IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

IEEE SYSTEMS, MAN AND CYBERNETICS SOCIETY SECTION

Received 20 May 2022, accepted 6 June 2022, date of publication 14 June 2022, date of current version 23 June 2022.

*Digital Object Identifier 10.1109/ACCESS.2022.3183102*

# A Survey on Audio-Video Based Defect Detection Through Deep Learning in Railway Maintenance

**LORENZO DE DONATO[1], FRANCESCO FLAMMINI[2], (Senior Member, IEEE), STEFANO MARRONE[1], CLAUDIO MAZZARIELLO[3], ROBERTO NARDONE[4], CARLO SANSONE[1], (Member, IEEE), AND VALERIA VITTORINI[1]**

[1]Department of Electrical Engineering and Information Technology, University of Naples Federico II, 80125 Naples, Italy
[2]Department of Computer Science and Media Technology, Linnaeus University, 351 95 Växjö, Sweden
[3]Digital and Data Driven Innovation Unit, Hitachi Rail STS, 80147 Naples, Italy
[4]Department of Engineering, University of Naples "Parthenope," 80143 Naples, Italy

Corresponding author: Francesco Flammini (francesco.flammini@lnu.se)

This research has received funding from the Shift2Rail Joint Undertaking (JU) under grant agreement No 881782 RAILS (Roadmaps for Artificial Intelligence (A.I.) integration in the raiL Sector). The JU receives support from the European Union's Horizon 2020 research and innovation programme and the Shift2Rail JU members other than the Union.

**ABSTRACT** Within Artificial Intelligence, Deep Learning (DL) represents a paradigm that has been showing unprecedented performance in image and audio processing by supporting or even replacing humans in defect and anomaly detection. The railway sector is expected to benefit from DL applications, especially in predictive maintenance applications, where smart audio and video sensors can be leveraged yet kept distinct from safety-critical functions. Such separation is crucial, as it allows for improving system dependability with no impact on its safety certification. This is further supported by the development of DL in other transportation domains, such as automotive and avionics, opening for knowledge transfer opportunities and highlighting the potential of such a paradigm in railways. In order to summarize the recent state-of-the-art while inquiring about future opportunities, this paper reviews DL approaches for the analysis of data generated by acoustic and visual sensors in railway maintenance applications that have been published until August 31st, 2021. In this paper, the current state of the research is investigated and evaluated using a structured and systematic method, in order to highlight promising approaches and successful applications, as well as to identify available datasets, current limitations, open issues, challenges, and recommendations about future research directions.

**INDEX TERMS** Computer vision, machine learning, fault detection, inspection, CNN, smart railways.

## I. INTRODUCTION

Railway components have a long operating life, often spanning over decades, with installations that may span over hundreds of kilometres and need to withstand harsh environmental conditions. Therefore, maintaining a railway in efficient working conditions is a task requiring a considerable amount of resources. Automating inspection, condition monitoring and eventually repairs to railway assets may lead to more efficient use of human skills and time as well as reduce maintenance mistakes and costs. Artificial Intelligence (AI) can help to analyse specific railway components through diverse sensors [1], and some works investigate to what extent it is possible to use non-intrusive data sources (e.g., LiDAR [2], X-Ray [3], light sensors [4], and fiber optic [5]) to improve maintenance effectiveness and efficiency. Among all, audio- and video-based inspection can provide effective means for observing the status of railway assets and selectively identifying items that need further attention. Indeed, automated and continuous audio-video inspection opens unprecedented ways to cope with railway maintenance that currently is the most investigated AI area in railways [6]. Several European projects have tackled the challenge to investigate and experiment with the adoption of AI techniques in the maintenance and inspection of railway systems, for example, the recent H2020 Shift2Rail (S2R) projects IN2SMART [7] and IN2SMART2 [8].

It is a matter of fact that current railway maintenance has been focusing on the use of special sensors, and many diagnostic tasks will continue to be addressed using those mature

The associate editor coordinating the review of this manuscript and approving it for publication was Huiyan Zhang.

methods and technologies. However, those approaches might be rather expensive in terms of infrastructures and data analysis. Moreover, inspection activities are often performed on a fixed schedule, possibly resulting in late anomaly detection. The use of those sensors may also be problematic in terms of intrusiveness due to the need for ''instrumenting'' components, pausing railway traffic, etc. In such a context, audio and video sources seem to have a huge potential and several advantages, as i) they do not need to be installed on the object they have to monitor, thus limiting their intrusiveness and any need for (re)certification of railway components, and ii) they might be needed or already installed for other reasons (e.g. safety/security surveillance [9]), thus enabling multiple uses of technologies for higher cost-effectiveness. In particular, Deep Learning (DL) represents a paradigm that has shown unprecedented performance in image and audio processing by supporting humans in defect and anomaly detection [10].

Therefore, although other types of non-intrusive sensors exist, this paper aims to provide a structured survey on the DL approaches for railway maintenance that leverage video (i.e., sequences of frames) and/or audio data. In doing so, we borrowed and tailored to our scope some practices from Systematic Literature Review (SLR) methodologies [11], [12], such as the definition of research questions (RQ), well-structured queries, and exclusion criteria. In particular, in this study we aim at answering the following research questions:

- RQ1 Which are the main railway applications and objectives addressed by current research on DL for audio and video analysis applied to railway maintenance?
- RQ2 Which specific DL approaches and datasets have been used in existing audio-video analytics applications for railway maintenance? Are those datasets publicly accessible?
- RQ3 What are the main open issues, challenges and opportunities in the field of DL applied to railway maintenance and inspection with audio and video sensors?

Although different terms with similar meanings are used in the literature to refer to maintenance activities supported by technologies such as the Internet of Things (IoT), cloud computing and data analytics, including *predictive maintenance*, *smart maintenance* and *computer*-aided maintenance, in this paper we will adopt those terms according to the usage made by the authors of the surveyed paper. Otherwise, we will simply use the word ''maintenance''. We emphasize that this work has been developed within the RAILS[1] H2020 research project [13], whose main objective is to provide recommendations and research directions for a fast take-up of AI in railways.

A structured literature review can be presented in several ways, according to the points of view chosen to cluster and discuss the findings. We have chosen to organize the results according to reference railway areas and thus provide one section for each area, from Section IV to VII. The internal organization of these sections is homogeneous, hence it is possible to read the paper focusing on specific areas of interest. Those sections summarize the findings of the review, and they are preceded by the following sections: Section II presents existing review papers surveying DL applications to maintenance; Section III provides background information, including a description of the reference railway areas, and presents the organization of their related sections. The presentation of the review findings is followed by: Section VIII, which reports some overall statistics on the findings; Section IX, which discusses the responses to the research questions; and Section X, which summarizes and discusses the main results, challenges, opportunities, and research directions for effective utilization of AI techniques in the railway domain. Finally, Section XI provides some closing remarks.

[1]Roadmaps for AI integration in the raiL Sector - https://rails-project.eu/

## II. RELATED WORK

The last few years have seen a growing interest in AI applications to railway systems. This trend is attested by industrial research and innovation initiatives, as well as by the growing number of scientific publications (Fig. 2), addressing the application of AI techniques to the rail sector. Technologies such as computer vision and audio processing, especially fostered by Machine Learning (ML) and DL approaches, will play an important role in providing effective methods to solve various problems, including intelligent surveillance, automatic train operation, timetable optimization and network management. Different authors tried to organise these works based on the task or application in the railway domain.

**Rail track maintenance** is among the most analysed aspects, with works focusing on both corrective and predictive studies. Focusing on the former, reference [14] reviews papers using both shallow and deep neural networks to detect structural defects, including those due to physical deterioration of the tracks, or geometrical irregularities (e.g., misalignment). The review concludes that, although DL algorithms have been widely adopted in recent years to identify structural defects, there are still shortcomings to address, such as i) the lack of benchmarks, ii) the lack of labelled datasets, and iii) the difficulty of finding defective observations. Moving to the latter, reference [15] reports a detailed SLR focused on data-driven methods (including ML and DL). The review was published in October 2020 and analyses about 109 papers covering the period from 2014 to 2019. Among the results of the SLR, it has been concluded that the application of data-driven models can help in avoiding the unnecessary replacement of track components. *Nevertheless, DL approaches accounted for only 8% of the reviewed models.* Two hot topics have been highlighted : i) the need to develop an automatic data labelling method (currently carried out manually in most cases) and ii) the importance of interpretability of black-box models.

**Predictive maintenance** has been also analysed in a broader sense in [16], where the authors provided a wide-spectrum SLR of suitable ML approaches. The analysis,

conducted on papers published from 2009 to 2018, highlights that predictive maintenance attracted an increasing interest from researchers starting from 2013. The work shows that only a few papers propose solutions specifically tailored to railway systems and suggests that appropriate sensors could help to avoid unnecessary replacements and hence lead to savings while ensuring safety, availability, and efficiency of maintenance processes. It is worth noting that a similar conclusion had been already stated in another work, dating back to 2010 [17].

An SLR on **Prognostic and Health Management** (PHM) has been presented in [18], where for each PHM subfield, i.e., Fault Detection, Fault Diagnosis, and Prognosis (Remaining Useful Life estimation), papers have been classified according to i) the DL approach (including convolutional/recurrent neural networks, autoencoders, etc.) and ii) the type of dataset they rely on (vibrational, time-series, imagery, and structured). The paper also emphasizes DL challenges concerning model complexity and its characterization, the lack of labelled data, and the fact that most of the papers were about datasets gathered from bench-scale experiments that possibly lead to poor applicability in the real world. It is also worth noting that no image-based approaches have been identified for prognosis.

More recently, researchers are also focusing on the use of **computer vision supported by ML and DL.** In [19], the authors provided a review of obstacle detection based on AI and DL for **risk reduction at Level Crossings**. The paper analyses the combination of obstacle detection with intelligent decision making as an opportunity to provide robust interlocking decisions. Another recent work [20] focuses on visual inspection based on image processing, including traditional and ML-based techniques, without a clear focus on DL.

All the works mentioned above provide literature reviews with different degrees of abstraction and focus. In some cases, DL has been analysed as one of the possible solutions, while in others the authors focused on different transport domains, data sources, maintenance topics, or specific railway areas. *To the best of our knowledge, no literature review specifically addresses DL for audio-video analytics in railway maintenance.*

## III. BACKGROUND AND REVIEW STRUCTURE

In this section, we provide background about railway areas and components that are relevant for maintenance, as well as an overview of DL approaches and maintenance activities.

### A. RAILWAY AREAS

Based on the results of our review, current research on railway maintenance focuses on the following main areas and components:

- *Rail Track*: the core element of the railway system infrastructure. It includes rails, switch points, sleepers, fasteners, and ballasts. It is also worth mentioning that the rails are, in some points, joined with welds [21];
- *Pantograph & Catenary*: the Pantograph and Catenary (PAC) system has a central role in electrical trains since they provide them with the necessary power to operate. The catenary is the fixed structure located near and upon (overhead line) the rail track; the pantograph is a component of the trains placed on the roof. Trains get the power through the contact between the pantograph slide plates and the catenary contact wire. Catenaries are composed of a high number of elements (e.g., screws, bolts, etc.). In section V, we divided this area into *Pantograph and Arcs*, *Catenary Support Device (CSD)*, and *Catenary Wires*;
- *Rolling Stock*: within rolling stocks, Bogies are undercarriages with four or six wheels pivoted beneath the end of a vehicle that could also provide traction and braking, while Electric Multiple Units (EMU) are multiple-unit trains consisting of self-propelled carriages using electricity as the motive power. Such category also includes a few other components of trains' body and frame (i.e., chassis);
- *Tunnel & Bridge*: encompassing all components of tunnels and bridges, including concrete structure and lining.

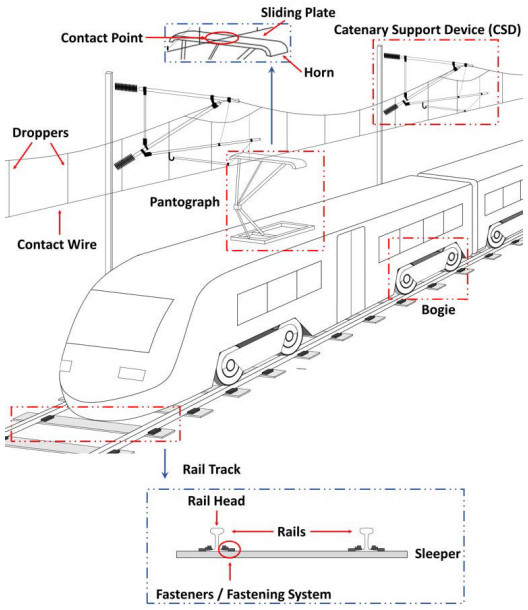



**FIGURE 1. Overview of railway components. The figure highlights the main components addressed by the research works surveyed in this paper.**

Fig. 1 provides an at-a-glance view of a railway system.

### B. DEEP LEARNING

Deep Learning is a subfield of Machine Learning encompassing a multitude of algorithms and models, including Convolutional Neural Networks (CNN), which specialise in image processing by learning how to extract a suitable set of features for a specific task [14]. Three are the main families of image processing tasks:

- Classification (C), in which a sample has to be assigned to a class within a fixed set of alternatives. This represents the earliest task CNNs have been used for,

as classification is commonly used in computer vision. Tens of models have been proposed, with some remarkable examples such as AlexNet [22], VGG [23], ResNet [24], DenseNet [25], etc. In this study we refer to those as state-of-the-art (SOTA) networks;
- Object Detection (OD), where CNN features are used to localise – using a bounding box – objects of interest within a scene. Literature includes many SOTA CNN-based models for OD, such as region-based CNN (e.g., R-CNN [26], Faster R-CNN [27], Mask R-CNN [28]) and those belonging to the You Only Look Once (YOLO) family (e.g., YOLO [29], YOLOv2 [30], YOLOv3 [31]);
- Semantic Segmentation (SS), which is the precise pixel-wise assignment of portions of an image to a given class. These architectures are often similar to autoencoders, as their output is a segmentation map having the same size as the input image. Very famous examples are the SegNet [32] and U-Net [33] architectures.

Interestingly, CNN can also be used for the analysis of audio signals. In this case, the idea is to analyse the frequency-phase spectrum as if they were images. This naive approach not only proved to be effective, but also allowed for the re-use of CNN architectures originally intended for images [34].

### C. MAINTENANCE TASKS

Despite the railway domain encompasses several maintenance tasks, all the works covered by this survey fall into two main categories:

- *Surface Defect Detection (SDD)*, focusing only on surface defects (e.g., cracks in tunnels, rails' heads, etc.);
- *Defect Inspection (DI)*, including all models intended to detect and analyse different types of faults (e.g., broken objects, missing parts, anomalies in components, etc.).

It is worth noting that we also reviewed papers that, although totally focused on the railway domain, are intended to support a maintenance process and not to address it (e.g., detection and/or classification of railway-related objects). We decided to keep those papers (gathered under the name ''Promising works'') in this survey, as they represent promising foundations for future works on maintenance.

### D. REVIEW METHODOLOGY

To evaluate the current status of the research and answer the questions highlighted in Section I, we followed the SLR methodology proposed in [12]. Therefore, we defined a research query by combining keywords from four sets :

- Railway domain: *"rail*", "metro", "subway";*
- Maintenance and defect/fault detection: *"smart maintenance", "predictive maintenance", "condition based maintenance", "fault diagnosis", "fault detection", "fault forecast*", "fault prediction", "fault prevention", "defect inspection", "defect detection", "health monitoring", "health status", "remaining useful life", "condition monitoring"*;
- Deep Learning: *"deep learning", "deep reinforcement", "deep neural", "convolutional neural network", "CNN*", "recurrent neural network", "LSTM", "autoencoder*", "generative adversarial", "GAN"*;
- Image/ audio processing: *"computer vision", "vision", "image*", "video*", "sound", "acoustic", "audio"*.

