# Pedestrian and Automatic Doors Abnormal Interactions Detection using Multi-Task Self-Supervised Learning

Olivier Laurendin[1]
olivier.laurendin@gmail.com
Sébastien Ambellouis[1]
sebastien.ambellouis@univ-eiffel.fr
Ankur Mahtani[2]
ankur.mahtani@railenium.eu
Anthony Fleury[3]
anthony.fleury@imt-lille-douai.fr

[1] Université Gustave Eiffel
20 Rue Élisée Reclus,
Villeneuve-d'Ascq, France

[2] Railenium Technological Research Institute
180 Rue Joseph-Louis Lagrange,
Valenciennes, France

[3] IMT Nord Europe
Rue Guglielmo Marconi,
Villeneuve-d'Ascq, France

**Abstract**

An anomalous event is commonly defined as an event sensibly distinct from the majority of its counterparts in a given context. Hence, video anomaly detection is often tackled as an out-of-distribution problem. Recent self-supervised state-of-the-art anomaly detection approaches are trained on object-centric descriptors extracted using object detectors. While these approaches are efficient for simultaneous detection and localization of single instance-related anomalies, they are not suited to identify anomalies related to different instance categories. In addition, anomalies that emerge from multiple instances interaction remain an open issue. In particular, we investigate the detection of anomalous interactions between pedestrians and automatic doors in the context of train video-surveillance.

We propose a three parts approach. A panoptic segmentation network extracts instance-aware semantic maps of pedestrians and automatic doors in the input video sequence. Two self-supervised multi-tasks networks are trained separately on each semantic maps sequence using a set of proxy tasks specifically tailored for the considered object categories. Finally, both networks anomalous binary responses are fused to provide a final interaction anomaly detection classification. We evaluate our method on a railway application dataset to detect doors-pedestrian anomalous interactions.

## 1 Introduction

Directly classifying a set of anomalous events following a fully-supervised scheme is not practical as anomalous events tend to be too rare and diverse to gather in sufficient quantities to train on. Such a training scheme makes the model incapable of identifying any unforeseen anomalous event. Anomalous events are instead commonly considered as outliers from a nominal distribution i.e. from a normality model trained on normal event instances. The nominal distribution is often learned by a self-supervised neural network trained on surrogate tasks designed to exploit certain consistent patterns within the training data. Proxy tasks can for instance learn temporal consistencies of motion information in consecutive frames (homogeneous motion), or visual consistencies such as the similar appearance of objects in the scene. A normality model is fit on normal training data to capture such assumptions, and detects any event that diverges from the normality model as anomalous.

Instead of fitting their normality model on the whole image, some recent state-of-the-art works [2, 3, 7, 10, 14] fit their normality model on object-centered bounding boxes extracted using object detectors [30, 31]. This approach is motivated by the prevalence of pedestrians in video surveillance anomalous events datasets documented in the literature, such as Avenue, ShanghaiTech, and UCSD Ped2 [20, 23]. However, it's important to note that these datasets predominantly contain isolated anomalies. These anomalies are primarily related to the movement of a single object class, such as individuals falling, or they may involve the presence of unexpected instance categories.

The recent release of the dataset FRailTRI20_DOD [16] provides a new unchallenged use-case: the detection of pedestrian and automatic doors interaction in a railway environment. To the best of our knowledge, it is the only dataset of the literature showcasing this kind of interaction anomalies.

Our contribution is multi-fold. We propose a new architecture that fuses the anomaly classifications provided by two punctual anomalies detection networks (ADN) to identify interaction anomalies. Both ADN are based on the self-supervised approach proposed in [10] and are respectively dedicated to identify doors-centric and pedestrian-centric punctual anomalies. We make use of the proxy tasks proposed in [10] to reduce the ADNs misalignment with respect to the anomaly detection task and we also introduce a new "optical flow prediction" task for both networks. In addition, we introduce the "doors states prediction" task for the doors anomaly detection network.

Finally, given that FRailTRI20_DOD is composed of top-down fish-eye images, we investigate the substitution of the object detector used in the original implementation of [10] by a version adapted to fish-eye images. Following [34], we leverage a panoptic segmentation [15] network of the literature [37] to extract doors and pedestrian masks. We then use simple heuristic to construct a bounding box from each pedestrian mask. We explore the use of different fish-eye specific bounding-box conventions to train our pedestrian anomaly detection network on.

## 2 Related Works

Several approaches leverage frames appearances consistency to model normality. Some methods use a CNN encoder as a feature extractor and train a one-class [35] or a binary [13, 21] classifier or use a reconstruction error as abnormality score [4, 11, 18, 24, 27, 28, 36]. Frames temporal consistency can also be used to model normality, either by predicting future frames from current data, using LSTMs [5] or generative approaches such as GANs [1, 6, 25, 26, 29, 33, 38]. Other approaches also use optical flow (OF) [6, 25, 29], or pixel-level motion between consecutive frames, to learn motion normality. While most approaches consider the entire image, "Object-centric" techniques focus their training on object-centric sequences produced by an object detector [2, 3, 7, 10, 14]. In particular, [3, 10] train a 3D

CNN following a multi-task learning strategy to yield a better normality model.

