truth labels are aggregated across the entire dataset to generate a single ROC curve.
  3. The area under this ROC curve is used to determine the Micro AUC-ROC for object-level evaluation.

- **Macro AUC-ROC:** For Macro AUC-ROC, the evaluation is conducted on a per-clip basis:

  1. For each clip c, the object-level AUC-ROC is computed based on the anomaly scores and ground truth labels of all objects within that clip.
  2. The Macro AUC-ROC is then calculated as the average of the object-level AUC-ROC values across all clips:
  $$\text{Macro AUC-ROC} = \frac{1}{N}\sum_{c=1}^{N} \text{AUC-ROC}_{c_{\text{object-level}}}$$
  where N is the total number of clips, and $\text{AUC-ROC}_{c_{\text{object-level}}}$ represents the object-level AUC-ROC for the c-th clip.

4 Experiments and Results

4.1 Experimental Setup

For our experiments, we use ResNet50 Mask-RCNN, pretrained on the MS-COCO dataset, to extract object bounding boxes. Low-confidence objects are filtered out by setting a confidence threshold of 0.9 for both the training and testing phases on the Vinkeveen dataset. Optical flow maps are generated using FlowNet2.0, providing motion-based features. Deep features are extracted using a pretrained ViT B-16 CLIP image encoder.

To process flow maps, we rescale them to $H_{\text{velocity}} \times W_{\text{velocity}} = 224 \times 224$. Bounding box dimensions are used to normalize the velocity magnitudes by the height of the bounding box for scale invariance . We evaluate the effect of this normalization on anomaly detection performance as part of our experiments. Velocity features are represented using $B = 8$ orientations.

We experiment with different parameter settings for the Gaussian Mixture Model (GMM) used for velocity features and k-Nearest Neighbors (kNN) used for deep representations. The final anomaly score for a frame is determined as the maximum anomaly score among all objects in that frame. Gaussian smoothing is applied to refine anomaly scores during post-processing.

All experiments were run on an NVIDIA A40-12C GPU where applicable, to accelerate feature extraction, density estimation, and anomaly scoring.

This setup ensures comprehensive evaluation of feature configurations, parameter choices, and the impact of normalization, enabling robust and accurate anomaly detection.

4.2 Experiments

4.2.1 Normalization of Velocity Features

To ensure scale invariance across objects of varying sizes, velocity magnitudes were normalized using the height of the corresponding bounding boxes. Specifically, the normalization was performed by dividing the L1 norm of the optical flow vectors within each bounding box by the bounding box height:

$$mag_{\text{norm}} = \frac{mag}{h_{\text{bbox}}}.$$

We compared normalized and unnormalized velocity features to analyze the effect of normalization on anomaly detection performance. Normalization is applied consistently across the training and testing phases. Flow maps are rescaled to 224×224 before extracting normalized velocity features.

Analysis Type	Velocity Normalization	Velocity Features	Deep Features	GMM	kNN	Sigma	Micro	Macro
Frame-Level		✓		5	-	7	**94.6**	**93.1**
Object-Level			✓	-	1	7	**55.5**	**55.0**

Table 2: Summary of Best-Performing Configurations for Frame-Level and Object-Level AUC-ROC (%) Scores on the Vinkeveen Dataset.

4.2.2 Feature Selection and Combination: Velocity, Deep, and Combined

This experiment assesses the effectiveness of velocity features, deep features, and their combination for anomaly detection:

- **Velocity Features:** Derived from optical flow maps and modeled using a Gaussian Mixture Model (GMM), these features focus on motion-specific patterns and dynamics within the video.
- **Deep Features:** Extracted using a pretrained ViT B-16 CLIP image encoder, these features are modeled with a k-Nearest Neighbors (kNN) density estimator. They capture high-level semantic and visual representations.
- **Combined Features:** Anomaly scores generated from both the velocity and deep feature models are combined through summation. This method harnesses the complementary strengths of motion-based and visual features, leading to improved anomaly detection performance.