The final search query is given by the OR among keywords of each set, combined with the AND of all sets. These keywords were selected by performing empirical trials in order to cover as many studies as possible when submitting the query to the considered scientific literature databases: IEEE Xplore Digital Library,[2] Scopus,[3] and ACM Digital Library.[4] Notably, given the limited number of *wildcards* (i.e., asterisks), we opted for some generic keywords (e.g., ''deep neural'') and not for specific ones (e.g. ''deep neural network'') to include possible keyword variations (e.g., ''network'', ''network**s**'', ''architecture'', ''architecture**s**'') that might not be encompassed otherwise. Lastly, the query was submitted to search within the *title*, *abstract*, and *keywords* data fields. Then, to filter the results, we applied the following exclusion criteria:

E1 works that are not written in English;
E2 works not related to maintenance and DL;
E3 works not focusing on the rail sector;
E4 works that present results leveraging data coming from sensors different from cameras or audio devices;
E5 works that are non-experimental;
E6 literature reviews.

The query was submitted to cover all years until August 31st, 2021. However, no relevant paper was found before 2014. Also, we removed all duplicated papers found in different libraries. After the application of all exclusion criteria, we got 95 relevant papers. Fig. 2 reports the publication trend per year; please note that the 13 papers published in 2021 are not included in the figure because the partial coverage of the year (until August 31st) would have led to a misleading trend. Fig. 3 shows the distribution of papers in the reference areas.

### E. REVIEW ORGANIZATION

The survey is structured into four sections, one for each relevant railway area: *Rail Track* (section IV), *Pantograph & Catenary* (section V), *Rolling Stock* (section VI), *Tunnel & Bridge* (section VII). For each section, we provide two tables: the first to gather the papers under the three groups identified in Section III-C (SDD, DI and Promising works), the second to summarise the maintenance task, DL approach, data acquisition system, dataset and obtained performance. The reported data is taken by preserving as much as possible the original format from the corresponding paper. Although this causes some inconsistencies (e.g., using different performance metrics in the same column) this ensures a fair and

[2]https://ieeexplore.ieee.org
[3]https://www.scopus.com
[4]https://dl.acm.org

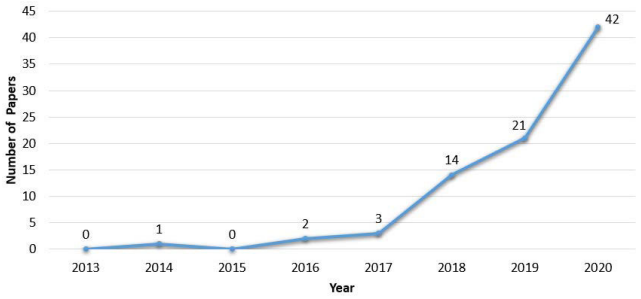



**FIGURE 2.** Publication trend of relevant papers from 2014 to 2020.

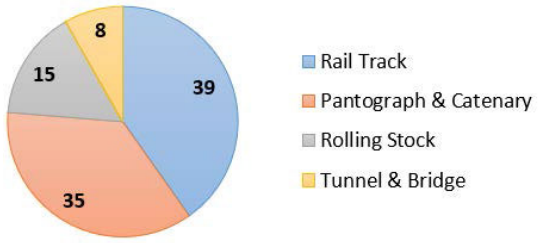



**FIGURE 3.** Overall distribution of papers. Notably, the count exceeds the number of papers as [35] addresses problems in both the railway areas of "Rail Track" and "Rolling Stock"; while [36] addresses problems in both the railway areas of "Rail Track" and "Tunnel & Bridge". Works addressing "Rail Track" and "Pantograph & Catenary" represent almost the 75% of the total.

transparent analysis. Nonetheless, for the sake of correctness, we omitted misleading results when possible. For example, we did not report the accuracy for semantic segmentation or for object detection tasks when other, fairer metrics, such as the IoU (Intersection over Union) or the mAP (mean Average Precision) were available. Those tables can be used to have all the important information at a glance, while some fine-grained details (where available) will be provided within the subsections.

## IV. RAIL TRACK

We found 39 research papers, reported in Tables 1 and 2, facing track-related issues introducing innovative methods to monitor and/or analyse various Rail Track components. It is worth noting that a given task (e.g., Defect Inspection), on a specific subject (e.g., fasteners) can be performed through object detection, semantic segmentation, classification, or a combination of them. For example, Surface Defect Detection can be performed by classifying the images in accordance with the labels or by detecting the defect within the image. Thus, the same reference may appear under different columns.

### A. RAILS' HEADS

Rails' heads surface defect detection is addressed by several papers. To classify cavities or small hills on the surface of rails' heads, in [41] a CNN-based classifier is trained on photometric stereo images while in [42] the authors used 3D laser camera data. Reference [63] proposes a CNN-based approach

**TABLE 1.** Maintenance tasks for rail track, by components.

| *Components* | *Surface Defect Detection* | *Defect Inspection* | *Promising Works* |
|---|---|---|---|
| Rails | [37] | [38], [39], [40], [35] | |
| Rails' Heads | [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [61], [62] | [63]†, [64] | |
| Fasteners/ Fastening Systems | | [48], [65], [66], [67], [68], [69], [70], [71] | [72], [55] |
| Welded Joints | | | [21]† |
| Sleepers | [36] | | |

The † mark indicates data obtained from specimens only (laboratory tests)

to classify rails' health status based on CNN and Acoustic Emission recognition. The tests have been obtained by stressing a rail test specimen until it broke, recording and labelling the audio events. Reference [54] proposes an architecture based on CNNs to identify surface rail defects at switch points. The proposed network leverages transfer learning on a dataset collected through a monitoring system previously developed by the same authors [75]–[77]. Reference [43] introduces a three-stage pipeline: the first stage detects and removes blurring from the images based on an Inertial Measurement Unit and on the Attitude and Heading Reference Systems algorithm; in the second stage, a CNN built from scratch was trained to detect surface defects (healthy or faulty image); in the third stage, the CNN was tested on new images. Experiments show that the whole process could be performed up to 40 frames per second which means a real-time operating speed of around 72Km/h. Reference [44] makes use of images collected by a robot running on the rail line. Local image processing and classification are performed onboard (on-the-fly) firstly. Once the onboard CNN architecture detected a defective image, the location is saved and the image is post-processed in the cloud. Reference [49] proposes an approach based on YOLOv3 architecture to improve the detection speed (about 0.15s), preserving a recognition rate of 97%. Training and test have been performed on images captured by a camera installed perpendicular to the rail. Reference [45] introduces a new network intended to operate (also) under poor lighting conditions. The model combines YOLOv3 for regression, multi-scale prediction and the loss computation method, with a feature extractor based on MobileNetV2 [86]. The resulting system operates up to 60 FPS. References [50] and [52] evaluate Faster R-CNN models against the use of other architectures as feature extractors. To eliminate the manual inspection, reference [64] proposes an IoT-based multi-robot able to detect rail tacks' defects in real-time. Multiple robots were involved in collecting data about defects through ultrasonic sensors and cameras. For surface cracks, a CNN is compared against SVM, ANN, and Random Forest trained with data obtained through the Speed Up Robust

**TABLE 2. Dataset details for rail track.**

| Paper | Availability | Collection Method | Task | ID | Dataset | Model | Performance |
|---|---|---|---|---|---|---|---|
| [37] | Proprietary | 4 CLCs surrounding the rail from a different corner | C | - | 10427 images (4096x3000 px) cropped to 256x256 px sub-images grouped into 7 classes (Non defective, Roll mark, Rolled-in scrap, Lack of material, Straightening mark, Wire, Other) | CNN (custom) | F1: 67.57-67.69% |
| [38] | Proprietary | RT18-D double track rail ultrasonic flaw detection vehicle | OD | - | 1500 B-scan images grouped in 2 classes (Normal srew hole, Screw hole crack) | LSTM (custom) | RR: >95% |
| [40] | Proprietary | Tensile testing machine (Zwick Z100) to stress rail specimen + Vallen AMSY-6 ASIP-2/A acquisition system ( 5 MHz) | C | - | 3200 crack non-overlapped signals each of which has 2048 sampling points with a length of 0.4096ms. Signals are grouped into 2 classes (Safe state, Unsafe) | LSGAN (custom) | SSIM: 0.8992 |
| [35] | Proprietary | Behringer B-5 sensor + Behringer U Phoria UMC204HD (24 Bit 92kHz resolution) | C | - | 228 30s audios grouped into 4 classes (dry_40, dry_60, wet_40, wet_60) | FCNN (custom) | ACC: 100% |
| [41] | Proprietary | Dark-field acquisition setup with photometric illumination (1 colour camera + 2 oblique lights) | C | - | 16x16 px patches (extracted from 2532 images) grouped into 2 classes (Normal, Defective) | CNN (custom) | ER: 0.556% |
| [42] | Proprietary | AT C5-1600CS19-500 3D laser camera with laser triangulation | C | - | 3D rail profile images (1600x50 px) grouped into 2 classes (Normal, Defective) | CNN (custom) | ACC: 98% |
| [51] | Type-I [73], [74] | As described in [73], [74] | SS | - | 67 images (201 after augmentation) grouped into 2 classes (Normal, Defective) | U-Net | ACC: 99.76% |
| [43] | Proprietary | Mako G032B/C Cameras + Xsens MTI-100 IMU data (for deblurring) | C | - | 1700 images (658x492 px) cropped to 100x30 px images grouped into 2 classes (Normal, Defective) | CNN (custom) | F1: 98% |
| [44] | Type-I [73], [74] | As described in [73], [74] | C | - | The original 67 images (1000x160 px) of the Type-I dataset were cropped into160x120 px sub-images and grouped into 2 classes (Normal, Abnormal) | CNN (custom) | ACC: 97% |
| [45] | Proprietary | Not specified by the authors | OD | - | 189889 images (224x224x3 px) grouped in 3 classes (Corrugation, Fatigue block, Stripping off block) | YOLOv3 + MobileNetV2 | mAP: 87.40% |
| [46] | Proprietary | COTS VTIS | SS | $D_1$ | 138 images (512x512 px, 4000 after augmentation) grouped into 2 classes (Normal, Defective) | U-Net | DSC: 0.99 |
| | | | C | $D_2$ | Rails extracted from the $D_1$ images | DenseNet | Avg ACC: 90.34% |
| [47] | Proprietary | CCD industrial cameras with high-resolution and high-speed line scanning (installed perpendicularly to the rails on IV) | SS | - | 127 images (1250x55x1 px) grouped into 5 classes (Crack, Regular circle, Irregularly defect, Small defect, Blurred defect) | SegNet | DR: 100% |
| [48] | Proprietary | Handhold DSLR camera (perpendicular to the fasteners, the distance from the fasteners is constant) | OD | $D_1$ | 322 track images (3456x5472 px), 1059 (416x416 px) after augmentation and resizing, grouped into 5 classes (Normal rail, Rail corrugation, Complete fastener, Broken fastener, Missing fastener) | YOLOv3 + DenseNet | mAP: 99.20% |
| | | | OD | $D_2$ | An extension of $D_1$ containing 3333 track images (after augmentation) grouped into 9 classes (Normal rail, Rail corrugation, Rail spalling, Complete GJ fastener, Broken GJ fastener, Missing GJ fastener, Complete CK-1 fastener, Missing CK-1fastener, Broken CK-2 fastener) | YOLOv3 + DenseNet | mAP: 99.61% |
| [49] | Proprietary | Cameras installed under the train (above the rails) | OD | - | 195 rail images (128 surface smooth images and 67 surface rough images) grouped in 2 classes (Normal, Defective) | YOLOv3 (ResNet-101) | DR: >97% |
| [50] | Proprietary | Not specified by the authors | OD | - | Images (not specified) grouped in 2 classes (Knot - K), Hole - H) | Faster R-CNN | $ACC_K$: ∼98%<br>$ACC_H$: ∼95% |
| [52] | Proprietary | Images captured from a fast lane and from ordinary and heavy haulage tracks | OD | - | Images (not specified) containing at least one defect each | Faster R-CNN | AP: 97.8% |
| [53] | Type-I and Type-II [73], [74] (Line-level labels available on GitHub[5]) | As described in [73], [74] | C | $D_1$ | 29084 pixel lines extracted from the 67 images in the Type-I dataset grouped into two classes (Defect line, Defect-free line). Performances were evaluated at line-level (LL) and defect-level (DL) | CNN + LSTM (custom) | $F1_{LL}$: 81.45%<br>$F1_{DL}$: 88.45% |
| | | | C | $D_2$ | 160000 pixel lines extracted from the 128 images in the Type-II dataset grouped in two classes (Defect line, Defect-free line). Performances were evaluated at line-level (LL) and defect-level (DL) | CNN + LSTM (custom) | $F1_{LL}$: 79.27%<br>$F1_{DL}$: 93.00% |
| [54] | Proprietary | AE-based monitoring system [75]–[77] consisting of 4 encapsulated piezoelectric (PZT) sensors | C | - | 5536 8-second audios grouped into 4 classes (Stage I, Stage II, Stage III, Stage IV) indicating the severity of the defect from intact to significantly cracked | CNN (custom) | avg F1: 97.5% |
| [56] | Proprietary | High-speed cameras located in front of special locomotives | C | - | 8069 rail image grouped in 2 classes (Defective, Intact) | Inception-V3 | F1: 92.30% |
| [57] | Proprietary | LS cameras at the bottom of the rail IV | C | - | 16179 images grouped into 4 classes (Normal, Abrasion, Scar, Crack, Corrugation) | ResNet-18 | ACC: 96.55% |
| | | | OD | - | 38000 images (4096x4096 px, resize to 600x600 px) containing rails | FPN (ResNet-50) | Pr@0.95: 85.67% |
| [58] | Proprietary | LS cameras mounted on the comprehensive IV | OD | - | 189988 images (resized to 224x224x3 px) grouped into 3 classes (Corrugation, Fatigue block, Stripping off block) | YOLOv3 (MobileNetV2) | mAP: 87.40% |
| [59] | Proprietary | A manually pulled cart equipped with three area scan cameras. | C | - | 11830 images (1024x512 px after cropping) grouped into 2 classes (Normal, Defective) | VAE + RLSGAN [78] | Not reported |
| [63]† | Proprietary | Tensile testing machine (Zwick Z100) to stress rail specimen + Vallen AMSY-6 ASIP-2/A acquisition system ( 5 MHz) | C | - | 1940 400-$\mu s$ audios (augmented to 9440) grouped into 2 classes (Safe state, Unsafe state) | CNN (custom) | Not reported |
| [64] | Available[6] | Robots with ultrasonic sensors (for cracks, corrugation, and squat defects + pi camera (for cracks and rust) | C | $D_1$ | 395 images grouped into 2 classes (Rail surface crack, Concrete surface without crack) | CNN (custom) | ACC: 95.04% |
| | | | C | $D_2$ | 200 images grouped in 2 classes (Rusted iron, Rust free) | CNN (custom) | ACC: 82.76% |
| [65] | Proprietary | GoPro Hero 7 Black mounted on the developed maintenance vehicle | OD | - | 19971 objects extracted from the original images grouped into 11 classes (Normal e_clip on wood crosstie, e_clip on concrete crosstie (normal, defective), e_clip covered, Normal spike, Normal fishplate, Normal slide-bed plate, Normal guardrail plate) | YOLOv3 | mAP: 84.08% |
| [69] | Proprietary | Flat car equipped with GoPro Hero 7 Black supported by 2 200W LEDs | OD | - | 45950 fasteners labelled in 1920x1080 px images and grouped into 13 classes: e_clip on wood crosstie (normal, defective), e_clip on concrete crosstie (normal, defective), e_clip covered, Normal spike, Fishplate (normal, defective), Normal slide-bed plate, Guardrail plate (normal, defective). | YOLOv3 | ACC: 83% |
| [66] | Proprietary | HD cameras (Zenmuse Z30) fixed on the GJI Matrice 600 UAV | OD | - | 517 rail track images extracted from videos (1080p) and grouped into 3 classes (Normal fastener, Fractured fastener, Missing Fastener) | FPN | mAP: 95.78% |
| [67] | Proprietary | Handhold DSLR camera (perpendicular to the fasteners, the distance from the fasteners is constant) | C | - | 450 fasteners images (128x256 px), extracted from track images (900x1800 px, or larger), grouped into 3 classes (Complete fastener, Broken fastener, Missing fastener) | VGG16 | ACC: 97.14% |
| | | | OD | - | 332 track images (3468x5472 px) containing fasteners grouped into 3 classes (Complete fastener, Broken fastener, Missing fastener) | Faster R-CNN | mAP: 97.90% |