Anomaly detection works in the field of train transportation include [19] which proposes an image masking strategy for unsupervised anomaly localization and supervised anomaly classification in an intercity train station environment. How use-case is sensibly different from their since they do not identify doors-related anomalies. They only use a specifically tailored door tracking-based key frame extraction method to capture the status of train doors. In addition their definition of anomaly is restricted to locating abnormal objects on the train station. They do not identify doors-centric anomalies nor interaction anomalies.

In our work, we propose a two-stream architecture based on [10], each stream focusing on both doors and pedestrians behaviors. We use a panoptic segmentation network inspired by [37] to extract doors and pedestrians masks. Moreover, we design the "doors states prediction" task as a self-supervised proxy task to classify sliding doors states which can easily be generalized to other door types.

# 3 Method

## 3.1 Motivation

We aim at designing a network to detect interaction anomalies between instances of two distinct categories. We apply it to the detection of pedestrians-automatic doors interaction but this strategy can be applied to other sets of categories as long as we design an ADN for each category of interest. Hence, we implement one object-centric punctual ADN for each category so as to focus on each category of interest separately. We then use the late fusion of their binary responses to identify interaction anomalies. We use the multi-task learning scheme from [10] as we expect each surrogate task to capture a valuable feature of normality.

## 3.2 Data Preprocessing

Each ADN is trained on a batch of object-centric temporal sequences (OCTS) produced by concatenating several instance-aware segmentations provided by an object detector. The YOLOv3 network used in the original implementation is replaced by a K-Net panoptic segmentation network [37] following the work in [34]. For training and inference, our architecture input is a set $S(t) = \{S_k, k \in [-3,3]\}$ of 7 object-centric frames $S_k$ extracted from 31 consecutive frames $(F_i, i \in [t-15, t+15])$ centered on the current frame $F_t$. Object-centric frames $S_k$ are extracted from the frames $F_i$ selected using a fixed time step $\delta$ such that $S_k = F_{t+k\delta}, \forall k \in [-3,3]$. The section 4.3.3 presents how each ADN proxy task performances vary w.r.t $\delta$. Some K-Net predictions are shown in fig. 3(a) and the full data preparation process is summarized in fig. 2.

**Doors OCTS** are composed of doors segmentation masks concatenated across multiple frames. We focus on modeling the closing state of the door without the need to represent their appearance. **Pedestrian OCTS** are generated by extracting in several frames the closest bounding box from each pedestrian mask in the current frame. The selection of this closest bounding box is determined following three heuristic approaches: **Axis-aligned bounding boxes**, i.e. the vertical or horizontal bounding box of a pedestrian mask. **Radius-aligned bounding box** [32] oriented w.r.t the axis from the center of the image to the pedestrian mask centroid. **Human-aligned bounding box** [8] oriented w.r.t the first characteristic vector of the pedestrian mask pixels coordinates Principal Component Analysis (PCA). Some examples and their properties are shown in fig. 1(a) and fig. 2. in the supplementary material.

Pedestrian and doors OF OCTS are cropped from the OF provided by a pretrained FlowNet2 [12] on two consecutive raw images. Doors OF are extracted by applying doors segmentation masks on the OF images while each pedestrian OF is extracted through its bounding box. For both doors and pedestrians, the OF norm is provided as ground truth label to the optical flow prediction head of their respective ADN, and all OCTS are reshaped as $64 \times 64$ pixels.

## 3.3 Networks architectures

As presented in 4, each ADN is a separate modified implementation of the network proposed in [10]. Each network is composed of a dedicated 3D CNN encoder shared respectively by 5 and 4 proxy prediction heads. Shallow and Narrow encoder and decoder structures are borrowed from this work for the proxy tasks 1 through 4. The proxy tasks are the following.

**Task 1: Arrow of time / Doors state prediction.** The **Arrow of time task**, only implemented for the pedestrian ADN, is trained to classify an OCTS as being played forward ($X^{(T_1)} = (S_{-3},...,S_0,...,S_3)$) as opposed to being played backwards ($X^{(T_1)} = (S_3,...,S_0,...,S_{-3})$). Since pedestrians walking motion is asymmetrical with respect to time, the arrow of time is expected to be harder to predict for those showcasing an anomalous motion. The backwards prediction probability is used for anomaly scoring during inference. Since the closing and opening of doors instead appear temporally symmetrical, we replace this task by the **Doors states prediction task** for the doors ADN. It is a self-supervised task trained to classify a doors OCTS between four normal doors states (**"Start Opening" (SO)**, **"Fully Opened" (FO)**, **"Start Closing" (SC)** and **"Fully Closed" (FC)**) in addition to an anomalous state (**"Stopped Midway" (SM)**). Possible transitions between these states are presented in fig. 1(c) in the supplementary material. The four normal doors states are trained in a supervised manner while the abnormal door state SM is an OCTS artificially crafted by repeating a SO or SC state doors segmentation masks $X^{(T_1)} = (S_0,...,S_0,...,S_0)$. The prediction probability of the SM state is used for anomaly scoring during inference.

**Task 2: Motion Irregularity.** During training, this task is a binary classification between a regular and an irregular motion. The regular motion sample is an untouched OCTS while the irregular motion sample is constructed by selecting 3 randomly chosen previous frames and 3 randomly chosen succeeding frames from the current frame $F_t$. Each frame is separated by random gaps in the range $[\sigma_{min}, \sigma_{max}]$, concatenated before and after the current frame. The selection of these $\sigma_{min}$ and $\sigma_{max}$ is the subject of the "Single" experimental setting presented in sec. 4.3.3. The irregular motion prediction probability is used for anomaly scoring during inference.