The experiment evaluates the individual contributions of velocity and deep features to anomaly detection and investigates whether their combination further enhances detection accuracy.

4.2.3 Model Parameter Optimization

This experiment focuses on optimizing the key parameters of the Gaussian Mixture Model (GMM) for velocity features and the k-Nearest Neighbors (kNN) estimator for deep features. Specifically, the number of components in the GMM and the number of neighbors in the kNN model are systematically tested to identify the configurations that best capture the underlying patterns in the data.

- **Velocity Features:** The Gaussian Mixture Model (GMM) is evaluated with different numbers of components:

  $$n_{\text{components}} = 1, 2, 3, 4, 5.$$

  This tests the impact of model complexity on capturing motion patterns. A smaller number of components results in coarse modeling of velocity feature distribution, potentially missing subtle motion variations, while a larger number captures finer granularity but risks overfitting with limited training data [32].

- **Deep Features:** The k-Nearest Neighbors (kNN) model is tested with varying numbers of neighbors:

  $$k_{\text{neighbors}} = 1, 3, 5, 7.$$

  This assesses the effect of neighborhood size on anomaly scoring. A smaller k value increases sensitivity to anomalies but may overreact to minor variations, causing higher false positives, while a larger k value smoothens the decision boundary, enhancing robustness but potentially missing subtle anomalies.

The results from these experiments help identify the optimal parameters for both models, ensuring robust and accurate anomaly detection.

4.2.4 Effect of Gaussian Smoothing

We investigate the role of Gaussian smoothing in refining anomaly scores. By reducing noise in frame-level and object-level scores, smoothing aims to improve detection robustness and minimize false positives. Gaussian smoothing is applied to the scores of velocity features, deep features, and their combination during post-processing. The smoothing parameter (σ) is varied to analyze its impact on anomaly score refinement.

$$\text{Smoothing parameter } (\sigma) = 1, 3, 5, 7.$$

4.3 Results

4.3.1 Overall Performance

The proposed anomaly detection framework demonstrated strong performance at the frame leve* on the Vinkeveen dataset, as summarized in Table 2. The best frame-level configuration, which utilized raw velocity features with 5 GMM components and $\sigma = 7$, achieved a Micro AUC-ROC of 94.6% and a Macro AUC-ROC of 93.1%, indicating a high ability to distinguish anomalies from normal patterns in frame-wise evaluation.

However, for object-level analysis, the performance was more limited. The best configuration, which employed deep features with 1 nearest neighbor (kNN) and $\sigma = 7$, achieved a Micro AUC-ROC of 55.5% and a Macro AUC-ROC of 55.0%. These results suggest that the framework struggled to effectively identify anomalies at the object level, highlighting the need for further improvements.

4.3.2 Effect of Velocity Features Normalization

- **Frame-Level Results:** The experiment using unnormalized velocity features achieved the best performance, with a Micro AUC-ROC of 94.6% and a Macro AUC-ROC of 93.1%, outperforming the normalized configuration, which achieved only 83.9% and 92.2%, respectively. While parameter optimization for the normalized data improved the results (89.3% Micro and 92.1% Macro), it still fell short of the unnormalized configuration (see Table 3).

Impact of Normalization on the Best-Performing Configuration		
Velocity Normalization	**Micro**	**Macro**
-	**94.6**	**93.1**
✓	83.9	92.2
Performance of the Best Experiment with Normalization Applied		
Velocity Normalization	**Micro**	**Macro**
✓	89.3	92.1

Table 3: Frame-Level AUC-ROC (%) Scores: The upper part compares experiments with and without velocity normalization, using velocity features only, $\sigma = 7$, and $GMM = 5$. The lower part highlights the performance of the best experiment with normalization applied (velocity features only, $\sigma = 7$, and $GMM = 3$). Best scores are shown in bold.