**TABLE 2.** *(Continued.)* **Dataset details for rail track.**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| [68] | Proprietary | Novel 3D Laser High-Speed Railway Detection System | C | - | More than $10^5$ fasteners images grouped into 5 classes (Intact fastener, Missing fastener, Upside broken fastener, Downside broken fastener, Both-side broken fastener) | CNN (custom) | ACC: 100% |
| [70] | Proprietary | LS cameras at the bottom of the track inspection car | C | $D_2$ | 1325 fasteners images, 5760 after augmentation, grouped into 6 classes (Missing, Broken, Offset, Reversed, Rotated, Normal) | VGG | ACC: 98.10% |
| | | | OD | $D_1$ | 21039 images (4096x4096 px, 512x512 px after resizing) containing fasteners | CenterNet (ResNet-18) | AP: 99.97% |
| [71] | Proprietary | CCD linear array camera mounted on an IV | OD | - | $sim$15000 fasteners samples, grouped into 3 classes (Intact, Lost, Fracture) created by merging original and GAN-generated samples | GAN + ResNet-18 | Avg ACC: 98.25% |
| [72]~ | Proprietary | 4 cameras installed on the bottom of a rail car | C | - | 25390 images (101375 after augmentation) grouped into 6 classes (DO2, DO3, KB, KBOP105, JBR, ARS) | VGG | ACC: 97.2% |
| [55]~ | Proprietary | Smartphone (under dark light conditions) | OD | - | 2180 objects (extracted from 575 images) grouped in 4 classes (Tack, Fastener, Bolt, Track damage) | Faster R-CNN (ZFNet) | mAP: 85.22% |
| [21]~,† | Proprietary | Images collected from theermite welded specimens | OD | - | 450 images (4032x3024 px, ∼900 after augmentation) containing thermite welded joint | YOLOv3 | ACC: 94.2% |
| [36] | Proprietary | A TMV equipped with a special LS camera and a very bright lightning system | SS | - | 1084 images of thin cracks in sleepers (5472x3648 px, resized to 480x320 after labelling), grouped in 9 classes (Ballast, Sleeper, Rail, e-clip fastener, Crack, Branch, Stone, Broadleaf, Cavity) | SegNet | IoU: 0.22 |
| [39] | Proprietary | On-field acquisition through three acoustic emission sensors (model Micro80D and F50$\alpha$ [79], [80]) + laboratory tests (INSTRON model 1334 testing machine [81]) | C | - | 5200 1-millisecond audio signals grouped into 3 classes (Crack propagation-induced, Impact-induced, Operational noise) | AlexNet | ACC: 99.52% |
| [60] | Proprietary | Monorail image acquisition system equipped with DALSA Lenea high-speed line-scan camera. | OD | - | 1773 images (2048x560 px, 6608 after augmentation) grouped into 2 classes (Spalling (S), Cracks (C)) | CornerNet [82] | F1 (S): 91.01% F1 (C): 88.53% |
| | Combination of Type I and Type-II [73], [74] | As described in [73], [74] | OD | - | 2764 (650x188) images obtained by cropping and collecting together images from the Type-I and Type-II datasets. The aim was to identify defects within the images. | CornerNet | F1: 92% |
| | NEU RSDDS-113 [83] | As described in [83] | OD | - | 1064 images from the EU RSDDS-113 dataset | CornerNet | F1: 97.97% |
| | KolektorSDD [84] | As described in [84] | OD | - | 851 plastic surface images from the KolektorSDD dataset | CornerNet | F1: 95.87% |
| [61] | Proprietary | Zenmuse H20 aerial camera mounted on a DJI Matrice RTK 300 drone | OD | - | 600 images (3840x2160 px and 1920x1080px, resized to 1280x720 px) | ResNet | F-measure: 96.7% |
| [62] | Available[7] | DalsaSpuder3 cameras mouted below the locomotive | C | - | 1838 rails images grouped in 4 classes (Healthy, Light squat, Severe Squat, Joint) | SqueezeNet [85] + MobileNetV2 | ACC: 97.10% |

The acronyms used for the tasks refer to Classification (C), Object Detection (OD) and Semantic Segmentation (SS). In the "collection method" the following acronyms have been used: CLC - Consrast Light Compensation; MTI - Metro Tunnel Inspection; IMU - Inertial Measurement Unit; COTS - Commercial-off-the-shelf; VTIS - Visual Track Inspection System; CCD - Charge CoupledDevice; IV - Inspection Vehicle; DSLR - Digital Single-Lens Reflex; LS - Line-Scan; AE - Acoustic Emission; HD - High-Definition; UAV - UnmannedAerial Vehicle; TMV - Track Measurement Vehicle. "Model" acronyms: CNN - Convolutional Neural Network; R-CNN - Region-based CNN; LSTM - Long Short-Term Memory; GAN - Generative Adversarial Network; LSGAN - Least-Squares GAN; FPN - Feature Pyramid Network; VAE - Variational Autoencoders; LSGAN - a combination of relativistic GANs and LSGANs. "Performance" acronyms: F1 -F1-Score; RR - Recognition Rate; RMS - Root Mean Square; SSIM - Structural Similarity Index; ACC - Accuracy; ER - Error Rate; AP - AveragePrecision; mAP - mean AP; DSC - Dice Similarity Coefficient; DR - Detection Rate; Pr - Precision; IoU - Intersection over Unit; Pr@x.y - Precision computed with a x.y IoU threshold. Lastly, the ∼ symbol identifies the papers we have put under the "Promising Work" column in Table 1; while the † symbol indicates data obtained from specimens only (laboratory tests)

Features (SURF) algorithm, with the CNN showing the best performance. When a defect is detected, an alert SMS is sent to an operator, while the GPS data and the image are sent to an IoT web server to be remotely available.

References [46], [47], and [51] focus on Semantic Segmentation. The first proposes a new network to face the high false alarm rate of the traditional commercial-of-the-shelf visual-based track inspection system, implementing a multi-stage approach: in the first stage, a U-Net is used to extract rail tracks and locate the Regions of Interest (ROI); then, the ROIs are cropped ; finally, the resulting patches are fed to a CNN which classifies them into True or False alarms. The authors highlighted three limitations: (i) only one type of defect was considered; (ii) the labelling of ROIs introduced unnatural noise; (iii) the network was tested only on vertically aligned rail tracks. In [47], the authors proposed a SegNet variant to cope with images of different surface defects, collected through CCD industrial cameras on a testing vehicle arranged near the rail, through a histogram matching method. Similarly, reference [51] proposes a combined approach of saliency cues [87] of the damaged area with a U-Net Network implementing a defect detection through segmentation.

Lastly, reference [53] presents two deep learning approaches for scenarios with a reduced number of samples available. The former relies on a CNN with a Long short-term memory and an attention module to select significant information; the latter removes the attention module and leverages a TD-LSTM [88].

### B. RAILS

Reference [37] presents a systematic procedure to configure a CNN-based classifier to evaluate the rails in the quality assessment step. Images were collected by four cameras able to capture the whole rail surface. Despite the images were manually labelled into six defect classes, the paper focuses on a binary classification by grouping the samples into defective and non-defective images. RNN has been used in [38], where the inspection was performed using an ultrasonic flaw detection vehicle collecting B-scan data of ten different types of defects. The idea was to turn the problem into identification and classification of ''sequence languages''.

[5]https://github.com/Charmve/Surface-Defect-Detection/blob/master/README.md (Dataset 11 - RSDDs: Rail Surface Defect Datasets)

[6]https://github.com/kriiyer/Dataset.git

[7]https://github.com/ilhanaydintr/RailDefectDetection

### C. FASTENERS/FASTENING SYSTEMS

Some of the reviewed works address *fasteners type recognition* and *fasteners defect detection*. Reference [72] focuses on the recognition of the rail fastening systems and proposes an intelligent information processing architecture able to automatically recognize the fasteners' type between six fastening systems (e.g., DO2, KB, etc.). Images were acquired through a rail detector car equipped with four cameras, and an automated workstation to pre-process the video stream. References [65] and [69] propose architectures to evaluate the faulty or normal status of the fastening system. Images were collected through a camera installed on an inspection vehicle equipped with a GPS module to mark the location of damaged fasteners. Reference [66] relies on images captured by an Unmanned Aerial Vehicle (UAV) flying at 30m on one side of the railway line. Authors tried different architectures for the defect detection, with a YOLOv3-based detector resulting to be the fastest with a detection time of 0.01s. Reference [67] compares three different approaches: the first is based on Dense Scale Invariant Feature Transform [89], bag-of-visual-word model [90], Spatial Pyramid Decomposition [91], and SVM ; the second is based on VGG16 convolutional layers; the last one is based on Faster R-CNN. The latter configuration results in better performances and in a faster comprehensive speed (0.23s for both detection and classification). Reference [48] focuses on both rails' heads and fasteners defect detection by relying on an architecture based on YOLOv3 with scale reduction and further feature connections between layers to achieve lower complexity and faster detection. In [68], the authors developed a 3D Laser Railway Detection System for automated railway fasteners detection on ballastless tracks. The system collects "depth images" to be used as input to a CNN-based classifier.

### D. WELDED JOINTS AND SLEEPERS

Two papers deal with *thermite welded joints* detection and *concrete and sleeper cracks* width estimation. Reference [21] proposes a YOLOv3-based object detector for the welded joints detection task, trained and tested on images collected in laboratory thermite welded specimen. Reference [36] proposes three Semantic Segmentation models, based on the SegNet model, to predict the width of cracks in concrete and sleepers. The authors made use of rectangular convolutional kernels to make the model more compliant with the shape of the cracks.

### E. FINDINGS ON RAIL TRACK

In most cases, researchers presented a single-target solution, i.e., focused on a specific component or a class of components (e.g., rails' heads), proposing new architectures or carrying out interesting comparative studies (e.g., [66], [67]). Nevertheless, we also found a comparative study presenting an interesting multi-target architecture able to cope with both defective fasteners and rails' heads defects [48].

Regarding rails' heads surface defect detection, the analysed papers mostly implement custom architectures from scratch. Instead, in Fasteners Defect Inspection most of the studies presented approaches based on SOTA object detectors (e.g., YOLOv3 and Faster R-CNN). Moreover:

- in two cases, the authors "simply" focused on items classification or detection, without looking at defects. In particular, [72] focused on fasteners type classification, while, [21] focused on welded joints detection;
- in two cases, tests were performed only on data coming from laboratory experiments (cf. [21], [63]);
- some studies use geolocation to detect the location of faulty components on the rail tracks (e.g., [44], [69], [65]).

A further finding regards the implementation of approaches based on Audio Processing for maintenance tasks: besides the results obtained by [63] on audio events generated in the laboratory, the authors in [35] leveraged audio data related to the wheel-rail interaction, while those in [54] leveraged Acoustic Emission data collected on the field pre-training their model on data coming from the well-known AudioSet dataset [34], [92].

### F. DATASETS

In most cases, data have been collected using custom systems, hand-labelled, and not made available. Some exceptions leverage third parties data, such as [46] using ImageNet [93] or [44], [51] and [53] using the Type-I [73] or the Type-II [74] datasets. Reference [36] leveraged the dataset proposed by [94], despite the fact that this includes cracks related to generic concrete structures. Lastly, in [54], the authors leveraged the AudioSet dataset [92]. Only in one case [64] the authors presented a new dataset, available on Github,[8] for rails' cracks and rust.

## V. PANTOGRAPH & CATENARY

In the following, we refer to three main components:

- *Catenary Support Device (CSD)*, supporting wires above the rail and consisting of many sub- components (e.g., insulators, brace sleeves, steady arm, etc.);
- *Pantograph and Arcs*: including pantograph defect inspection and detection of arcs, i.e., abnormal sparkling events occurring between the pantograph slide plate and the contact wire;
- *Catenary Wires*: including all papers dealing with wires (e.g., droppers, contact wire, message wire, etc.) and related defect detection.

Tables 3 and 4, whose structure and aim are described in Section III-E, summarise the papers we found within this Railway Area.

### A. CATENARY SUPPORT DEVICE (CSD)

Different DL-based solutions perform defect inspection on CSD components. In [134], the authors applied a DL-based

[8]https://github.com/kriiyer/Dataset.git

**TABLE 3. Maintenance tasks for pantograph & catenary, by components.**

| *Components* | *Surface Defect Detection* | *Defect Inspection* | *Promising Works* |
|---|---|---|---|
| Catenary Support Device | [95], [96], [97] | [98], [99], [100], [101], [102], [103], [104], [105], [106], [107], [108], [109], [110] | [111], [112], [113], [114], [115], [116] |
| Pantograph & Arcs | [117], [118] | [101], [119], [120], [121] | [122], [123], [124] |
| Catenary Wires | | [125], [126], [127], [128], [109] | [129] |

approach to detect the insulators before performing the defect detection through Contour Features and Gray similarity matching. Reference [114] uses both a classical approach (based on SURF) and an AlexNet-based method to detect insulators and bolts in images of CSD in metro power supply systems. Reference [113] proposes a method based on the Markov Random Field (MRF) and on Faster R-CNN models to perform loose strands diagnosis in isoelectric lines. Differently, a three-stage DL-based model, combining SSD, YOLO, and CNN operating at >60 FPS, is proposed in [99] to detect fastener defects on the catenary support device. Reference [95] proposes a two-stage method to detect surface defects in catenary insulators leveraging a Faster R-CNN (based on VGG16), a deep denoising autoencoder (DDAE), and a deep material classifier (DMC, a segmentation approach inspired to [135] and [136]) trained through a multi-task learning approach, sharing the first convolutional layers. A two-stage approach is also described in [98] to perform defect detection in contact wire clamps and split pins by using a custom catenary components segmentation network (CCSN) inspired to the Mask R-CNN architecture. With the same aim, reference [100] presents a three-stage method adopting an improved YOLOv3 network and DeepLabV3+, achieving the best performance with respect to the state of the art. Reference [103] proposes a new two-stage approach to detect defects of catenary insulators. Composed of a Basic Localization Network and of an Oriented Region Proposal Network, it uses a post-refinement method based on Generative Adversarial Network (GAN) to detect defects within the images, associating a defect score to each faulty image. Tests showed that, despite being effective, it operates at <0.3 FPS. Reference [101] proposes a solution consisting of a Spatial Transformer Network [137], to learn invariant representations of the original image, and two parallel networks to extract features on which classification is performed. Different couples were tested, with VGG16+VGG19 obtaining the best accuracy. Reference [102] presents a pyramid fusion architecture focused on bird preventing, relying on ResNet-101 as backbone network.

Some works also address CSD component detection. In [112], the authors proposed a comparative study between four frameworks to perform catenary support components detection. Despite the b est results were obtained by an R-CNN with ResNet-101, none of the tested architectures achieved good precision on small components, underlining the difficulties of such frameworks in extracting the features of small-scale targets. Reference [111] presents an approach based on the Faster R-CNN to detect brace sleeve screws. Faster R-CNN has been improved by discrimination feature maps and Proposal feature maps. Then, VGG16 and ResNet-101 have been tested as feature extractors, with the former resulting slightly better both in terms of mAP and AP, even if a little bit slower than the latter. Finally, in [115], the authors proposed to detect rod insulators within images captured through two drones. Analyses were conducted considering two types of drones, custom and off-the-shelf, evaluating different pre-processing strategies to obtain optimal images.