**Task 3: Middle bounding box prediction.** Each ADN is trained to learn to reconstruct the central frame $S_0$ content from the previous and succeeding frames contents in the OCTS i.e. $(S_{-3},...,S_{-1},S_1,...,S_3)$. The L1 loss between the central frame and the decoder prediction is used as loss function during training and as anomaly scoring during inference.

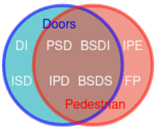
**Task 4: Model distillation.** During training, the pedestrian ADN is trained to predict the last layer features of a ResNet-50 pre-trained on ImageNet and the "person" class prediction probability provided by K-Net from the pedestrian OCTS center frame $(S_0)$. The model thus learns normal events features distribution, and we expect great discrepancies between the model and the teachers predictions for pedestrians with an unusual appearance or other objects than were wrongly detected by K-Net. During inference, similar to [10], we only use the L1 loss between the model prediction and K-Net "person" class prediction probability as anomaly scoring. This task is only learned for the pedestrian ADN as the doors ADN only processes doors segmentation masks, and not their visual appearance.

**Task 5: Optical flow prediction.** During training, each model is trained to reconstruct the norm of the full optical flow temporal sequence from the entire OCTS but the last frame $X^{(T_5)} = (S_{-3}, ..., S_2)$. Since each optical flow models the pixel level motion between two OCTS consecutive elements, this task amounts to an optical flow reconstruction for the first to the second to last optical flows and an optical flow prediction for the last one. The L1 loss between the ground truth and predicted optical flow temporal sequences are used as loss function during training and as anomaly scoring during inference.

Each ADN resulting loss function is the sum of its decoders losses and half of the knowledge distillation loss [10] for pedestrian ADN. During inference, each ADN anomaly score is the mean of its decoders anomaly scores normalized over the whole test set.

### 3.4 Fusion

The anomaly score provided by each ADN during inference is thresholded following an optimal threshold value $T$ maximizing the geometric mean of the True Positive Rate (TPR) and True Negative Rate (TNR) over the test set. Each ADN thus provides a binary classification for each frame of the evaluation set. The final interaction anomaly prediction is obtained by applying the "logical OR" operator on both outputs (fig. 2).



(a)

| Evaluated Set | Abnormal Events | | | Normal Events |
|---|---|---|---|---|
| | Doors Only | Combined | Ped Only | |
| Full | 1 | 1 | 1 | 0 |
| Doors Only | 1 | None | 0 | 0 |
| Combined | None | 1 | None | 0 |
| Ped Only | 0 | None | 1 | 0 |
| Doors | 1 | 1 | 0 | 0 |
| Ped | 0 | 1 | 1 | 0 |

(b)

Figure 1: 1(a) Venn diagram of each hazardous events dependencies w.r.t pedestrians and doors. 1(b) Ground truth annotations for each evaluation subset. The "None" annotations are subtracted from the evaluation data to avoid overlap between the "Doors Only" and "Ped" subsets and "Ped Only" and "Doors" subsets.

# 4 Experiments

### 4.1 FRailTRI20_DOD dataset description

The FRailTRI20_DOD dataset [16] is composed of video footage taken from a fish-eye camera placed on the ceiling of a train in front of train automatic doors. It showcases videos of boarding and unboarding pedestrians and opening and closing doors. A multi-label annotation is provided from each frame which includes the doors opening state and a set of anomalous events. Anomalies can either be related to doors only, such as the events **Doors Interrupted while closing (DI)** when the automatic doors get stuck due to a mechanical mishap, or **Instance Stuck in the Doorway (ISD)** when a miscellaneous object blocks the doors. Others are related to pedestrians only such as an **Interrupted Passenger Exchange (IPE)** when simultaneous boarding and unboarding passengers bump into each other or the **Fall of a Passenger (FP)**. Finally some events result from an abnormal pedestrian-doors

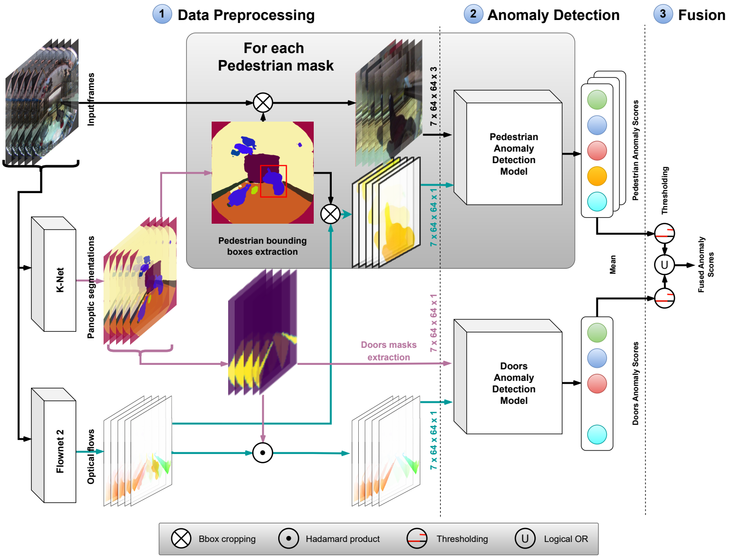
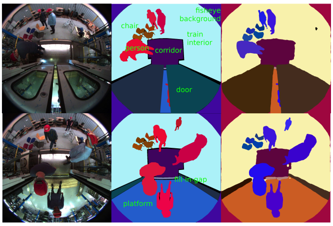


Figure 2: Full model diagram.