- **Object-Level Results:** Normalization had minimal impact on object-level results, with both raw and normalized features achieving the same Micro AUC-ROC of 55.5%, indicating similar performance across configurations.

Impact of Normalization on the Best-Performing Configuration		
Velocity Normalization	**Micro**	**Macro**
-	55.5	55.0
✓	55.5	55.0
Performance of the Best Experiment with Normalization Applied		
Velocity Normalization	**Micro**	**Macro**
✓	55.5	55.0

Table 4: Object-Level AUC-ROC (%) Scores: The upper part compares experiments with and without velocity normalization, using deep features only, $\sigma = 7$, and $k = 1$. The lower part highlights the performance of the best experiment with normalization applied (deep features only, $\sigma = 7$, and $k = 1$).

4.3.3 Effect of Feature Selection and Combination

- **Frame-Level Results:** Velocity features proved to be the most effective at the frame level, achieving a Micro AUC-ROC of 94.6%. This highlights their critical importance in detecting motion-related anomalies, particularly for cyclists, where motion patterns are a key indicator of safety-critical events. While combining velocity and deep features showed slightly lower performance (Micro AUC-ROC of 94.0%), it demonstrated complementary strengths, which may provide additional robustness in certain case.

Velocity Features	Deep Features	Micro	Macro
✓	-	**94.6**	**93.1**
-	✓	94.0	92.3
✓	✓	76.0	90.3

Table 5: Frame-Level AUC-ROC (%) Scores: The table compares the effects of different feature combinations on frame-level anomaly detection. (1) Using velocity features only (GMM = 5, $\sigma = 7$); (2) using deep features only (GMM = 5, $k = 1$, $\sigma = 7$); (3) using both features ($k = 1$, $\sigma = 7$). Best scores are shown in bold.

- **Object-Level Results:** At the object level, there was minimal improvement, with deep features achieving a Micro AUC-ROC of 55.5% and velocity features scoring 55.2%. The limited change suggests challenges in learning distinctive patterns for object-level anomalies.

4.3.4 Parameter Optimization

Parameter tuning experiments indicated:

- **GMM for Velocity Features:** The optimal configuration was achieved with 5 components, yielding a Micro AUC-ROC of 94.6% and a Macro AUC-ROC of 93.1% using only velocity features for frame-level comparisons. While the effect can also be investigated at using combination of velocity and deep features, only velocity features were selected for reporting to isolate their contribution to anomaly detection performance. A comparison of the effect of the number of components on performance is provided in Table 6. Notably, the performance consistently improved as the number of components increased, reaching a peak at 5 components for frame-level anomaly detection.

$n_{\text{components}}$	Micro	Macro
1	88.7	92.1
2	91.9	92.5
3	93.9	92.8
4	94.4	**93.4**
5	**94.6**	93.1

Table 6: Effect of the Number of GMM Components on Frame-Level AUC-ROC (%). (Without velocity normalization, Velocity feature only and $\sigma = 7$)

- **kNN for Deep Features:** Parameter tuning for the number of neighbors in kNN revealed minimal variations in performance at the frame level. The best configuration, with a neighborhood size of 7, achieved a Micro AUC-ROC of 76.0% and a Macro AUC-ROC of 90.3%. As shown in Table 7, increasing the neighborhood size from 1 to 7 resulted in a slight improvement in Micro AUC-ROC, while the Macro AUC-ROC remained relatively stable around 90.3%-90.5%.

$k_{\text{neighbors}}$	Micro	Macro
1	75.2	**90.5**
3	75.6	90.4
5	75.8	90.3
7	**76.0**	90.3

Table 7: Effect of the Number of Neighbors (k) in kNN on Frame-Level AUC-ROC (%). (Deep features only and $\sigma = 7$) Best scores are shown in bold.

Reporting is focused on frame-level analysis, as the model exhibits learning challenges at the object level, resulting in parameter optimization producing negligible variations in performance across configurations.