### B. PANTOGRAPH & ARCS

The constant contact between the pantograph sliding plate and the catenary contact wire can cause erosion on one or both the components, resulting in a sparkling event called Arc. Reference [119] proposes an Arc detection system based on CNNs to analyse frames captured by a camera located on the top of the train pointing to the catenary-pantograph contact point. In reference [120], the authors proposed a Faster R-CNN based architecture able to detect both pantograph and arcs within high-speed train lines. The proposed system involved the computing of the height of the pantograph (to detect a bow drop) and a SIFT-based approach to geolocate the fault on the railway line. The authors of reference [122] proposed a detection method for systems with double contact lines. The authors also introduce a new dataset, called PAC-TPL2020, collected using a high-resolution system installed on the roof of the train. The authors of reference [121] proposed a two-stage object detection to detect pantograph horns as well as faults such as pantograph loss and deviation. Lastly, reference [117] proposes an approach to estimate the wear of pantograph sliding plate on metro line images collected with a handhold DSLR camera. They also presented a procedure to estimate the thickness of the pantograph sliding plate.

### C. CATENARY WIRES

With respect to Catenary Wires, the *droppers* have caught the attention of researchers as their failure (e.g., bending, break, deformations, etc.) can lead to unpleasant consequences such as the burnout of the contact wire [125]. Reference [125] proposes a complex system for droppers defect inspection subdivided into (evident) faults and (tiny) defects. If a dropper is identified as faulty, a first system produces that output, otherwise, it is analysed by a second system to identify tiny defects (e.g., little foreign objects, broken strands). The same task is addressed in [126], with the authors proposing a two-stage detection system based on a light version of YOLOv3, a Faster R-CNN based network, and on traditional image processing (OTSU, digital morphological operation and self-define linear difference detection) to detect non-stress defects. Differently, in reference [127], the authors mostly focused on droppers detection as they can be easily obscured by other catenary components. With the same aim,

**TABLE 4. Datasets details for pantograph & catenary.**

| Paper | Availability | Collection Method | Task | ID | Dataset | Model | Performance |
|---|---|---|---|---|---|---|---|
| [95] | Proprietary | 2 groups of HR cameras and auxiliary light devices MotR of the KCIS-01 IV | C | - | 500000 patches (16x16, 32x32, and 48x48 px) from 1000 insulators images (see OD task) | DMC | $AUC_{16}$: 0.968<br>$AUC_{32}$: 0.998<br>$AUC_{48}$: 0.999 |
| | | | C | - | 200000+ patches (16x16, 32x32, and 48x48 px) for foreground (insulators) and background from 1000 insulators images (see OD task) | DDAE | $AUC_{16}$: 0.989<br>$AUC_{32}$: 0.895<br>$AUC_{48}$: 0.977 |
| | | | OD | - | ∼18000 catenary images (4920x3280 px) containing 6 types of key components (Insulator (I), Swivel clevis (SC), Clamp (C), Messenger wire holder (MWH), Steady arm support (SAM), Clevis end holder (CEH)) | Faster R-CNN | $mAP_I$: 99.1%<br>$mAP_{SC}$: 99.5%<br>$mAP_C$: 99.1%<br>$mAP_{MWH}$: 98.9%<br>$mAP_{SAM}$: 99.5%<br>$mAP_{CEH}$: 99.4% |
| [96] | Proprietary | High-precision Catenary Checking Monitor Device consisting of multiple cameras MotR of the patrolling IV | OD | $D_1$ | 5000 images (5120x5120 px) grouped into 2 classes (horizontal insulators (hi), oblique insulators (oi)) | Faster R-CNN (ResNet-50) | $mAP_{hi}$: 96.4%<br>$mAP_{oi}$: 95.7% |
| | | | OD | $D_2$ | 654 insulators rotated horizontally from $D_1$, grouped nto 2 classes (Normal, Defective) | Faster R-CNN (ResNet-50) | F1: 87% |
| | | | OD | $D_3$ | As $D_2$, but the insulators were not rotated | Faster R-CNN (ResNet-50) | F1: 77% |
| [98] | Proprietary | XLN4C-02 Catenary IV equipped with 2 groups of 9 HQ cameras | SS | $D_1$ | 2000 images grouped into 2 classes (Split Pin (SP), Contact wire clamp (CWC)) | Mask R-CNN (ResNet-50) + BNN | $mIoU_{CWC}$: 90.1%<br>$mIoU_{SP}$: 89.6% |
| | | | SS | $D_2$ | 600 images grouped into 2 classes (Clear image, Blur image) for uncertainty evaluation | Mask R-CNN (ResNet-50) + BNN | F1: 100% |
| | | | SS | $D_3$ | 585 clear images of CWC and SP extracted from $D_2$ including 3 kinds of defects (Nuts missing CWC, Nuts loosen CWC, Missing SP) | Mask R-CNN (ResNet-50)+BNN | $F1_{CWC}$: 95.5%<br>$F1_{SP}$: 98.2% |
| [99] | Proprietary | Cameras MotR of the XLN4C-01 IV (NT) | C | $D_3$ | 560 bolt images grouped into 16 classes: Normal and Missing states for Nut and ScrewB fasteners; then, Normal, Missing, and Latent Missing states for PinB, PinA, ScrewA, and Puller bolt. | CNN (custom) | mAP: 92.78 |
| | | | OD | $D_1$ | 8563 CSD images (66000x4400 px, resized to 660x440), plus 4487 images for testing (collected from another railway line) grouped into 4 classes (Diagonal tube up, Diagonal tube down, Clevis, Double tube joint) | SSD | mAP: 92.16% |
| | | | OD | $D_2$ | 3500 images (448x488 px) grouped into 6 classes (Nut, ScrewA, ScrewB, PinA, PinB, Puller bolt) | YOLO | mAP: 96.72% |
| [100] | Proprietary | 4C detection vehicle equipped with HD cameras (NT) | OD | $D_1$ | 11016 catenary images (resized to 416x416 px) for joint components (MB, T_UP, T_DOWN, clevis) detection | YOLOv3 | mAP: 95.26% |
| | | | OD | $D_2$ | 60094 split pins images for split pins detection (within the macro components images) | YOLOv3 | mAP: 99.06% |
| | | | SS | $D_3$ | 72874 split pins images labelled into 3 zones (TYPE_A_SP, TYPE_B_SP, TYPE_C_SP) | DeepLabV3+ | Pixel ACC: 91.26% |
| [102] | Proprietary | Imaging IV equipped with HD cameras (NT) | OD | - | 1000 images (6560x4384 px, resized to 1500x1000 px) grouped into 2 classes (Bird preventing loosening, Broken open pin) | ResNet-101 | mAP: 81.2% |
| | VOC2012 [130] | As described in [130] | OD | - | As described in [130] | ResNet-101 | mAP: 80.1% |
| [103] | Proprietary | Images collected by the JX-300 inspection vehicle | C | - | 100000 insulator patches (not specified) | GAN | overall F1: 94.6% |
| | | | OD | - | 2479 (6600x4400 px) catenary images including ∼7450 insulators | Faster R-CNN (ResNet-101+FPN) | AP@0.5: 99.9%<br>AP@0.8: 99.2%<br>AP@0.9: 80.4% |
| [104] | Proprietary | IV | C | - | 100000 unlabeled images of isoelectric lines + 100 isoelectric lines images grouped into 2 classes (Normal, Faulty) | GAN | Pr: 96% |
| | | | OD | - | 6742 CSD images (not specified) | Faster R-CNN (ResNet-101) | AP: 85.31% |
| [105] | Proprietary | HR cameras MotR of the IV | C | - | 100000 images between isoelectric lines and Insulators + 100 images containing Normal and Faulty isoelectric lines (IL) and insulators (In) | GAN | $F1_{IL}$: 97.8%<br>$F1_{In}$: 100% |
| | | | OD | - | 6742 images (6600x4400 px, resized to 660x440 px) containing Isoelectric lines (IL) and Insulators (In)) | Faster R-CNN | $AP_{IL}$: 85.31%<br>$AP_{In}$: 91.03% |
| [106] | Proprietary | Image acquisition system MotR of the catenary IV | SS | - | 2000 images containing insulators | Mask R-CNN | mask IoU: ∼93% |
| | | | SS | - | 8000 insulator piece images (resized to 256x64 px) | RCCAEN | Not specified |
| [107] | Proprietary | Aerial camera Zenmuse Z30 (produced by DJI-Innovations) fixed on the UAV DJI Matrice 600 | OD | - | 1411images (1920x1080 px) containing 17 objects (Nut (normal), PinA (normal, latent, missing), PinB (normal, latent, missing), PinC (normal, latent, missing), PinD (normal, latent, missing), Puller bolt (normal, missing), Screw (normal, missing)) | Faster R-CNN (ResNet-101) | mAP: 76.2% |
| | VOC2007 [131] | As described in [131] | OD | - | As described in [131] | Faster R-CNN (ResNet-101) | mAP: 80.1% |
| [108] | Proprietary | 12 cameras MotR of the IV: 6 cameras were used to collect 3D images in the front area, while the others to obtain 3D images in the back area | C | - | 50760 data points grouped into 16 classes by considering 4 components (Cantilever support connection-down and connection-up, Registration arm support connection-down and connection-up) and, for each of these, four states (normal, pin loss, screw loss, screw loose) | CNN (custom) | F1: 87.03% |
| [111]∼ | Catenary-5000 [132] | Cameras MotR of the GJ-2 IV | OD | - | 5022 images (resized to 900x600 px) grouped into 12 classes (Insulator, Rotary double ear, Binaural sleeve, Brace sleeve, Steady arm base, Bracing wire hook, Double sleeve connector, Messenger wire base, Windproof wire ring, Insulator base, Isoelectric line, Bracing sleeve screw (BSS)) | Faster R-CNN | mAP: 85.47%<br>$AP_{BSS}$: 58.15% |
| [112]∼ | Proprietary | HR cameras and LED light MotR of a IV | OD | - | About 40000 objects extracted from 6000 images grouped into 12 classes (Insulator, Rotary double ear, Binaural sleeve, Brace sleeve, Steady arm base, Bracing wire hook, Double sleeve connector, Messenger wire base, Windproof wire ring, Insulator base, Isoelectric line, Bracing sleeve screw) | Faster R-CNN (ResNet-101) | mAP: 79.7% |
| | | | | | | Faster R-CNN (VGG16) | mAP: 62.7% |
| | | | | | | SSD | mAP: 66.2% |
| | | | | | | YOLOv2 | mAP: 52.1% |

**TABLE 4.** ***(Continued.)*** **Datasets details for pantograph & catenary.**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| [113]~ | Proprietary | 12–16 HD industrial cameras MotR of the JX-300 IV (NT) | OD | - | 549 images (6600x4400 px resized to 640x440 px) containing isoelectric lines | Faster R-CNN (ZFNet) | ACC: 96.54% |
| [114] | Proprietary | 4 image acquistion equipments with different angles MotR of an operating train | OD | - | 4776 images (resized to 227x227x1 px) grouped into 8 classes (Top bolt (normal, abnormal), Side bolt (normal, abnormal), Bottom bolt (normal, abnormal), Insulator (normal, abnormal)) | AlexNet | Not specified |
| [115]~ | Proprietary | 2 drones: a Copting Transformer UAV Hexacopter equpped with a Sony Alpha 7 RII (7952x5304 px, 1-3 FPS) and a DJI Mavic Pro with an integrated camera (4096x2160 px, 24 FPS) | OD | - | 395 (containing 1062 insulators) and 752 (showing relevant scenes without assets) high-resolution images. 44000 images in total after augmentation. Tests were performed using full-resolution (FR) images and cropped (C) images (600x600 px) | SSD + MobileNetV1 | $mAP@0.5_{FR}$: 96.8% $mAP@0.5_{C}$: 98% |
| [116]~ | Proprietary | High-speed cameras MotR of the catenary IV | OD | - | 4644 images (6600x4400 px, resized to 990x660 to improve the detection efficiency) grouped into 2 classes (Defective messenger wire bases (DMW), Bracing wire hooks (BWH)) | Faster R-CNN | $AP_{MWB}$: 90.1% $AP_{BWH}$: 89.1% |
| [117] | Proprietary | Handhold DSLR camera | C | - | 238 images (both 5472x3649 px RGB and 4904x328 px grayscale, resized to 256x256 px) grouped into 5 classes (Normal, Over abrasion, Partial abrasion, Groove-shaped abrasion, Others) | CNN (custom) | ACC: 90.63% |
| [118] | Proprietary | Image acquisition equipment installed beside the track + images (with relatively obvious defects) captured by the metro maintenance department | OD | - | 257 pantograph images grouped into 4 classes (Normal pantograph, Over-abrasion, Irregular-abrasion, Burning) | Faster R-CNN | mAP: 97.63% |
| [119] | Proprietary | A camera MotR of the train | OD | - | 811 frames extracted from a video for Arcs detection (if present) | CNN (custom) | AP: 97.52% |
| [120] | Proprietary | Camera MotR of the train collecting videos under five different conditions | OD | - | 1200 frames from 10 Pantograph (P) videos and 4233 frames from 30 Arc (A) videos (320x240x3 px) | Faster R-CNN (ZFNet) | $AP_P$: 99.54% $AP_A$: 99.48% |
| [121] | Proprietary | 4 cameras: 2 MotR of the train (vehicle orientation) and 2 fixed on the trackside (orbital orientation) | OD | - | Multiple sets of 100 sequences from vehicle orientation cameras, 800 images from the orbital orientation cameras, 500+ images containing defective horns, grouped into 3 classes Normal horn orbital, Normal horn vehicle, Faulty horn) | SSD | mIoU: 90.75% |
| [101] | Proprietary | Data collected from Sidney Train Maintenance Center | OD | - | 2336 images (2048x5400 px and 37932x2730 px) from which 1546 images were extracted, labelled, resized to 224x224 px, and grouped into 10 classes (Splice (standard and V-ware), Knuckle (standard and misinstalled), Kline Insulator (standard and twisted), Dropper1 (standard and defective), Dropper2 (standard and broken) | STN + VGG16 + VGG19 | ACC: 94.09% |
| [122]~ | Proprietary (PAC-TPL2020) | HR cameras MotR of the train | OD | - | 5000+ images (1920x1080 px) grouped into 5 classes (Cantilever and bridge frame interference, Severe weather, Arcing, Foreign light interference, Other) | ResNet | ACC: 99.97% |
| [123]~ | Proprietary | 2 sources: Imager MotR of a high-speed operating train, and the Internet | OD | - | 3600 images (not specified) | YOLOv3 | ACC: 97.4% |
| [124]~ | Proprietary | Shutter cameras and flash lamp devices MotR of the service train | OD | - | 800 images (1952x1280 px) containing pantographs (not speficied) | CNN (custom) | mbIoU: 93.4% |
| | | | SS | - | As for OD | DeepLab (ResNet-50) | mIoU: 90.2% |
| [125] | Proprietary | HD camera MotR of the catenary IV | OD | - | 4800 dropper images grouped into 4 classes (Dropper (normal), C-Dropper-1 (evident fault), C-Dropper-2 (evident fault), S-Dropper (tiny fault)) | Faster R-CNN (ResNet-18) | ACC: >95% |
| [126] | Proprietary | Imaging IV equipped with HD cameras (NT) | OD | - | 5120x3840 px resolution images (number not specified) adjusted, augmented, and resized to 2500x1200 px, grouped into 3 classes (Foreign body (FB), Hard bending (HB), No-stress (NS)) | Faster R-CNN | $PR_{FB}$: 91% $PR_{HB}$: 90% $PR_{NS}$: 39% |
| [127] | Proprietary | Training images extracted from the videos captured by an IV | C | - | 1600 images (not specified) of dropper bodies grouped into 2 classes (Normal, Abnormal) | ResNet-34 | ACC: 99% |
| | | | OD | - | 4500 images (1024x1024 px) after augmentation and resizing, containing Current Carrying Rings (CCRs) grouped into 2 classes (Normal, Faulty ring) | CenterNet | AP@50: 89.9% |
| [129]~ | Proprietary | High-speed rail 2C system for engineering tests | OD | - | 1465 HR images (resized to 1333x800 px for training and 960x800 px for testing) containing dropper | Faster R-CNN (ResNet-101) | mAP@0.5: 86.8% mAP@0.7: 83.9% |
| | VOC2012 [130] | As described in [130] | OD | - | As described in [130] | Faster R-CNN (ResNet-101) | mAP@0.5: 74.9% |
| | MS-COCO 2014 [133] | As described in [133] | OD | - | The MS-COCO 2014 contains 11540 images with 80 object categories | Faster R-CNN (ResNet-101) | mAP@0.5: 50.8% |
| [128] | Proprietary | IV under various visual conditions | OD | - | 5460 images containing several Current-carry rings grouped into 2 classes (Normal, Faulty) | CenterNet | AP: 55.9% AP@0.5: 78.1% |
| [97] | Proprietary | Acquistion device of the CCLM consisting of two sets of cameras and flashers MotR of the IV | C | - | 160000 small insulator patches (32x32 px) grouped into 2 classes + 1000 insulator images for testing | GAN | AUC: 0.98 AD F1: 95% |
| | | | SS | - | 800 normal insulator images (not specified) | DPDN [124] + DeepLabV3 (ResNet-18) | mIoU: 94% |
| [109] | Proprietary | 2C System | OD | - | 2768 between small, medium, and large defective objects ugmented to 3948 and divided in 3 classes (Pole number plate malfunction, Foreign body invasion, Unstressed dropper) | Faster R-CNN | mAP: 89.61% mean F1: 93.75% |
| [110] | Proprietary | Catenary images were provided by the China Railway Group Co., Ltd. | C | - | 2447 insulator images (360x360 px, after augmentation) grouped into 3 classes (Normal (N), Damaged (D), Missing (M)) | VGG16 | F1 (N): 96.66% F1 (D): 90.66% F1 (M): 88.52% overall mAP: 93.46% |
| | | | OD | - | 1000 images (3968x2976 px), 4000 (500x375 px) after resizing and augmentation, each of which contains at least three insulators | RPN (custom) | AP: 94.23% |