(a)

| Dataset | cat. | PQ | RQ | SQ |
|---|---|---|---|---|
| FPDM | person<br>door | 89.5<br>98.0 | 91.9<br>98.0 | 97.4<br>100 |
| Train door | person<br>door | 79.6<br>97.0 | 87.3<br>97.0 | 91.2<br>100 |
| combined | person<br>door | 83.3<br>97.7 | 89.1<br>97.7 | 93.6<br>100 |

(b)

Figure 3: 3(a) Input image, panoptic segmentation ground truth and K-Net predictions. 3(b) K-Net panoptic segmentation results on the FPDM, Train doors and combined datasets expressed for the categories "person" and "door".

interaction, such as an **Instance Present in the Doorway during closing (IPD)** which can become a **Passenger Stuck in the Doorway (PSD)**, a **Passenger's Bag Stuck in the Doorway while being Inside the train (BSDI)** or a **Passenger's Bag Stuck in the doorway while being on the Station (HE.BSDS)**. Anomalous events are summarized in fig. 1(a).

## 4.2 Object detector

We use the K-Net model with the swin [22] backbone variety pretrained on the COCO dataset and finetuned for 30 epochs on a combination of the FPDM dataset and an annotated subset of the FRailTRI20_DOD dataset called the Train Door dataset [9]. Panoptic segmentation results are provided in fig. 3(c) using the Panoptic Quality (PQ) metric [15], the geometric mean of the segmentation quality (SQ) and recognition quality (RQ) metrics.

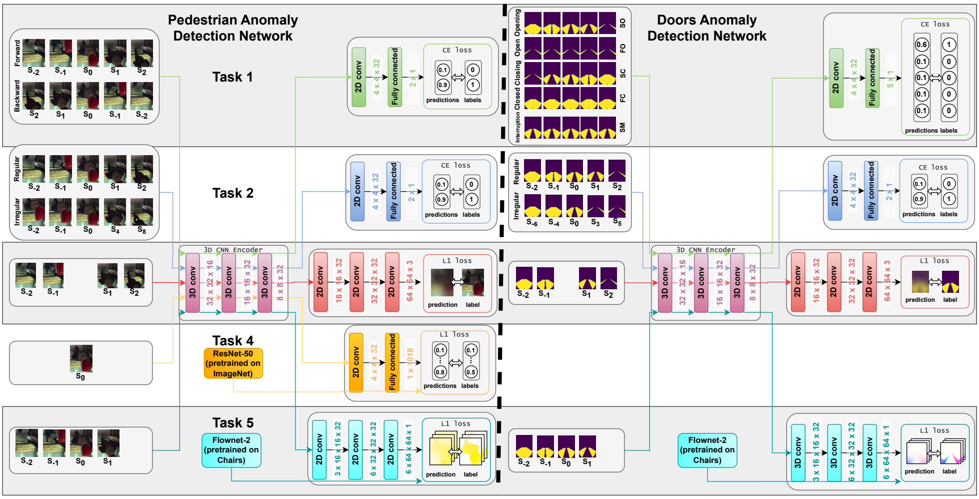


Figure 4: Pedestrians and doors anomaly detection models summaries using the shallow and narrow encoder variant. Each encoder layer is followed by a batch normalization layer, a ReLU activation and a 3D max-pooling layer and the encoder ends with a temporal pooling layer. 2D CNN Decoders layers are followed by nearest neighbors upsampling layers. 3D decoders CNN layers are followed by 3D transpose conv layers.

## 4.3 Anomaly detection networks

### 4.3.1 Evaluation metrics

We evaluate our approach using the AUC ROC (Area Under the Curve, Receiver Operating Characteristic) metric computed with respect to the frame-level ground-truth annotations. Each ADN must be evaluated on its ability to distinguish the specific subset of abnormal events it was customized to detect from normal events. We therefore design specific AUC ROC metrics for 4 different abnormal events subsets called "Doors", "Ped", "Doors Only" and "Ped only". Each subset ground truth annotations are presented in fig. 1(c).

### 4.3.2 Frames extractions

K-Net detections with a confidence higher than 0.1 and an area greater than 500 pixels are kept for the train and test sets. Similar to [10], we use the first 85% of each training sequence to train and the last 15% to validate our Ped ADN. The Doors ADN is trained and validated on a stratified shuffle split of the training sequences with respect to doors states annotations.

### 4.3.3 Individual ADN Experiments

Following [10], two series of experiments shown in table 2 are performed for each ADN.

**"Single" experiments** involve a single-single encoder-decoder scheme for each prediction head to show their independent effectiveness for anomaly detection. We also experiment with different values for the $\delta$ and $\sigma$ parameters for the prediction heads 1-3 to test them to the best of their ability. These results are reported in table 1 and the optimal models in terms of AUC ROC on the "Ped Only" (resp. "Doors") subset are transposed in table 2.