4.3.5 Effect of Gaussian Smoothing

σ	Micro	Macro
1	91.3	93.2
3	93.0	**93.6**
5	93.9	93.4
7	**94.6**	93.1

Table 8: Effect of Sigma (σ) on Frame-Level AUC-ROC (%). (Without velocity normalization, Velocity feature only and GMM = 5)

σ	Micro	Macro
1	48.5	50.2
3	50.8	52.7
5	52.3	53.4
7	**55.5**	**55.0**

Table 9: Effect of Sigma (σ) on Object-Level AUC-ROC (%).(Without velocity normalization, Deep feature only and $k = 1$)

The experiments with Gaussian smoothing demonstrated its significant impact on both frame-level and object-level anomaly detection performance. A smoothing parameter (σ) of 7 provided the highest performance across both analyses. For frame-level evaluation, this configuration achieved a Micro AUC-ROC of 94.6% and a Macro AUC-ROC of 93.1%(Table 8), while for object-level evaluation, it achieved a Micro AUC-ROC of 55.5% and a Macro AUC-ROC of 55.0% (Table 9). Gaussian smoothing proved effective in filtering out short-term noise in anomaly scores, reducing false positives by smoothing isolated anomalous predictions, and improving the stability and robustness of results.

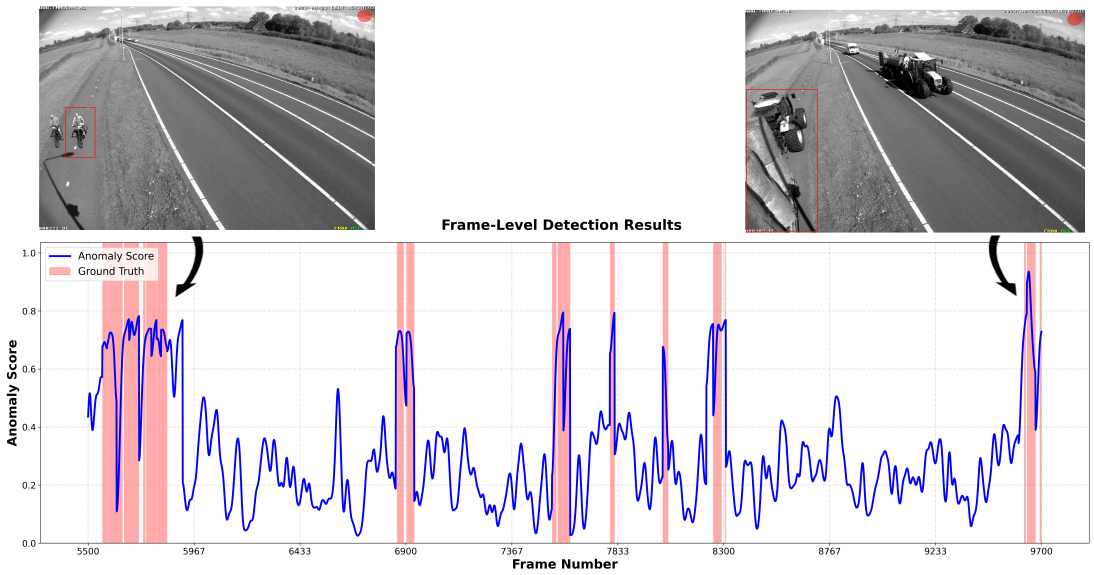

Fig. 3: Frame-level anomaly detection results on Vinkeveen dataset. The top left image highlights lateral positioning anomaly behavior, while the top right shows an a tractor entering the bike lane which is a also another anomalous scene. In the graph below, the blue curve represents the anomaly score, and the red shaded regions indicate ground truth (GT) annotations for anomalies. The framework's anomaly score aligns with the GT during anomalous events, capturing the transitions effectively. Minor fluctuations in normal frames reflect the framework's sensitivity, while overall robustness ensures accurate anomaly detection.