The acronyms used for the tasks refer to Classification (C), Object Detection (OD) and Semantic Segmentation (SS). In the "collection method" column it has been briefly reported the acquisition setup by using the following acronyms: HR - High-Resolution; IV - Inspection Vehicle; MotR - Mounted on the Roof (of the train or IV); HQ - High-Quality; NT - Night Time; HD - High-Definition; UAV - Unmanned Aerial Vehicle; DSLR - Digital Single-Lens Reflex; CCLM - Catenary-Checking on-Line Monitor Device. "Model" acronyms: DMC - Deep Material Classifier; DDAE - Deep Denoising Autoencoder; CNN - Convolutional Neural Network; R-CNN - Region-based CNN; BNN- Bayesian Neural Network; GAN - Generative Adversarial Network; RCCAEN - Reconstruction and Classification Convolutional Autoencoder Network; STN - Spatial Transformer Network; RPN - Region Proposal Network. "Performance" acronyms: AUC - Area Under Curve; AP - Average Precision; mAP - mean AP; F1 - F1-Score; IoU - Intersection over Unit; mIoU - mean IoU; ACC - Accuracy; mAP@x.y (or AP@x.y) - mAP (or AP) computed with a x.y IoU threshold; mbIoU - mean bounding boxes IoU; PR - Precision Rate. Lastly, the ~ symbol identifies the papers we have put under the "Promising Work" column in Table 3.

the authors of reference [129] implemented an approach for high-speed railways based on the Faster R-CNN architecture. The model was tested on images from the high-speed rail 2C system, showing to be able to operate only up to ∼8 FPS.

### D. FINDINGS ON PANTOGRAPH & CATENARY

What most characterises the Pantograph & Catenary area compared to that of the Rail Track is the presence of very small components. This represents, especially with regard to CSDs, one of the greatest challenges to be faced as applying a defect detection approach directly to the entire image will result in poor accuracy. Therefore, in many cases, multi-stage approaches have been presented to identify the component under examination as a first step and then apply a defect identification mechanism, whether based on traditional image processing techniques (e.g., [134]), semantic segmentation (e.g., [98], [100]), classification networks (e.g., [99]), or other approaches such as deep denoising autoencoders (DDAE) and deep material classifier (DMC) as in [95]. The architectures mostly used as the first stage were those belonging to the YOLO family, given their detection speed, and the Faster R-CNN, given their accuracy. Clearly, these frameworks have been modified to suit the case under examination. In this respect, a mention goes to the work described in reference [103] as it proposes an innovative methodology to better identify insulators and applied GAN to identify defects.

Interestingly, while in some cases other DL or image processing approaches have been implemented downstream of these identifiers to detect anomalies, in other cases the presented approach is limited to target detection. For instance, [115] focuses on insulators detection, [111] focuses on brace sleeve screws identification, while [127] and [129] propose architectures for dropper detection. It is worth noting that ''simple'' object detection applications can also directly involve defect detection, such as the identification of Arcs (i.e., an object), which indicates possible deterioration for pantographs or contact wires (e.g., [119], [120]). Finally, to the best of our effort, we did not find any paper making use of audio data for this task.

### E. DATASETS

Researchers mostly built datasets from scratch or relied on data that are not publicly available. Only in a few cases, the authors used existing datasets or made available those created by them. In [129], the authors tested their model on two well known datasets (VOC2012 [130] and MS-COCO 2014 [133]) to compare its performances with those achieved by other SOTA architectures. Differently, in [120], the authors used the ImageNet dataset [93] and VOC2007 [131] to pre-train their architectures, so transferring knowledge from big datasets and then adapting it to their purposes. Reference [111] used the Catenary-5000 [132] to test the proposed model for the detection of brace sleeve screws, however, it is not clear whether the dataset is available. The same issue applies to the PAC-TPL2020 dataset used in [122].

**TABLE 5. Maintenance tasks for rolling stock, by components.**

| *Components* | *Surface Defect Detection* | *Defect Inspection* | *Promising Works* |
|---|---|---|---|
| Bogie & Frame | [138] | [35], [139], [140], [141], [142], [143], [144], [145], [146], [147] | [148] |
| EMU Train Key Components | | [149], [150] | |
| Others | | [151] | |

## VI. ROLLING STOCK

In this study, we found a few papers mostly related to small objects (in relation with the whole train) detection, and their defect identification, within bogies and train's body and frame. Tables 5 and 6, whose structure is defined in Section III-E, report their characteristics.

### A. BOGIE & FRAME

Reference [139] proposes a defect detection architecture to deal with low image quality and complex background, sliding a set of anchors over multi-level convolutional feature maps. The model was tested on four datasets related to 1) Cut-out cock handle, 2) Dust collector, 3) Fastening bolts, and 4) Bogie block key. The same authors improved the approach in [140] to make it suited for real-time fault detection by introducing a novel multilevel feature fusion strategy. They also proposed an image acquisition system composed of a device placed in the middle of the rails and two devices on the rails' sides composed of CCD cameras and an auxiliary light device. When a train passes through, images were acquired and sent to a server system for the analysis. A multi-defect (i.e., lost pin, lost bolt, lost rivet, broken chain, broken wire, and foreign object) detection system was proposed in [141]. The authors built a three-stage model based on ResNet-101 trained on data collected through 12 high-speed cameras placed on the tracks. Reference [144] proposes a DL-based two-stage component defect detection architecture to recognize anomalies in train bogies. In the first stage, a hierarchical object detection scheme is implemented to detect and localize small and large objects, while the second pipeline is tailored to detect smaller components. Reference [142] focuses on bolt defect detection. The authors proposed a system for bolt-loosening detection in key bogie components (Axle Box, Traction Motor and Gear Box) acquiring data in an online fashion, i.e. while trains were running. First, a CNN-based system detects regions of interest, then another CNN extracts bolts edges; lastly, to detect single or multiple bolt loosening, a 3D reconstruction method was used to calculate the distance between the bolt cap and the mounting surface. Results show optimal performances with a relative error smaller than 1.42%, moreover, the processing time to complete the fault detection considering a train with 8 cars was about 4.6 minutes, 1.1s per image (within China's requirement of 5 minutes). Lastly, in reference [143], the authors implemented a two-stage approach to determine the

**TABLE 6. Datasets details for rolling stock.**

| Paper | Availability | Collection Method | Task | ID | Dataset | Model | Performance |
|---|---|---|---|---|---|---|---|
| [35] | Proprietary | Behringer B-5 sensor + Behringer U Phoria UMC204HD | C | - | 228 30-second audios grouped in 4 classes (dry_40, dry_60, wet_40, wet_60) | FCNN (custom) | ACC: 100% |
| [139] | from [152] | As described in [152] | OD | $D_1$ | 1665 images (700x512 px) grouped into 2 classes (No-fault, Fault) | Faster R-CNN (GoogLeNet+HyperNet) | CDR: 99.18% |
| | | | OD | $D_2$ | Dust collector dataset, 1665 (700x512 px) grouped into 2 classes (No-fault, Fault) | Faster R-CNN (GoogLeNet+HyperNet) | CDR: 100% |
| | | | OD | $D_3$ | Fastening bolts dataset, 3626 images (700x512 px) grouped into 2 classes (No-fault, Fault) | Faster R-CNN (GoogLeNet+HyperNet) | CDR: 100% |
| | | | OD | $D_4$ | Bogie block key, 8337 images (700x512 px) grouped into 2 classes (No-fault, Fault) | Faster R-CNN (GoogLeNet+HyperNet) | CDR: 98.76% |
| [140] | from [152] (as [139]) and [153] | As described in [152] | OD | $D_1$ | Cut-out cock handle dataset, 1665 images (700x512 px) grouped into 2 classes (No-fault, Fault) | Faster R-CNN (GoogLeNet+HyperNet) | CDR: 96.24% |
| | | | OD | $D_2$ | Dust collector dataset, 1665 (700x512 px) grouped into 2 classes (No-fault, Fault) | Faster R-CNN (GoogLeNet+HyperNet) | CDR: 99.53% |
| | | | OD | $D_3$ | Fastening bolts dataset, 3626 images (700x512 px) grouped into 2 classes (No-fault, Fault) | Faster R-CNN (GoogLeNet+HyperNet) | CDR: 100% |
| | | | OD | $D_4$ | Bogie block key, 8337 images (700x512 px) grouped into 2 classes (No-fault, Fault) | Faster R-CNN (GoogLeNet+HyperNet) | CDR: 99.86% |
| | | Not specified by the authors | OD | $D_5$ | Angle cock dataset, 4026 images (700x512 px) grouped into 2 classes (no-fault, fault) | Faster R-CNN (GoogLeNet+HyperNet) | mean CDR: 99.13% |
| [141] | Proprietary | 12 HS cameras located on the railways (under the train) 4 for each side and 4 in the middle | OD | $D_1$ | 307 images containing at least 1 defect grouped in 6 classes (Lost pin, Lost bolt, Lost rivet Foreign object, Broken chain, Broken wire) | ResNet-101 | avg F1: 79.32% |
| | | | OD | $D_2$ | 580 images (after data augmentation) grouped in 2 classes (Scratch, Oil leak) | ResNet-101 | avg F1: 88% |
| [142] | Proprietary | Binocular stereo vision (2 GC1380H cameras and 2 Cingeo1.4 lenses) | OD | - | 1908 images grouped into 6 classes (Non-fault, Bolt-loosening, and Bolt-missing of Axle Box, Traction Motor, and Gear Boxcomponents) | CNN (custom) | RDRR: 100% |
| | BSD500 [154] | As described in [154] | SS | - | As described in [154] | CNN (custom) | AbsE: <0.08mm RE: <1.23% |
| [143] | Proprietary | HPLAC installed in a MIS | C | - | 1066 Lidbolt images, grouped into 2 classes (Normal, Abnormal) | OC-CNN (VGG16) | AUC: 1.0 |
| | | | C | - | 1325 Groundbolt images, grouped into 2 classes (Normal, Abnormal) | OC-CNN (VGG16) | AUC: 0.9648 |
| | | | C | - | 286 Mgorundbolt images, grouped into 2 classes (Normal, Abnormal) | OC-CNN (VGG16) | AUC: 0.9927 |
| | | | OD | - | 261 axle box cover images containing 3 types of bolts (Lidbolt, Groundbolt, Mgroundbolt) | Faster R-CNN | ACC: 100% |
| [144] | Proprietary | TFDS | C | - | 1200 images grouped into 4 classes (Normal bearing; Defective bearing (db), Normal spring, defective sprin1g (ds)). | Detection Model + DenseNet | $Pr_{db}$: 87.5% $Re_{db}$: 92.5% $Pr_{ds}$: 88.3% $Re_{ds}$: 93.3% |
| | | | OD | - | 1156 images (1400x1024 px) containing 12207 train components grouped in 15 classes (b-plate, l-plate, bearing, collector, flange, spring, group, fixator, valve, nut-s, screw-s, nut-f, screw-f, bolt, plug) | Faster R-CNN (ResNet-101) | Detection mAP: 96.3% |
| | | | OD | - | 1980 images grouped into 4 classes (normal b-plate, defective b-plate with screw missing - ms, normal collector, defective collector with bolt missing - mb) | Faster R-CNN (ResNet-101) | $Pr_{ms}$: 89.5% $Re_{ms}$: 100% $Pr_{mb}$: 91.5% $Re_{mb}$: 99.1% |
| | | | SS | - | 1200 images grouped into 4 classes (normal l-plate, defective l-plate - dl, normal valve, defective valve - dv) | Detection Model + U-Net | $Pr_{dl}$: 88.8% $Re_{dl}$: 99.1% $Pr_{dv}$: 92.9% $Re_{dv}$: 98.3% |
| [151] | Proprietary | COTS VTIS | SS | - | 74 images (1024x1400 px resized to 256x256 px) grouped in to 4 classes (Axle bolts - A, Break - B, Gasket - G, Screws - S) | U-Net | $IoU_A$: 0.8336 $IoU_B$: 0.7991 $IoU_G$: 0.9066 $IoU_S$: 0.8757 |
| | | | SS | - | 73 images (1024x1400 px resized to 256x256 px) grouped into two classes (Normal, Faulty) | U-Net | IoU: 0.8417 FVD ACC: 97.26% |
| [149], [150] | Proprietary | EMU Failures Detection System: matrix cameras on tracks and linear cameras on both sides of the rails (night time) | C | - | 7000 bolt images grouped into 6 classes (B1, B2, M1, M2, L1, L2) from 700 bolt plates | Inception-ResNet-v2 | mAP@0.5: 79.2% |
| | | | OD | - | 3000 images containing 15000 objects grouped into 6 classes (Brake disc, Brake calliper, Tractor, Side suspension, Under suspension, Plate bolt) and dimensions (small (S), medium (M), large (L)) | Faster R-CNN (ResNet-101) | $ACC_S$: 93.8% $ACC_M$: 97.6% $ACC_L$: 98.4% |
| | Available[9] | As described on the GIT page | C | - | As described on the GIT page | Inception-ResNet-v2 | mAP@0.5: 76.5% |
| [148]~ | Proprietary | TFDS | OD | - | 3884 (1400x1024 px) non-fault images gropued into 4 classes (Sleeper spring - SS), Side Frame Keys - SKF), Shaft Bolts - SB), End Bolts - EB) | CNN (custom) | $Re_{SS}$: 100% $Re_{SKF}$: 98.95% $Re_{SB}$: 100% $Re_{EB}$: 98.95% |
| [146] | From [139], [140] | As described in [139], [140] | OD | $D_1$ | Cock handle dataset, 1665 images grouped into 2 classes (No-fault, Fault) | SqueezeNet | mCDR: 98.60% |
| | | | OD | $D_2$ | Dust collector dataset, 1665 images grouped into 2 classes (No-fault, Fault) | SqueezeNet | |
| | | | OD | $D_3$ | Fastening bolts dataset, 3626 images grouped into 2 classes (No-fault, Fault) | SqueezeNet | |
| | | | OD | $D_4$ | Bogie block key, 8337 images grouped into 2 classes (No-fault, Fault) | SqueezeNet | |
| | | | OD | $D_5$ | Angle cock dataset, 4026 images grouped into 2 classes (no-fault, fault) | SqueezeNet | |
| | | | OD | $D_6$ | Brake show key, 9600 images grouped into 2 classes (no-fault, fault) | SqueezeNet | |
| | ImageNet [93] | As described in [93] | OD | - | As described in [93] | SqueezeNet | Top1 ACC: 64.4% |
| | Pascal VOC 2007 [131] | As described in [131] | OD | - | As described in [131] | SqueezeNet | mAP: 70.01% |
| | MS-COCO [133] | As described in [133] | OD | - | As described in [133] | SqueezeNet | AP@0.5: 19.7% |
| [145] | Proprietary | TFDS | OD | - | 4518 (1400x1024 px) images grouped into 6 classes (Bottom bolt (defective, normal), Side bolt (defective, normal), Retaining key (defective, normal) | SSD+VGG16 | mAP: 94.19% |