**"Combined" experiments** are single encoder-multiple decoders schemes used to evaluate the decoders performance when used in a multi-task setting. In addition, the contribution of the optical flow prediction head is evaluated for each ADN by comparing two "Combined" variants. The "Full" variant contains all prediction heads while the "Real-time" variant contains all of them safe from the optical flow prediction head. The latter does not rely on third party network predictions so as to be used in a real-time setting. The bounding boxes selection contribution is also evaluated for the ped ADN.

Each neural network is trained using mini-batches of 64 samples during 200 epochs for the "Single" experiments, 400 epochs for the "Combined" experiments, both using the Adam optimizer with an initial learning rate of $10^{-3}$ and keeping other parameters to default values.

### 4.3.4 Individual ADN Results

**"Single" experiments results** in table 1 show the importance of the choice of the $\delta$ and $\sigma$ parameters to insure the proxy tasks 1 through 3 alignment with the anomaly detection task. It is particularly visible in the case of the doors ADN tasks prediction heads 1 and 2 which results are greatly improved with a greater $\delta$ value. It seems that greater values of $\delta$ leads to sharper motion, more closely related to doors being stopped in their motion. Similarly, the task 2 prediction head systematically achieves better results when $\sigma_{min} = 0$. In these cases, the model is taught to distinguish stationary doors from moving doors.

**"Combined" experiments results** in tables 2(1) and 2(2) show that every ADN reliably reaches greater results in terms of AUC ROC than 0.5 (random chance) on their dedicated subset and less than 0.5 on their complementary subset (except for the pedestrian ADN prediction head 3). We observe great results of the Doors ADN on the Doors subset with a slight improvement in mean error and greater interactions between decoders for the "Full" variant compared to the "Real-time" variant. In the case of the pedestrian ADN, the "Full" variant using the axis-aligned bounding boxes convention achieves the best results. As such, these two networks are the ones selected for the fusion step. In addition, each punctual ADN is evaluated in terms of AUC ROC all on the "Full" subset so as to be compared to other methods of the literature in 2(3). The doors ADN on its own achieves state-of-the-art results while the pedestrian ADN does not. This is a predictable outcome since doors-related anomalies are overrepresented in the dataset, see [16].

| Models | $\delta$ | $\sigma$ | Val | | | Test | |
|---|---|---|---|---|---|---|---|
| | | | Accuracy | | MAE | AUC ROC all | |
| | | | Task 1 | Task 2 | Task 3 | Doors | P. O. |
| Task 1 | 1 | | 0.94 | | | 0.43 | 0.7 |
| | 2 | | 0.89 | | | 0.42 | 0.7 |
| | 3 | | 0.88 | | | 0.43 | 0.67 |
| | 4 | | 0.91 | | | 0.46 | 0.67 |
| | 5 | | 0.92 | | | 0.46 | 0.64 |
| Task 2 | 1 | (1-4) | | 0.9 | | 0.58 | 0.58 |
| | 1 | (0-4) | | 0.97 | | 0.58 | 0.62 |
| | 2 | (1-2) | | 0.86 | | 0.52 | 0.78 |
| | 2 | (0-2) | | 0.96 | | 0.42 | 0.88 |
| | 3 | (0-1) | | 0.99 | | 0.45 | 0.75 |
| Task 3 | 1 | | | | 0.04 | 0.39 | 0.71 |
| | 2 | | | | 0.04 | 0.32 | 0.81 |
| | 3 | | | | 0.05 | 0.29 | 0.85 |
| | 4 | | | | 0.05 | 0.28 | 0.86 |
| | 5 | | | | 0.05 | 0.28 | 0.86 |

| Models | $\delta$ | $\sigma$ | Val | | | Test | |
|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | AUC ROC all | |
| | | | Task 1 | Task 2 | Task 3 | Doors | P. O. |
| Task 1 | 1 | | 0.99 | | | 0.82 | 0.19 |
| | 2 | | 0.99 | | | 0.81 | 0.28 |
| | 3 | | 0.99 | | | 0.82 | 0.24 |
| | 4 | | 0.99 | | | 0.94 | 0.12 |
| | 5 | | 0.99 | | | 0.93 | 0.18 |
| Task 2 | 1 | (1-4) | | 0.99 | | 0.13 | 0.65 |
| | 1 | (0-4) | | 0.99 | | 0.55 | 0.53 |
| | 2 | (1-2) | | 0.99 | | 0.51 | 0.47 |
| | 2 | (0-2) | | 0.99 | | 0.93 | 0.37 |
| | 3 | (0-1) | | 0.99 | | 0.94 | 0.42 |
| Task 3 | 1 | | | | 0.99 | 0.86 | 0.3 |
| | 2 | | | | 0.99 | 0.91 | 0.3 |
| | 3 | | | | 0.99 | 0.91 | 0.24 |
| | 4 | | | | 0.99 | 0.93 | 0.31 |
| | 5 | | | | 0.99 | 0.94 | 0.25 |

Table 1: Single decoders experiments results for tasks 1 through 3 using different $\delta$ and $\sigma$ for the Ped ADN (left) and Doors ADN (right). Optimal $\delta$ and $\sigma$ values are in yellow.