5 Discussion, Conclusion & Future Work

5.1 Discussion

In this work, we proposed a video anomaly detection framework combining explicit features (velocity) with deep residual features to make interpretable anomaly detection, which will help traffic safety professionals enhance cyclist safety. The framework is designed to automate the identification of safety-critical events in video footage, reducing the need for time-consuming manual review. By isolating the most relevant and anomalous segments, the system enables stakeholders to focus their efforts on high-risk areas and behaviors, supporting targeted interventions and evidence-based urban planning. While the framework does not directly notify individual cyclists or ensure immediate municipal action, it provides the tools to streamline the decision-making process for traffic safety improvements. The results highlight a disparity between frame-level and object-level performance. At the frame level, the framework achieved a robust AUC-ROC of 94.6%, demonstrating its effectiveness in analyzing the overall scene context. In contrast, object-level performance was significantly lower, with an AUC-ROC of 55.5%, reflecting the challenges in accurately detecting object-level anomalies.

The lower object-level performance can be attributed to limitations in the off-the-shelf object detector used during feature extraction. Specifically, the detector often generated empty bounding boxes, even at high confidence thresholds, and struggled with inconsistencies caused by overlapping objects and frequent occlusions in urban environments. These inaccuracies introduced noise into the data, particularly from unnecessary feature extraction of non-existent objects, which impaired the model's ability to identify true anomalies at the object level.

On the other hand, the frame-level anomaly detection demonstrated strong performance, achieving significantly higher Micro and Macro AUC-ROC scores compared to object-level detection. This highlights the effectiveness of analyzing the overall context of the scene, where the combined analysis of all objects provide a more comprehensive representation of potential irregularities. The strong frame-level results indicate the framework's robustness in identifying anomalies at a broader scale, even in complex urban environments.

While the technical contributions of this framework are significant, addressing key ethical and societal considerations is critical for its responsible implementation. Continuous video monitoring of cyclists and other road users raises privacy concerns. To address this, the framework should incorporate anonymization techniques, such as blurring faces and license plates, to protect individual identities. Furthermore, transparent communication about how data is collected, stored, and used is essential for building public trust and ensuring the system is used exclusively for safety purposes.

The societal concern regarding increased surveillance is another critical issue. By utilizing existing infrastructure rather than advocating for additional cameras, the framework minimizes the risk of promoting excessive surveillance. Transparent deployment policies, coupled with public engagement and consultation, can foster trust and ensure the responsible use of the system. Additionally, strict ethical guidelines must be established to limit data usage exclusively to safety-related applications and explicitly prohibit commercial exploitation or profiling of individuals.

The societal concern regarding increased surveillance is another critical aspect. By leveraging existing infrastructure rather than advocating for additional cameras, the framework can avoid promoting excessive surveillance. Transparent deployment policies, coupled with public engagement and consultation can be established to to build trust and ensure responsible use of the system. Additionally, strict ethical guidelines must be established to restrict data usage exclusively for safety-related applications and prohibit commercial exploitation or profiling of individuals.

Data security is equally important. Video footage must be securely stored and transmitted using encryption and strict access controls, ensuring adherence to data protection regulations such as GDPR. These measures are vital to prevent unauthorized access and misuse of sensitive information.

Sustainability also plays a key role in the responsible deployment of AI systems. Their implementation must balance computational demands with broader environmental and societal goals. Efforts to optimize resource consumption by using more energy-efficent computational models can be a key in aligning the framework with sustainability objectives.

Finally, to maximize its practical impact, the framework's findings must be accessible and interpretable for traffic safety professionals

and urban planners. Results should be presented clearly and in an actionable manner, enabling stakeholders to prioritize interventions and make informed decisions for improving cycling infrastructure and urban traffic safety. By streamlining the analysis of video data and providing actionable insights, this framework can serve as a valuable tool for enhancing cyclist safety and fostering safer urban environments.