**TABLE 6.** *(Continued.)* **Datasets details for rolling stock.**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| [138] | Proprietary | Industrial cameras | OD | - | About 1200 (416x416 px) images after cropping and augmentation operations grouped in 4 classes (Steel Crack, Dent, Inclusion, Scratch) | YOLOv3 | mAP: 88.3% |
| [147] | Proprietary | Isaw sports action camera (240FPS) | OD | - | 7 datasets containing frames (480x848x3 px) extracted from the recorded videos. D1 (14688 frames), D2 (233280 frames), D3 (202080 frames), D4 (185280 frames),D5 (195600 frames), D6 (108000 frames), D7 (233280 frames). D1-6 are defect free, and the frames are grouped in 6 classes (Axial box, Spring, Binding screw, Support beam, Flange, Support rod). D7 contains defective samples e.g. defective spring, defective binding screw. | YOLOv2 | mAP: 84.05% |

The acronyms used for the tasks refer to Classification (C), Object Detection (OD) and Semantic Segmentation (SS). In the "collection method" column it has been briefly reported the acquisition setup by using the following acronyms: HS - High Speed; HPLAC - High-Precision Linear Array Camera; MIS - Metro Inspection Shed; TFDS - Running Freight Train Detection System; COTS - Commercial-off-the-shelf; VTIS - Visual Track Inspection System; EMU - Electric Multiple Units. "Model" acronyms: FCNN - Full Connected Neural Network; CNN - Convolutional Neural Network; R-CNN - Region-based CNN; OC-CNN - One-Class CNN. "Performance" acronyms: ACC - Accuracy; CDR - Correct Detection Rate; avg - average; F1 - F1-Score; RDRR - Region Detection Recall Rate; AbsE - Absolute Error; RE - Relative Error; AUC - Area Under Curve: Pr - Precision; Re - Recall; IoU - Intersection over Unit; FVD - Fault Valve Detection; AP - Average Precision; mAP - mean AP; mAP@x.y (or AP@x.y) - mAP (or AP) computed with a x.y IoU threshold. Lastly, the $\sim$ symbol identifies the papers we have put under the "Promising Work" column in Table 5.

status (Normal/Abnormal) of Axle Box Cover bolts. Images were collected through an acquisition system called ''metro inspection shed'' equipped with different line-scan digital cameras with high precision. In the first stage, a Faster R-CNN architecture was implemented in order to identify bolts within the component under examination. A one-class convolutional architecture (OC-CNN) was used to discern normal samples against defective ones, to deal with the lack of negative examples. Results show that the whole architecture requires 265ms to process a single image.

### B. EMU TRAIN KEY COMPONENTS

With the term EMU Train key components, the authors of [149] and [150] referred to brake disk, brake caliper, tractor, side suspension, under suspension, and plate bolt. The authors proposed a system able to cope with EMU components defect detection inspired by the Faster R-CNN architecture with a feature extractor based on ResNet-101 (instead of VGG16). Extracted regions were cropped and processed in order to obtain ''super-resolution images'' leveraging, among the others, a Generative Adversarial Network (SRGAN) [155].

### C. OTHER COMPONENTS

Reference [151] proposes a faulty rail-valve detection through a two-step segmentation approach leveraging U-Net. Images were collected through a VTIS system which allowed an image collection always from the same distance. The first network performed the segmentation on the whole image, detecting the segmented mask which highlights the valve. These were then processed by the second network that performed a high-resolution segmentation. The mask in output was then processed by an image processing algorithm to detect the fault. Despite the good performances, the system is able to work only with train-valve binary classification and only if there is a single valve per image oriented in a specific way.

### D. FINDINGS ON ROLLING STOCK

From the above analysis, it is clear that the bogies have been the most investigated components. As for the Pantograph & Catenary area, the bogies are composed of multiple small elements that are very hard to identify given the complex background. Therefore, in most cases, multi-stage approaches based on Object Detection frameworks (mainly Faster R-CNN) were proposed to cope with Bogie and EMU train components defect inspection. Regarding Audio Processing, only the authors in [35] leveraged audio data to estimate the adhesion conditions between wheels and rails. Lastly, transfer learning approaches have been applied, as mentioned in the following subsection.

### E. DATASETS

Regarding the used datasets, references [139] and [140] leverage those introduced in [152] to train and test their models. In [139], the authors also used the ImageNet dataset [93] to pre-train the model implementing, de facto, a transfer learning approach. Likely, in [142], the authors relied on the BSD500 dataset [154] to pre-train their bolt edge detection model. In two cases, the models proposed for bogie or EMU train components defect detection were also tested on other kinds of data to evaluate their generalization performances. Indeed, in [141], the model, trained to identify defects such as lost pin, lost rivet, broken chain, etc. was also tested fine-tuning it to cope with defects such as oil leaks and scratches; nevertheless, it is not clear if these datasets are available to the researchers' community. Differently, reference [149] proposes a model to address EMU Train Key Components defect detection, which was also tested on the Northeastern University (NEU)[10] [156] surface defect dataset.

[9]Northeastern University (NEU) dataset: http://faculty.neu.edu.cn/yunhyan/NEU_surface_defect_database.html

[10]http://faculty.neu.edu.cn/yunhyan/NEU_surface_defect_database.html

**TABLE 7. Maintenance tasks for tunnel & bridge, by components.**

| *Components* | *Surface Defect Detection* | *Defect Inspection* | *Promising Works* |
|---|---|---|---|
| Tunnels' lining | [36]* [157] [158] [159] [160] [161] [162] | | |
| Bridges | [163] | | |

The * mark indicates data not related to the railway domain.

Lastly, reference [151] mentioned the Singapore Mass Rapid Transit (SMRT) dataset, however, also in this case, the dataset seems not to be accessible.

## VII. TUNNEL & BRIDGE

Focusing on the Tunnel & Bridge area, a few papers have been found. Tables 7 and 8, whose structure and aim have been described in Section III-E, report their characteristics. Finally, it is worth noting that we reported in this section also a paper (i.e., [36]) already analysed in section IV, as it addresses concrete crack analyses by using a setting compatible with the estimation of cracks in tunnels' lining.

### A. TUNNEL

As from Table 7, all but one work target tunnels' lining defects, performing such analysis at different levels. In [157], the authors used a Kinect to estimate cracks' defect intensity, leveraging a DL-based approach (a CNN classifier) to classify the faulting lines in vertical, horizontal, upward tilt or downward slope. Reference [161] uses Mask R-CNN to perform tunnel inspection and crack detection. Data were collected by a smart vision measurement system composed of different cameras, each capturing a segment of the tunnel. Segments were arranged in a circular array to obtain the tunnel cross-section. The authors considered both 2D and 3D information: the former were used to train a Mask R-CNN, with ResNet-50 as backbone, in order to obtain crack segmentation and detection; the latter were used to build a 3D model of the tunnel using a self-developed robust B-spline algorithm.

Reference [158] proposes a two-stage Semantic Segmentation approach using two Fully Connected architectures, based on VGG16, to detect (possibly overlapping) cracks and leakages. Images for training and testing were acquired by a Moving Tunnel Inspection (MTI-200a) system, developed by the authors, which captured the tunnel's surface images through 6 cameras while moving on the rails. Similarly, reference [159] uses a two-stages approach (inspired by the R-FCN architecture) to detect leakages, cracks, and scratches. Images were collected through the Movable Tunnel Inspection (MTI-100) system and two datasets were built. The first dataset was used to pre-train the Fully Convolutional Network (based on GoogLeNet and VGG16) composing the first stage of the model. This network has been considered as a traditional classifier (with a last fully connected layer) and has been trained to classify images according to five classes: leakage, crack, segment joint, pipeline and lining. Once the proposed network was trained, the last FC layers have been dropped and replaced to obtain the last feature map used as input for the next stage performing object classification and localization. Tests showed that, although the detection accuracy was almost the same, the location accuracy increased with smaller images.

Finally, reference [160] proposes a Semantic Segmentation based approach to detect cracks in tunnels' lining referring to the DeepLab-v3 architecture. The authors built two datasets starting from the images captured by a proprietary image acquisition system equipped with eight industrial linear array CCD cameras: the first contains only cracks (tunnel crack dataset); the second contains cracks, structural seams, water stains, and scratches (augmented tunnel crack dataset).

### B. BRIDGE

In [163], the authors used a YOLOv3 network, pre-trained on the VOC2007 dataset [131], to detect cracks on bridges' beams. Since YOLO is not so good with small-scale objects, images acquired by a CCD camera were first cropped in 64 smaller images. To improve generality, images' brightness and contrast were modified to simulate different lighting and weather conditions. It is worth highlighting that, despite the authors indicated that the accuracy was greater than 95%, the recall rate was less than 10%.

### C. FINDINGS ON TUNNEL & BRIDGE

In the Tunnel & Bridge area, tunnels are subject to some complex situations such as uneven illumination, image noise, etc. In order to address those problems, most of the reviewed papers leveraged existing state-of-the-art architectures. In almost all cases, the Surface Defect Detection task has been addressed as an Object Detection or Semantic Segmentation problem. Beyond reference [36], which also focused on width estimation, and [157], which implemented Image Processing approaches, the reviewed papers have not implemented an ''estimation'' mechanism of the cracks or defect intensity through DL, instead, they mostly focused on detecting, classifying and locating these defects in the concrete structures. Most of them focused on cracks, except reference [159], where the authors also included leakages and scratches, and reference [158] where the authors proposed an architecture composed of two FCNs to detect cracks and leakages.

### D. DATASETS

Concerning datasets, as for the previous areas, the authors used some existing datasets to build transfer learning approaches: reference [163] used the VOC2007 dataset [131] to pre-train the proposed architecture, while in reference [36], the model has been trained on the [94] dataset, although in our understanding this is related to generic concrete structures cracks. On the other hand, [158] built a dataset from scratch related to defects on tunnels' lining from a proprietary acquisition system.

**TABLE 8. Dataset details for tunnel & bridge.**

| Paper | Availability | Collection Method | Task | ID | Dataset details | Model | Performance |
|---|---|---|---|---|---|---|---|
| [36]* | from [94] | Smartphone with a constant distance of ∼0.2m from the concrete | SS | $D_1$ | 2750 images of thick cracks in concrete (1008x740 px, resized to 480x360 px) | SegNet | ACC: 97% IoU: 0.81 |
| [157] | Proprietary | Kinect sensor | C | - | 440 images, grouped in 4 classes (Transverse, Longitudinal, Upward tilt, Downward slope) | CNN (custom) | ACC: 91.7% |
| [158] | On request[a] | MTI-200a with 6xLSC (GigE) and 19xLEDs | SS | - | 1380 defective images (3000x24576 px, cropped to 3000x3000 px), with perimeter annotations for 6 object types (Leakage-only, Crack-only, Crack-TDN, Leakage-TDN, Crack-TDO, Leakage-TDO) | VGG16 | AvgE: 0.8% |
| [159] | Proprietary | MTI-100 with 6xHR Linear CCD and 12xLED | C | - | 9520 images (256x256 px), grouped in 5 classes (Leakage, Crack, Segment joint, Pipeline, Lining) | CNN (GoogLeNet+VGG16) | ACC: 95.84% |
| | | | OD | - | 4193 images (3000x3720 px), with box annotations for three object types (Crack, Leackage, Scratch) | R-FCN (GoogLeNet+VGG16) | ACC: 86.6% |
| [160] | Proprietary | 8xSILA CCD and IGL | SS | $D_1$ | 15718 tunnel's crack images (512x512 px) extracted from 643 crack images (4096x4096 px) | DeepLabV3 (ResNet-18) | IoU: 0.6836 |
| | | | SS | $D_2$ | augmented tunnel crack images, with other images added to include other defects | DeepLabV3 (ResNet-18) | IoU: 0.4111 |
| [163] | Proprietary | 4 CCD | OD | - | 3000 fracture samples extracted from 100 fracture images (2456x2058 px) | YOLOv3 | ACC: ∼95% |
| [161] | Proprietary | MV with 10xDCA and LEDs + LIDAR | SS | - | 30 images containing only cracks | Mask-RNN (ResNet-50) | ACC: 85.79% |
| [162] | Proprietary | Novel MTI with 8xHR LS | OD | - | Hundreds of tunnel surface images (20000x20000 px cropped to 2000x2000 px), with box annotations for three object types (Leakage, Falling block, Crack) | Faster R-CNN | mAP: 88.5% |

The acronyms used for the tasks refer to Classification (C), Object Detection (OD) and Semantic Segmentation (SS). In the "collection method" column it has been briefly reported the acquisition setup by using the following acronyms: MTI - Metro Tunnel Inspection (a type of inspection system for data collection); LSC - Linear Sensor Camera; HR - High-Resolution; CCD - Charge Coupled Device (a type of camera device); SILA - Synchronised Industrial Linear Array; IGL - Industrial grade lens; MV - Moving Vehicle; DCA - Digital Camera Array; LS - Line-Scan (a type of camera). In the "dataset details" column it has been reported the case of two (or more) overlapping and of non-overlapping defects (TDN - Two-defect-nonoverlapping; TDO - Two-defect-overlapping). "Model" acronyms: CNN - Convolutional Neural Networks; R-FCN - Region-based Fully Convolutional Networks; R-CNN - Region-based CNN. "Performance" acronyms: ACC - Accuracy; IoU - Intersection over Unit; AvgE - Average Error; mAP - mean Average Precision. Lastly, the * symbol indicates data not related to the railway domain.
[a]https://www.sciencedirect.com/science/article/abs/pii/S0886779817310258?via%3Dihub#ec-research-data

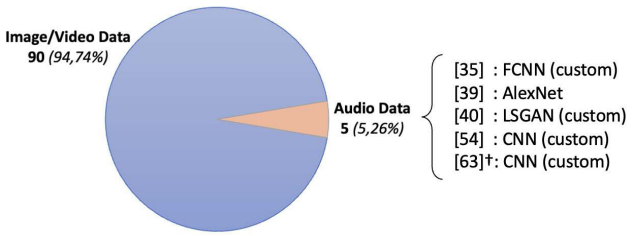



**FIGURE 4. Distribution of reviewed papers by data type (Image/Video or Audio). For audio, the figure also reports the used DL approaches.**

## VIII. STATISTICAL ANALYSIS

In this survey, we reviewed 95 papers distributed as shown in Fig. 3; interestingly, only 5 leverage audio data (Fig. 4).