### 4.3.5 Fusion experiment

The Fused Ped and Doors results are presented as the geometric mean of the TPR and TNR in table 2(4). It encompasses both the combined analysis for all anomalies and an assessment for each individual anomaly category. The evaluation metrics for each individual anomaly category are computed by isolating the abnormal events related to each considered anomaly category. Each individual ADN achieves better results than average on their dedicated anomalies (except for the fall of a passenger, FP) the fused model benefits from each ADN on all anomaly subsets for a small loss of performances.

(1) Pedestrian Anomaly Detection Network Experimental Results

| | Models | $\delta$ | $\sigma$ | Bbox Type | AUC ROC all Doors | | | | | | AUC ROC all Ped Only | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Mean | T 1 | T 2 | T 3 | T 4 | T 5 | Mean | T 1 | T 2 | T 3 | T 4 | T 5 |
| Single | Task 1 | 1 | | Axis | | 0.43 | | | | | | 0.7 | | | | |
| | Task 2 | 2 | (0-2) | Axis | | | 0.42 | | | | | | 0.88 | | | |
| | Task 3 | 5 | | Axis | | | | 0.28 | | | | | | 0.86 | | |
| | Task 4 | 1 | | Axis | | | | | 0.51 | | | | | | 0.79 | |
| | Task 5 | 1 | | Axis | | | | | | 0.32 | | | | | | 0.86 |
| Combinés | Full | | | Axis | 0.38 | 0.43 | 0.41 | 0.28 | 0.47 | 0.29 | 0.86 | 0.67 | 0.85 | 0.84 | 0.72 | 0.89 |
| | Real-time | | | Axis | 0.51 | 0.42 | 0.47 | 0.27 | 0.58 | | 0.8 | 0.74 | 0.85 | 0.84 | 0.64 | |
| | Full | | | Radius | 0.39 | 0.41 | 0.44 | 0.29 | 0.51 | 0.27 | 0.86 | 0.73 | 0.81 | 0.84 | 0.77 | 0.86 |
| | Real-time | | | Radius | 0.41 | 0.42 | 0.39 | 0.29 | 0.58 | | 0.84 | 0.74 | 0.82 | 0.84 | 0.71 | |
| | Full | | | Human | 0.34 | 0.39 | 0.42 | 0.28 | 0.6 | 0.28 | 0.85 | 0.72 | 0.78 | 0.84 | 0.72 | 0.87 |
| | Real-time | | | Human | 0.47 | 0.46 | 0.44 | 0.28 | 0.54 | | 0.8 | 0.74 | 0.81 | 0.84 | 0.71 | |

(2) Doors Anomaly Detection Network Experimental Results

| | Models | $\delta$ | $\sigma$ | AUC ROC all Doors | | | | | | AUC ROC all Ped Only | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Mean | T 1 | T 2 | T 3 | T 4 | T 5 | Mean | T 1 | T 2 | T 3 | T 4 | T 5 |
| Single | Task 1 | 5 | | | 0.93 | | | | | | 0.18 | | | | |
| | Task 2 | 3 | (0-1) | | | 0.94 | | | | | | 0.42 | | | |
| | Task 3 | 5 | | | | | 0.94 | | | | | | 0.25 | | |
| | Task 5 | 1 | | | | | | | 0.93 | | | | | | 0.37 |
| Co. | Full | | | 0.96 | 0.93 | 0.95 | 0.94 | | 0.92 | 0.28 | 0.18 | 0.27 | 0.25 | | 0.33 |
| | Real-time | | | 0.96 | 0.89 | 0.9 | 0.95 | | | 0.28 | 0.22 | 0.42 | 0.23 | | |

(3) Punctual ADN comparisons

| AUC ROC all Full | |
|---|---|
| Doors ADN | 0.87 |
| Ped ADN | 0.47 |
| [16] | 0.8 |
| [17] | 0.82 |

(4) Fusion Experimental Results, fused binary responses optimal PR threshold

| Models | Doors Only | | Combined | | | | Ped Only | |
|---|---|---|---|---|---|---|---|---|
| | DI | ISD | PSD | BSDI | BSDS | IPD | IPE | FP |
| Doors ADN | 0.86 | 0.97 | 0.91 | 0.71 | 0.97 | 0.73 | 0 | 0 |
| Ped ADN | 0 | 0.01 | 0.01 | 0.02 | 0 | 0.03 | 0.67 | 0.17 |
| Fused | 0.85 | 0.96 | 0.9 | 0.71 | 0.96 | 0.73 | 0.66 | 0.17 |

Table 2: "Single" and "Combined" experiment results for the Ped (1) and Doors (2) ADN. Ped and Doors combined ADN variants selected for the fusion step are in green, orange. (3) Punctual ADNs compared to methods of the literature on the "Full" abnormal events. Both methods are based on [25]. (4) Fusion experiment results for each anomalous event.

# 5 Conclusion

In this paper, we propose a multi object-centric networks architecture to deal with anomaly detection in human object interactions. We validate and evaluate our architecture to detect doors-pedestrian anomalous interactions in a railway application context. We also present two new proxy tasks and show their effectiveness and limitations for each ADN. We performed a thorough study of the impact of the choice of bounding boxes convention and the impact of preprocessing on the misalignment between the self-supervised proxy tasks and the anomaly detection task. In future work, we will explore other fusion strategies to fill the performance gap between the pedestrian and the doors anomaly detection networks. Adding trained connections between the streams will be investigated.