5.2 Conclusion

The proposed video anomaly detection framework demonstrated strong performance in identifying anomalous behaviors, particularly at the frame level, underscoring its potential for practical applications in cyclist safety. Quantitative results showed that the framework effectively captured anomalies, highlighting its reliability and robustness in real-world urban scenarios. The integration of explicit velocity features and deep residual representations proved to be complementary, offering both interpretability and enhanced detection performance.

For qualitative analysis, interpretability can be further developed by improving object-level scores. Enhanced object-level accuracy would allow for a more precise identification of the reasons behind detected anomalies, offering deeper insights into the nature of events. This level of interpretability is essential for understanding and addressing specific safety concerns in urban environments.

Despite these strengths, certain limitations remain, particularly in object-level detection, where issues such as noise introduced by empty bounding boxes and dataset imbalances impact performance. These findings suggest potential areas of improvement to enhance anomaly detection accuracy.

5.3 Future Work

While the current framework achieves promising results, several areas for further exploration and improvement are identified:

- **Improved Object Detection:** Integrating a state-of-the-art object detector or optimizing the current system could address challenges such as the generation of empty bounding boxes and noise, which negatively impact object-level evaluations. Improving the accuracy and reliability of object detection would lead to better localization and identification of anomalies at the object level, reducing false positives and enhancing detection robustness.

- **Pose Feature Utilization:** Incorporating pose features for anomaly detection has the potential to capture nuanced behaviors, particularly those related to human motion and activity. However, this approach requires higher-quality data where cyclists are captured at closer distances with sufficient resolution to extract meaningful pose information.

- **Improved Object-Level Scores:** Efforts focused on refining object-level anomaly detection scores could bridge the current performance gap between frame-level and object-level evaluations. Enhancing the model's ability to isolate object-specific anomalies would significantly improve interpretability and detection accuracy.

- **More Data for Training:** Expanding the training dataset by including more non-empty normal frames, sampling from diverse locations, and addressing dataset biases could strengthen the model's ability to learn robust normal patterns. This would help mitigate the impact of false positives, improve generalizability, and ensure better detection accuracy across varied environments..

Acknowledgements

- This project is financially supported by Regieorgaan SIA (part of NWO) under the RAAK Publiek project BikeSafeAI.



References

[1] Lama Ayad, Hocine Imine, Claudio Lantieri, and Francesca De Crescenzio. Pedal towards safety: The development and evaluation of a risk index for cyclists. *Infrastructures*, 9(1), 2024.

[2] *Cyclist safety: an information resource for decision-makers and practitioners*. World Health Organization, Genève, Switzerland, October 2020.

[3] Kay Teschke, M. Anne Harris, Conor C.O. Reynolds, Meghan Winters, Shelina Babul, Mary Chipman, Michael D. Cusimano, Jeff R. Brubacher, Garth Hunte, Steven M. Friedman, Melody Monro, Hui Shen, Lee Vernich, and Peter A. Cripton. Route infrastructure and the risk of injuries to bicyclists: A case-crossover study. *American Journal of Public Health*, 102(12):2336–2343, 2012. PMID: 23078480.

[4] Fred Wegman and Paul Schepers. Safe system approach for cyclists in the netherlands: Towards zero fatalities and serious injuries? *Accident Analysis and Prevention*, 195, 2023.

[5] SWOV. Cyclists. SWOV-Fact sheet, The Hague, February 2023. Fact sheet number: 16415.

[6] Tarek Sayed, Mohamed Zaki, and Jarvis Autey. Automated safety diagnosis of vehicle–bicycle interactions using computer vision analysis. *Safety Science*, 59:163–172, 11 2013.

[7] Kevin Gildea, Daniel Hall, Clara Mercadal-Baudart, Brian Caulfield, and Ciaran Simms. Computer vision-based assessment of cyclist-tram track interactions for predictive modeling of crossing success. *Journal of Safety Research*, 87:202–216, 2023.