Fig. 2 describes the temporal progress of papers published during the last years and reviewed in this paper, with the first work found on the subject published in 2014. The trend shows an almost exponential growth, which demonstrates an increasing interest of researchers in this field (year 2021 was not included in the graph due to partial coverage showing a misleading trend).

Fig. 5 summarizes the different DL approaches that have been adopted by researchers to solve specific maintenance tasks in reference Railway Areas. Notably, some architectures have been leveraged to deal with different problems in different areas and have been also exploited for different DL tasks (i.e., Classification, Object Detection, and Semantic Segmentation). However, some pairs are missing:

- Rolling Stock - Surface Defect Detection: the studies were oriented to detect defective parts of relevant components (e.g., broken or missing bolts) and not to the identification of surface defects such as scratches on the train body;
- Tunnel & Bridge - Defect Inspection: as expected, the studies were oriented to identifying only surface defects on tunnel/bridge linings;
- Rolling Stock, Tunnel & Bridge - Object Identification: in all papers, the authors proposed a methodology aimed at the detection of defects, while no study addressing object identification was found.

It is worth mentioning that we also found a single paper [164] focusing on track-side equipment anomaly detection. Besides it is not related to any of the four main relevant railway areas, a mention is deserved, as it uses a DL approach (MDNet [165], [166]) trained on 600 images ($2048 \times 1011$ px each) to identify defective objects in trackside components. The approach achieves a True Positive Rate of 98.9%.

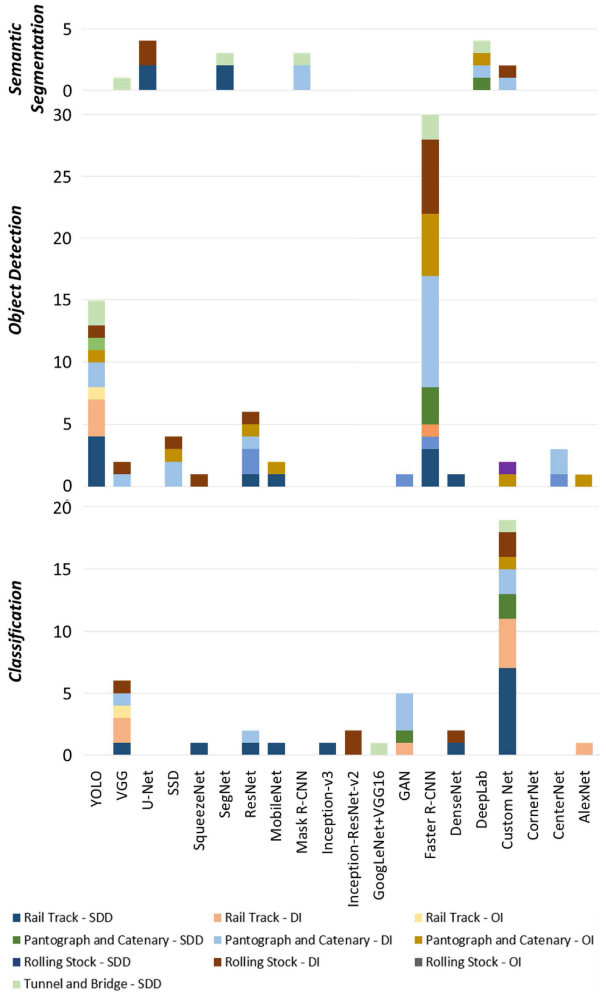



**FIGURE 5.** DL Architectures. The chart shows the cumulative number of times that the specific network or architecture has been used to address a Classification, Object Detection, or Semantic Segmentation task in the reference railway areas for Surface Defect Detection (SDD), Defect Inspection (DI) or Object Identification (OI). OI refers to the papers we have identified as "Promising Works", which only perform object detection (no defects) or classification. Notably: i) the Faster R-CNN is an architecture involving a backbone network to extract features, with networks belonging to the ResNet series being the most adopted to this aim; ii) YOLOv3 was the most exploited YOLO architecture; iii) with "Custom Net" we indicate approaches that were not (explicitly) based on SOTA networks.

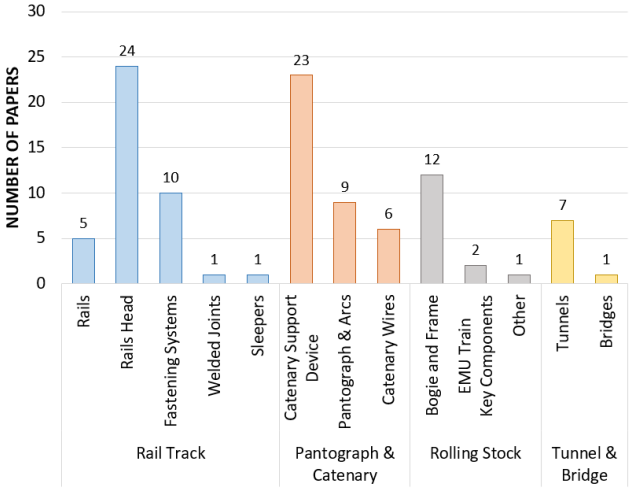



**FIGURE 6.** Distribution of papers by sub-components. The histogram shows the number of studies found for each sub-component of the Railway Areas. The total number of items in each area may be greater than that indicated in Fig. 3 as some papers deal with multiple sub-components.

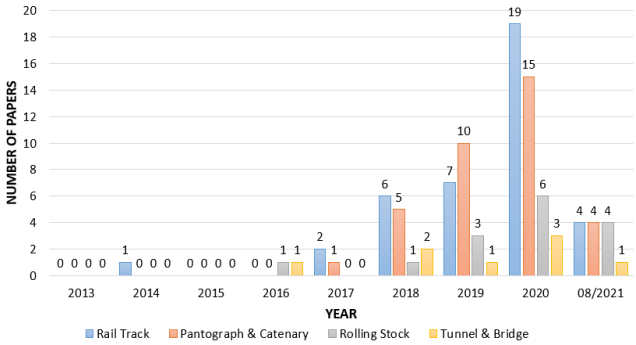



**FIGURE 7.** Trends of the papers by Railway Areas. For each Railway Area, the chart shows the trend of studies over the years. Interestingly, the Rail Track, Pantograph & Catenary, and the Rolling Stock areas show more or less the same increasing trend, even though studies in the Rolling Stock area 'started' one year later. On the other hand, the 'few' studies in the Tunnel & Bridge area show a fairly flat trend starting from 2016.

## IX. RESPONSES TO RESEARCH QUESTIONS

In Sections from IV to VII, we reviewed DL approaches that have been adopted in the context of Rail Tracks, Catenary & Pantograph, Rolling Stock, and Tunnel & Bridge, to address defect detection and inspection. Specifically, Tables 2, 4, 6, and 8 summarise the main characteristics of these approaches, while Tables 1, 3, 5, and 7 report the maintenance tasks per railway component (e.g., CSD, fastening systems, etc.). From these analyses, we derived the following answers to the three RQs we have previously identified.

### A. RESPONSE TO RQ1

We reviewed 95 papers distributed as shown in Fig. 3. The "Rail Track" and "Pantograph & Catenary" are the most investigated areas. To better understand trends and research directions, Figs. 6 and 7 report the papers grouped by components within each area. The most investigated components are "Rails' Heads" and "Catenary Support Device", discussed in Sections IV-A and V-A respectively, with papers providing suitable and effective approaches. Differently, for other sub-components, the outcome was not expected, especially for cracks in tunnels and bridges (which account for only 8 papers out of 95), since such defects have already been analysed in other domains. Evidence for this are: i) a study related to the road sector [167]; ii) the datasets it uses (CRACK500 [168], [169], GAPs384 [170], and CFD [171]), which are related to different types of cracks; iii) two approaches based on transfer learning we reviewed in this SLR, i.e., reference [36], using a general crack dataset [94], and reference [163], using the VOC2007 dataset [131]. Fig. 7 also highlights interesting aspects of

publication trends per area. In particular, papers focusing on both ''Rail Track'', ''Pantograph & Catenary'' and ''Rolling Stock'' show a growing trend (with the latter mostly pushed by papers on Bogies, suggesting this to be a hot topic in the future), while those focusing on ''Tunnel & Bridge'' appears almost steady.

### B. RESPONSE TO RQ2

Several object detection approaches have been implemented for Rail Tracks, Bogies and, especially, Catenaries. Indeed, the latter are composed of very different parts (even very small), often causing ineffective defect detection on the acquired images. To deal with this, different multi-stage approaches have been proposed, with the first stage providing the detection of the area of the component under inspection (e.g., [99], [100]). Among the various object detection methods, the most used approaches are those belonging to the YOLO (mostly Yolov3) family, as they seem to be among the fastest architectures, and those belonging to the region-based CNN (mostly Faster R-CNN) family, which have been shown to be (generally) more precise even if less fast than the previous [172]. Such models have been used both as defect detectors (e.g., [118], [125]), evaluating defects starting from the original images, and as a first-stage approach in the multi-stage solutions, to detect the portion of the image related to the components of interest (e.g., [95], [105]). This aspect can be easily derived from Fig. 5.

As for the datasets, unfortunately, most of the studies leveraged ad-hoc datasets built from scratch; additionally, data have been collected through different methods and sensors. However, as from Tables 2, 4, 6, and 8, some works also make use of (new or already existing) datasets that are narrowed to specific railway components (e.g., Type-I and Type-II [73], [74], Catenary-5000 [132] and BSD500 [154], etc.). Additionally, only a few works made available online the datasets they built (e.g., [62], [64]). The heterogeneity of datasets represents an obstacle to identifying the best approach among those proposed in the relevant literature.

### C. RESPONSE TO RQ3

We have identified the following main research directions based on the issues highlighted by the works examined and their proposed solutions:

#### 1) DEFECTS IN CONCRETE INFRASTRUCTURES

Considering the analysis conducted on concrete infrastructures in section VII, we can state that only a few studies have been carried out in this direction. As already mentioned, these studies are *scattered* over time, and it seems that no particular emphasis has been placed on this topic in the railway sector. Therefore, starting from existing studies (including those not directly related to the rail sector [173]) and datasets, further analyses could be carried out on the detection of surface defects in concrete infrastructures, especially considering that, as expressed in some of the reviewed studies, delineating defects inside tunnels is not straightforward given some unfavourable environmental conditions (e.g., uneven illumination).

#### 2) SMALL-SCALE OBJECT DETECTION

The small-scale object detection introduces difficulties in tasks such as defect inspection of CSD and bogies components, leading some researchers to focus only on component detection (e.g., [111], [115], [129]). Building multi-stage approaches as discussed in the response to RQ2 could be a suitable solution to deal with this issue. Additionally, the ground laid by some of the reviewed studies could lead to further improved methods for detecting small-scale objects, and thus further improve maintenance tasks. In this respect, the authors in reference [103] presented a quite innovative approach to improve insulator detection by leveraging rotated anchors instead of horizontal or vertical ones.

#### 3) MULTI-TARGET SYSTEMS

Most of the reviewed papers focused on a single target or a class of targets such as rails' heads, fasteners, and insulators (or a few of the CSD components). Future research should aim at implementing multi-target systems focusing on more than one component. That can be done by combining and improving some of the proposed methods in order to detect defects of multiple components by leveraging the same video and audio data. Some steps have already been taken in this direction by evaluating rails' heads and fasteners' defects at the same time [48]. The same approach could be applied to CSD, pantographs, and arcs.

#### 4) HARMONISED DATASETS AND BENCHMARKS

The availability and quality of data have always been one of the major challenges when it comes to properly characterise ML and DL models. Additionally, most of the studies built an ad-hoc dataset from scratch to conduct their experiments, also to solve the same task on the same rail component (e.g., rails' heads). This means that: i) it is not possible to immediately identify the best approach to solve a given problem; ii) it is not said that if a model works well on a given dataset, it will be able to perform properly on another dataset, possibly collected in a different way. Hence, besides DL-related concepts such as the ''generalization'' of the model, it would be advisable to define some guidelines for data collection and dataset construction. These guidelines should allow the creation of harmonised datasets, i.e., datasets that share similar characteristics, including data format, data quality, and data variety (e.g., under different weather or illumination conditions). Notably, data quality should not only address the resolution of the image, but it should also take into account the number of collected samples and how relevant they are to solve the task. Harmonised datasets, in our view, would facilitate performance evaluation and enable comparison among different models. Reference datasets, possibly managed by reputable railway stakeholders, would enable reliable benchmarks to assess the performance of DL solutions and effectively support future research activities.

#### 5) LEVERAGING AUDIO DATA

Our results showed a strong deficiency in the use of audio-based techniques to support maintenance activities in railways (Fig. 4). Indeed, in only a few of the reviewed papers, mainly related to the Rail Track area, the authors leveraged audio data: reference [35] proposed a model to identify the wheel-rails adhesion conditions; reference [39] proposed an approach to identify rail crack status and, similarly, reference [40] aimed at detecting rail cracks; in reference [63], laboratory experiments are described to detect rails head defects, while, to address the same task, in reference [54], the authors used acoustic emissions recorded on the field and implemented a more complex approach. Therefore, based on our analysis and due to its known potential of warning maintenance operators about defects, the exploitation of audio data through DL is an aspect worthy of further investigation and extensive evaluation.

#### 6) COMBINING AUDIO AND VIDEO DATA

No study has attempted to combine data from both audio and video sources. However, in the field of data-driven approaches and DL, there is plenty of feature fusion techniques that may be exploited to merge data from multiple sources, therefore, such a combination has a huge potential. In fact, by focusing on the same target, the models built on the two different data sources could be synergistic and complete each other to reduce their shortcomings and ultimately lead to more accurate defect detection.

## X. DISCUSSION

The common idea shared by a number of the reviewed works (e.g., [48], [140], and [143]) has been to move towards data-driven methods implementing mechanisms to improve traditional inspection and maintenance activities, which are often time-consuming and limited by the experience of the operators. It is worth noting that, although our scope was to review works addressing DL approaches to perform or facilitate maintenance and inspection activities by Audio-Video Analytics in railways, search results also included several approaches based on traditional Image Processing techniques for image pre-processing (e.g., image enhancing, noise reduction), feature extraction and image segmentation (e.g., [67], [134], [114]), and defects evaluation (e.g., [46], [113], [125], [126], [151]). Although some of those works are very recent and interesting, they tend to show limited performance when compared against DL, especially in unfavourable conditions, such as in tunnels (e.g., uneven illumination, image noise, etc.). On the other hand, DL approaches often require more data and computational power to be trained, as well as hardware acceleration (e.g., GPU, Graphics Processing Unit) even for the inference phase. This latter point is particularly critical for an embedded application designed to operate in near real-time and/or at very high frame rates, often by using batteries (e.g., on track monitoring, without direct access to the power grid) or to a suitable cooling system. Despite this, the achievements made by deep neural networks in recent years are strongly pushing the research towards new procedures, techniques and devices able to cope with the aforementioned points, making DL more and more appealing even for inspection and maintenance applications in many domains, including railway transport. In particular, the growing interest in DL in the railway domain is demonstrated by the studies we have reported in Tables 2, 4, 6, and 8, although our work focuses on works leveraging DL for audio and video analyses, motivated by the necessity to analyse the current state-of-the-art on the usage of non-intrusive, possibly cheap and available sensors.