# References

[1] AnoPCN | Proceedings of the 27th ACM International Conference on Multimedia, . URL https://dl.acm.org/doi/10.1145/3343031.3350899.

[2] Continual Learning for Anomaly Detection in Surveillance Videos, . URL https://www.computer.org/csdl/proceedings-article/cvprw/2020/09150686/1lPHfcdfpDi.

[3] Antonio Barbalau, Radu Tudor Ionescu, Mariana-Iuliana Georgescu, Jacob Dueholm, Bharathkumar Ramachandra, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B. Moeslund, and Mubarak Shah. SSMTL++: Revisiting self-supervised multi-task learning for video anomaly detection. *Computer Vision and Image Understanding*, 229:103656, March 2023. ISSN 1077-3142. doi: 10.1016/j.cviu.2023.103656. URL https://www.sciencedirect.com/science/article/pii/S107731422300036X.

[4] Liyang Chen, Zhiyuan You, Nian Zhang, Juntong Xi, and Xinyi Le. UTRAD: Anomaly detection and localization with U-Transformer. *Neural Networks*, 147:53–62, March 2022. ISSN 0893-6080. doi: 10.1016/j.neunet.2021.12.008. URL https://www.sciencedirect.com/science/article/pii/S0893608021004810.

[5] Yong Shean Chong and Yong Haur Tay. Abnormal Event Detection in Videos Using Spatiotemporal Autoencoder. volume 10262, pages 189–196, Cham, 2017. Springer International Publishing. ISBN 978-3-319-59080-6 978-3-319-59081-3. doi: 10.1007/978-3-319-59081-3_23. URL http://link.springer.com/10.1007/978-3-319-59081-3_23. Book Title: Advances in Neural Networks - ISNN 2017 Series Title: Lecture Notes in Computer Science.

[6] Fei Dong, Yu Zhang, and Xiushan Nie. Dual Discriminator Generative Adversarial Network for Video Anomaly Detection. *IEEE Access*, 8:88170–88176, 2020. ISSN 2169-3536. doi: 10.1109/ACCESS.2020.2993373. Conference Name: IEEE Access.

[7] Keval Doshi and Yasin Yilmaz. Any-Shot Sequential Anomaly Detection in Surveillance Videos. pages 4037–4042. IEEE Computer Society, June 2020. ISBN 978-1-72819-360-1. doi: 10.1109/CVPRW50498.2020.00475. URL https://www.computer.org/csdl/proceedings-article/cvprw/2020/09151050/1lPHyrhzTPO.

[8] Zhihao Duan, M. Ozan Tezcan, Hayato Nakamura, Prakash Ishwar, and Janusz Konrad. RAPiD: Rotation-Aware People Detection in Overhead Fisheye Images. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2700–2709, Seattle, WA, USA, June 2020. IEEE. ISBN 978-1-72819-360-1. doi: 10.1109/CVPRW50498.2020.00326. URL https://ieeexplore.ieee.org/document/9150668/.

[9] Rémi Dufour, Cyril Meurie, Clément Strauss, and Olivier Lézoray. Instance segmentation in fisheye images. In *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6, November 2020. doi: 10.1109/IPTA50016.2020.9286623. ISSN: 2154-512X.

[10] Mariana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Khan, Marius Popescu, and Mubarak Shah. Anomaly Detection in Video via Self-Supervised and Multi-Task Learning. pages 12737–12747, June 2021. doi: 10.1109/CVPR46437.2021.01255.

[11] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, and Larry S. Davis. Learning Temporal Regularity in Video Sequences, April 2016. URL https://arxiv.org/abs/1604.04574v1.

[12] 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 *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1647–1655, July 2017. doi: 10.1109/CVPR.2017.179. ISSN: 1063-6919.

[13] Radu Tudor Ionescu, Sorina Smeureanu, Bogdan Alexe, and Marius Popescu. Unmasking the abnormal events in video, May 2017. URL https://arxiv.org/abs/1705.08182v3.

[14] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. Object-Centric Auto-Encoders and Dummy Anomalies for Abnormal Event Detection in Video. June 2019. doi: 10.1109/CVPR.2019.00803. URL https://www.scinapse.io.

[15] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollar. Panoptic Segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9396–9405, Long Beach, CA, USA, June 2019. IEEE. ISBN 978-1-72813-293-8. doi: 10.1109/CVPR.2019.00963. URL https://ieeexplore.ieee.org/document/8953237/.

[16] Olivier Laurendin, Sébastien Ambellouis, Anthony Fleury, Ankur Mahtani, Sanaa Chafik, and Clément Strauss. Hazardous Events Detection in Automatic Train Doors Vicinity Using Deep Neural Networks. In *2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–7, November 2021. doi: 10.1109/AVSS52988.2021.9663863.

[17] Olivier Laurendin, Anthony Fleury, Sebastien Ambellouis, Sanaa Chafik, and Ankur Mahtani. Deep Hazardous Events Detection in Top-Down Fish-Eye Images for Railway Applications. June 2022.

[18] Yunseung Lee and Pilsung Kang. AnoViT: Unsupervised Anomaly Detection and Localization with Vision Transformer-based Encoder-Decoder, March 2022. URL https://arxiv.org/abs/2203.10808v1.