[8] A. R. G. Littek. *Explainable Anomaly Detection in Surveillance Videos: Autoencoder-based Reconstruction and Error Map Visualization*. Dissertation, Uppsala University, 2024. Retrieved from the Uppsala University Publications.

[9] UDRIVE Project Team. European naturalistic driving and riding for infrastructure vehicle safety and environment (udrive), 2017. FP7 Project, Grant Agreement ID: 314050.

[10] R. Jansen, T. Lotan, M. Winkelbauer, J. Bärgman, J. Kovaceva, M. Donabauer, A. Pommer, O. Musicant, A. Harel, S. Wesseling, M. Christoph, N. van Nes, and M. de Goede. *UDRIVE deliverable 44.1 Interactions with vulnerable road users, of the EU FP7 Project UDRIVE*. UDRIVE Consortium, 1 edition, July 2017.

[11] Jillian Strauss and Luis F. Miranda-Moreno. Speed, travel time and delay for intersections and road segments in the montreal network using cyclist smartphone gps data. *Transportation Research Part D: Transport and Environment*, 57:155–171, 2017.

[12] Bikealyze Project Team. Bikealyze: Naturalistic cycling study for infrastructure and safety analysis, 2016. Project Lead: Salzburg Research, Project Duration: 04.2015 – 10.2016.

[13] Christine Chaloupka-Risser and Elisabeth Fussl. The importance of communication between cyclists and other traffic participants and its potential in reducing traffic safety-critical events. *Transactions on Transport Sciences*, 8(1):24–30, 2017.

[14] Rashmiranjan Nayak, Umesh C. Pati, and Santos Das. A comprehensive review on deep learning-based methods for video anomaly detection. *Image and Vision Computing*, 106:104078, 11 2020.

[15] Naeem Seliya, Azadeh Abdollah Zadeh, and Taghi M. Khoshgoftaar. A literature review on one-class classification and its potential applications in big data. *Journal of Big Data*, 8, 2021.

[16] André Paulo Ferreira Machado, Ricardo Emanuel Vaz Vargas, Patrick Marques Ciarelli, and Celso Jose Munaro. Improving performance of one-class classifiers applied to anomaly detection in oil wells. *Journal of Petroleum Science and Engineering*, 218:110983, 2022.

[17] Yongwei Nie, Hao Huang, Chengjiang Long, Qing Zhang, Pradipta Maji, and Hongmin Cai. Interleaving one-class and weakly-supervised models with adaptive thresholding for unsupervised video anomaly detection, 2024.

[18] Pankaj Raj Roy, Guillaume-Alexandre Bilodeau, and Lama Seoud. Local anomaly detection in videos using object-centric adversarial learning, 2020.

[19] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection - a new baseline. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6536–6545, 2018.

[20] Or Hirschorn and Shai Avidan. Normalizing flows for human pose anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13545–13554, October 2023.

[21] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 589–597, 2016.

[22] Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Ubnormal: New benchmark for supervised open-set video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20143–20153, June 2022.

[23] Tal Reiss and Yedid Hoshen. Attribute-based representations for accurate and interpretable video anomaly detection. *arXiv preprint arXiv:2212.00789*, 2022.

[24] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1975–1981, 2010.

[25] 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 (ICCV)*, pages 2720–2727. IEEE, 2013.

[26] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos, 2019.

[27] Tal Reiss and Yedid Hoshen. Attribute-based representations for accurate and interpretable video anomaly detection. *arXiv preprint arXiv:2212.00789*, 2022.

[28] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017.

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.

[30] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021.

[32] Markus Wicker, Cihan Ates, Max Okraschevski, Simon Holz, Rainer Koch, and Hans-Jörg Bauer. Modeling multivariate spray characteristics with gaussian mixture models. *Energies*, 16(19), 2023.