[19] Ruikang Liu, Weiming Liu, Mengfei Duan, Liang Wang, Liang Mao, Qisheng Qiu, and Guangzheng Ling. Intercity Rail Transit Platform Anomaly Detection Using Door Tracking-Based Key Frame Extraction and AnoDet Network. *IEEE Transactions on Instrumentation and Measurement*, 72:1–16, 2023. ISSN 1557-9662. doi: 10.1109/TIM.2023.3269755. Conference Name: IEEE Transactions on Instrumentation and Measurement.

[20] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future Frame Prediction for Anomaly Detection – A New Baseline. *arXiv:1712.09867 [cs]*, March 2018. URL http://arxiv.org/abs/1712.09867. arXiv: 1712.09867.

[21] Yusha Liu, Chun-Liang Li, and B. Póczos. Classifier Two Sample Test for Video Anomaly Detections. 2018. URL https://www.semanticscholar.org/paper/Classifier-Two-Sample-Test-for-Video-Anomaly-Liu-Li/6869ab1f42e30b415829de9928f7e4a606113601.

[22] 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, August 2021. URL http://arxiv.org/abs/2103.14030. arXiv:2103.14030 [cs].

[23] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal Event Detection at 150 FPS in MATLAB. page 8.

[24] Weixin Luo, Wen Liu, and Shenghua Gao. Remembering history with convolutional LSTM for anomaly detection. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 439–444, July 2017. doi: 10.1109/ICME.2017.8019325. ISSN: 1945-788X.

[25] Trong Nguyen Nguyen and Jean Meunier. Anomaly Detection in Video Sequence With Appearance-Motion Correspondence. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1273–1283, October 2019. doi: 10.1109/ICCV.2019.00136. ISSN: 2380-7504.

[26] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning Memory-guided Normality for Anomaly Detection, March 2020. URL http://arxiv.org/abs/2003.13228. arXiv:2003.13228 [cs].

[27] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context Encoders: Feature Learning by Inpainting, November 2016. URL http://arxiv.org/abs/1604.07379. arXiv:1604.07379 [cs].

[28] Jonathan Pirnay and Keng Chai. Inpainting Transformer for Anomaly Detection. In Stan Sclaroff, Cosimo Distante, Marco Leo, Giovanni M. Farinella, and Federico Tombari, editors, *Image Analysis and Processing – ICIAP 2022*, Lecture Notes in Computer Science, pages 394–406, Cham, 2022. Springer International Publishing. ISBN 978-3-031-06430-2. doi: 10.1007/978-3-031-06430-2_33.

[29] Mahdyar Ravanbakhsh, Moin Nabi, Enver Sangineto, Lucio Marcenaro, Carlo Regazzoni, and Nicu Sebe. Abnormal event detection in videos using generative adversarial nets. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1577–1581, September 2017. doi: 10.1109/ICIP.2017.8296547. ISSN: 2381-8549.

[30] Joseph Redmon and Ali Farhadi. YOLOv3: An Incremental Improvement. *arXiv:1804.02767 [cs]*, April 2018. URL http://arxiv.org/abs/1804.02767. arXiv: 1804.02767.

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *arXiv:1506.01497 [cs]*, January 2016. URL http://arxiv.org/abs/1506.01497. e1;e2;e3;e4;e5;e6;5 FPS ; 1xGPU 300 region proposals; https://github.com/rbgirshick/py-faster-rcnn.

[32] Masato Tamura, Shota Horiguchi, and Tomokazu Murakami. Omnidirectional Pedestrian Detection by Rotation Invariant Training. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1989–1998, January 2019. doi: 10.1109/WACV.2019.00216. ISSN: 1550-5790.

[33] Yao Tang, Lin Zhao, Shanshan Zhang, Chen Gong, Guangyu Li, and Jian Yang. Integrating prediction and reconstruction for anomaly detection. *Pattern Recognition Letters*, 129:123–130, January 2020. ISSN 0167-8655. doi: 10.1016/j.patrec.2019.11.024. URL https://www.sciencedirect.com/science/article/pii/S0167865519303447.

[34] Mohamed Thioune, Sanaa Chafik, Ankur Mahtani, Olivier Laurendin, and Safia Boudra. FPDM: Fisheye Panoptic segmentation dataset for Door Monitoring. In *2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8, November 2022. doi: 10.1109/AVSS56176.2022.9959151.

[35] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. Learning Deep Representations of Appearance and Motion for Anomalous Event Detection, October 2015. URL https://arxiv.org/abs/1510.01553v1.

[36] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Reconstruction by inpainting for visual anomaly detection. *Pattern Recognition*, 112:107706, April 2021. ISSN 0031-3203. doi: 10.1016/j.patcog.2020.107706. URL https://www.sciencedirect.com/science/article/pii/S0031320320305094.

[37] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-Net: Towards Unified Image Segmentation.

[38] Yiru Zhao, Bing Deng, Chen Shen, Yao Liu, Hongtao Lu, and Xian-Sheng Hua. Spatio-Temporal AutoEncoder for Video Anomaly Detection. In *Proceedings of the 25th ACM international conference on Multimedia*, MM '17, pages 1933–1941, New York, NY, USA, October 2017. Association for Computing Machinery. ISBN 978-1-4503-4906-2. doi: 10.1145/3123266.3123451. URL https://doi.org/10.1145/3123266.3123451.