In Section I, we already discussed why non-intrusive and cost-effective sensors, such as cameras and microphones, could be more appealing and advantageous than ''special sensors'' (e.g., laser, ultrasonic, etc.) that might be expensive in terms of infrastructures and data analysis. Thus, cameras and microphones enable a transition from scheduled inspection activities using expensive inspection vehicles to continuous monitoring of railway assets. Furthermore, modern smart-cameras used for railway monitoring and surveillance feature high definition, on-board computation capabilities, and are often equipped with quality microphones or microphone inputs, thus it is be possible to collect multiple types of data with a single device. Therefore, independent DL models, which elaborate diverse data can be built to increment detection performance, as discussed in Section IX-C6. As for the small-scale object detection issue (Section IX-C2), it depends mainly on the data acquisition system; by collecting track images through UAVs (e.g., [66]), it is clear that the fasteners will occupy only a small portion of the image. Hence, this issue is shared among almost all areas and their corresponding components; the same holds for the multi-target problem (Section IX-C3). Components such as CSD, bogies, and tracks are typically composed of a large number of ''sub-components''; by evaluating the status of only one or few of them, it would not be sufficient to ensure an adequate safety level for the whole component. Image analysis has the potential to capture characteristics related to multiple sub-components at once; therefore, it would be possible to build systems capable of detecting multiple defects to improve maintenance and inspection activities.

However, there might be some limitations related to the usage of audio and video sensors. For example, captured images might be blurred [43] given the train speed or vibration, objects of interest may be occluded by other objects [99], and – as already mentioned above – there might be a foreground-background imbalance [60] as railways components usually occupy only a part of the image. Similarly, audio might be distorted due to other noise and microphones might need a specific orientation [35]. On the other hand, from the data perspective, collected samples may be altered by natural phenomena (e.g., weather conditions, light reflection – especially when considering rails' heads [60]), or, given the rarity and different shapes/classes of the various possible defects,

datasets built from scratch may result to be unbalanced[11] [56], [71]. Part of these issues (e.g., blurring, noises) represent technological challenges which are related to the choice of the sensor(s). The remaining issues, instead, are related to methodological aspects encompassing both data acquisition and the choice of the DL model. Part of these challenges may be overcome by means of harmonised datasets and collection guidelines (Section IX-C4), and indicating the adequate set of sensors for the given task. Notably, using harmonised datasets, intended as a balanced collection of samples gathered under different weather and light conditions, may bring two positive side effects: they can be used as benchmarks to compare different approaches or as reference datasets for transfer learning.

As a final consideration, it is interesting to note that, despite some works use IoT for data acquisition, during our analysis, we did not find any work relying on edge computing, i.e., the use of embedded (possibly custom) hardware to run AI applications where needed. This is partly expected as almost all the approaches rely on powerful and energivorous GPUs, however recent development of embedded GPU and architectures is more and more opening for massive use of AI at the edge, which is sometimes referred to as edge intelligence [174]. Thus, despite not yet being exploited, we believe this will probably be a very promising research direction in the near future.

## XI. CONCLUSION

Within AI, artificial vision and audio signal recognition are two extremely relevant research areas with many promising applications in several domains [175]. Those areas are among the ones that are expected to benefit more from modern deep learning paradigms. Among the many domains and applications where audio-video analytics based on deep learning could succeed, railway maintenance has been explored by researchers and practitioners with the aim of making it more: effective, i.e., capable of detecting more faults and defects before they can cause failures and harmful consequences, and being less error-prone compared to human inspection; and efficient, i.e., requiring less time, resources and efforts through remote monitoring and cost-effective automation. Those recent and planned advances in railway maintenance are sometimes referred to as smart maintenance, including predictive analytics, usage of IoT sensors, intelligent cameras and microphones, autonomous drones (see e.g. [176]), etc. However, one of the main issues limiting the development of effective AI solutions in the railway domain is the lack of harmonised, well-described, open-source or public available datasets and benchmarks for the broad set of applications experts have to cope with. In this paper, we have summarized the state-of-the-art in this field by adopting a review approach whose results are reproducible and the risk of missing relevant results is minimum. We have provided a thorough classification, representation and extensive discussion of the review results throughout the paper, by highlighting current challenges and some pointers to promising research directions and future developments. It is our expectation that this work can effectively support research and development connected with the usage of smart audio and video sensors to significantly aid railway maintenance by leveraging novel deep learning technologies.

[11] High ratio between different classes in terms of the number of samples.

## DISCLAIMER

The information and views set out in this document are those of the author(s) and do not necessarily reflect the official opinion of Shift2Rail Joint Undertaking (JU) which funded the research. The JU does not guarantee the accuracy of the data included in this document. Neither the JU nor any person acting on the JU's behalf may be held responsible for the use which may be made of the information contained therein.

## REFERENCES

[1] S. Yella, M. S. Dougherty, and N. K. Gupta, "Artificial intelligence techniques for the automatic interpretation of data from non-destructive testing," *Insight-Non-Destructive Test. Condition Monitor.*, vol. 48, no. 1, pp. 10–20, Jan. 2006.

[2] X. Tu, C. Xu, S. Liu, S. Lin, L. Chen, G. Xie, and R. Li, "LiDAR point cloud recognition and visualization with deep learning for overhead contact inspection," *Sensors*, vol. 20, no. 21, p. 6387, Nov. 2020.

[3] J. Suzumura, Y. Sone, A. Ishizaki, D. Yamashita, Y. Nakajima, and M. Ishida, "*In situ* X-ray analytical study on the alteration process of iron oxide layers at the railhead surface while under railway traffic," *Wear*, vol. 271, nos. 1–2, pp. 47–53, May 2011.

[4] H. Cui, J. Li, Q. Hu, and Q. Mao, "Real-time inspection system for ballast railway fasteners based on point cloud deep learning," *IEEE Access*, vol. 8, pp. 61604–61614, 2020.

[5] D. Sasi, S. Philip, R. David, and J. Swathi, "A review on structural health monitoring of railroad track structures using fiber optic sensors," *Mater. Today*, vol. 33, pp. 3787–3793, Jan. 2020.

[6] R. Tang, L. De Donato, N. Bešinović, F. Flammini, R. M. P. Goverde, Z. Lin, R. Liu, T. Tang, V. Vittorini, and Z. Wang, "A literature review of artificial intelligence applications in railway systems," *Transp. Res. C, Emerg. Technol.*, vol. 140, Jul. 2022, Art. no. 103679. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0968090X22001206

[7] *Intelligent Innovative Smart Maintenance of Assets by IntegRated Technologies (IN2SMART)*. Accessed: May 18, 2022. [Online]. Available: https://projects.shift2rail.org/s2r_ip3_n.aspx?p=IN2SMART

[8] *Intelligent Innovative Smart Maintenance of Assets by integRated Technologies 2 (IN2SMART2)*. Accessed: May 18, 2022. [Online]. Available: https://projects.shift2rail.org/s2r_ip3_n.aspx?p=IN2SMART2

[9] G. Bocchetti, F. Flammini, and A. Pappalardo, "Dependable integrated surveillance systems for the physical security of metro railways," in *Proc. 3rd ACM/IEEE Int. Conf. Distrib. Smart Cameras (ICDSC)*, Aug. 2009, pp. 1–7.

[10] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature Med.*, vol. 25, no. 1, pp. 24–29, Jan. 2019.

[11] B. V. Wee and D. Banister, "How to write a literature review paper?" *Transp. Rev.*, vol. 36, no. 2, pp. 278–288, Mar. 2016.

[12] B. Kitchenham, "Procedures for performing systematic reviews," Dept. Comput. Sci., Keele Univ., Keele, U.K., Tech. Rep. 0400011T.1, 2004, pp. 1–26, vol. 33.

[13] *Roadmaps for AI Integration in the Rail Sector (RAILS)*. Accessed: May 18, 2022. [Online]. Available: https://rails-project.eu/

[14] M. C. Nakhaee, D. Hiemstra, M. Stoelinga, and M. van Noort, "The recent applications of machine learning in rail track maintenance: A survey," in *Proc. Int. Conf. Rel., Saf., Secur. Railway Syst.* Cham, Switzerland: Springer, 2019, pp. 91–105.

[15] J. Xie, J. Huang, C. Zeng, S.-H. Jiang, and N. Podlich, "Systematic literature review on data-driven models for predictive maintenance of railway track: Implications in geotechnical engineering," *Geosciences*, vol. 10, no. 11, p. 425, Oct. 2020.
[16] T. P. Carvalho, F. A. A. M. N. Soares, R. Vita, R. D. P. Francisco, J. P. Basto, and S. G. S. Alcalá, "A systematic literature review of machine learning methods applied to predictive maintenance," *Comput. Ind. Eng.*, vol. 137, Nov. 2019, Art. no. 106024.
[17] H. M. Hashemian, "State-of-the-art predictive maintenance techniques," *IEEE Trans. Instrum. Meas.*, vol. 60, no. 1, pp. 226–236, Jan. 2011.
[18] L. Zhang, J. Lin, B. Liu, Z. Zhang, X. Yan, and M. Wei, "A review on deep learning applications in prognostics and health management," *IEEE Access*, vol. 7, pp. 162415–162438, 2019.
[19] M. A. B. Fayyaz, A. C. Alexoulis-Chrysovergis, M. J. Southgate, and C. Johnson, "A review of the technological developments for interlocking at level crossing," *Proc. Inst. Mech. Eng., F, J. Rail Rapid Transit*, vol. 235, no. 4, pp. 1–11, 2020.
[20] S. Liu, Q. Wang, and Y. Luo, "A review of applications of visual inspection technology based on image processing in the railway industry," *Transp. Saf. Environ.*, vol. 1, no. 3, pp. 185–204, Dec. 2019.
[21] Y. Liu, X. Sun, and J. H. L. Pang, "A YOLOv3-based deep learning application research for condition monitoring of rail thermite welded joints," in *Proc. 2nd Int. Conf. Image, Video Signal Process.*, Mar. 2020, pp. 33–38.
[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097–1105.
[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[25] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[26] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 580–587.
[27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," 2015, *arXiv:1506.01497*.
[28] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.
[29] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
[30] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.
[31] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[32] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Jan. 2017.
[33] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.
[34] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 131–135.
[35] S. Shrestha, A. Koirala, M. Spiryagin, and Q. Wu, "Wheel-rail interface condition estimation via acoustic sensors," in *Proc. Joint Rail Conf.*, Apr. 2020, pp. 1–6.
[36] J. S. Lee, S. H. Hwang, I. Y. Choi, and Y. Choi, "Estimation of crack width based on shape-sensitive kernels and semantic segmentation," *Struct. Control Health Monitor.*, vol. 27, no. 4, p. e2504, 2020.
[37] D. F. García, I. García, F. J. D. Calle, and R. Usamentiaga, "A configuration approach for convolutional neural networks used for defect detection on surfaces," in *Proc. 5th Int. Conf. Math. Comput. Sci. Ind. (MCSI)*, Aug. 2018, pp. 44–51.
[38] Q. Xu, Q. Zhao, G. Yu, L. Wang, and T. Shen, "Rail defect detection method based on recurrent neural network," in *Proc. 39th Chin. Control Conf. (CCC)*, Jul. 2020, pp. 6486–6490.
[39] D. Li, Y. Wang, W.-J. Yan, and W.-X. Ren, "Acoustic emission wave classification for rail crack monitoring based on synchrosqueezed wavelet transform and multi-branch convolutional neural network," *Struct. Health Monit.*, vol. 20, no. 4, pp. 1563–1582, 2021.
[40] K. Wang, X. Zhang, Q. Hao, Y. Wang, and Y. Shen, "Application of improved least-square generative adversarial networks for rail crack detection by AE technique," *Neurocomputing*, vol. 332, pp. 236–248, Mar. 2019.
[41] D. Soukup and R. Huber-Mörk, "Convolutional neural networks for steel surface defect detection from photometric stereo images," in *Proc. Int. Symp. Vis. Comput.* Cham, Switzerland: Springer, 2014, pp. 668–677.
[42] Y. Santur, M. Karaköse, and E. Akin, "A new rail inspection method based on deep learning using laser cameras," in *Proc. Int. Artif. Intell. Data Process. Symp. (IDAP)*, Sep. 2017, pp. 1–6.
[43] Y. Santur, M. Karaköse, and E. Akin, "An adaptive fault diagnosis approach using pipeline implementation for railway inspection," *Turkish J. Elect. Eng. Comput. Sci.*, vol. 26, no. 2, pp. 987–998, 2018.
[44] N. AlNaimi and U. Qidwai, "IoT based on-the-fly visual defect detection in railway tracks," in *Proc. IEEE Int. Conf. Informat., IoT, Enabling Technol. (ICIoT)*, Feb. 2020, pp. 627–631.
[45] H. Yuan, H. Chen, S. Liu, J. Lin, and X. Luo, "A deep convolutional neural network for detection of rail surface defect," in *Proc. IEEE Vehicle Power Propuls. Conf. (VPPC)*, Oct. 2019, pp. 1–4.
[46] A. James, W. Jie, Y. Xulei, Y. Chenghao, N. B. Ngan, L. Yuxin, S. Yi, V. Chandrasekhar, and Z. Zeng, "TrackNet—A deep learning based fault detection for railway track inspection," in *Proc. Int. Conf. Intell. Rail Transp. (ICIRT)*, Dec. 2018, pp. 1–5.
[47] Z. Liang, H. Zhang, L. Liu, Z. He, and K. Zheng, "Defect detection of rail surface with deep convolutional neural networks," in *Proc. 13th World Congr. Intell. Control Automat. (WCICA)*, 2018, pp. 1317–1322.
[48] X. Wei, D. Wei, D. Suo, L. Jia, and Y. Li, "Multi-target defect identification for railway track line based on image processing and improved YOLOv3 model," *IEEE Access*, vol. 8, pp. 61973–61988, 2020.
[49] S. Yanan, Z. Hui, L. Li, and Z. Hang, "Rail surface defect detection method based on YOLOv3 deep learning networks," in *Proc. Chin. Automat. Congr. (CAC)*, Nov. 2018, pp. 1563–1568.
[50] Y. Cheng, D. HongGui, and F. YuXin, "Effects of faster region-based convolutional neural network on the detection efficiency of rail defects under machine vision," in *Proc. IEEE 5th Inf. Technol. Mechatronics Eng. Conf. (ITOEC)*, Jun. 2020, pp. 1377–1380.
[51] J. Lu, B. Liang, Q. Lei, X. Li, J. Liu, J. Liu, J. Xu, and W. Wang, "SCueU-Net: Efficient damage detection method for railway rail," *IEEE Access*, vol. 8, pp. 125109–125120, 2020.
[52] X. Chen and H. Zhang, "Rail surface defects detection based on faster R-CNN," in *Proc. Int. Conf. Artif. Intell. Electromechanical Automat. (AIEA)*, Jun. 2020, pp. 819–822.
[53] D. Zhang, K. Song, Q. Wang, Y. He, X. Wen, and Y. Yan, "Two deep learning networks for rail surface defect inspection of limited samples with line-level label," *IEEE Trans. Ind. Informat.*, vol. 17, no. 10, pp. 6731–6741, Oct. 2021.
[54] S.-X. Chen, L. Zhou, Y.-Q. Ni, and X.-Z. Liu, "An acoustic-homologous transfer learning approach for acoustic emission–based rail condition evaluation," *Struct. Health Monit.*, vol. 20, no. 4, pp. 2161–2181, 2021.
[55] S. Li, P. Li, Y. Zhang, and X. Zhao, "Detection of component types and track damage for high-speed railway using region-based convolutional neural networks," *Smart Mater., Adapt. Struct. Intell. Syst.*, vol. 51951, Nov. 2018, Art. no. V002T05A012.
[56] L. Shang, Q. Yang, J. Wang, S. Li, and W. Lei, "Detection of rail surface defects based on CNN image recognition and classification," in *Proc. 20th Int. Conf. Adv. Commun. Technol. (ICACT)*, Feb. 2018, pp. 45–51.
[57] X. Li, Y. Zhou, and H. Chen, "Rail surface defect detection based on deep learning," in *Proc. 11th Int. Conf. Graph. Image Process. (ICGIP)*, Jan. 2020, Art. no. 113730.
[58] J. H. Feng, H. Yuan, Y. Q. Hu, J. Lin, S. W. Liu, and X. Luo, "Research on deep learning method for rail surface defect detection," *IET Electr. Syst. Transp.*, vol. 10, no. 4, pp. 436–442, Dec. 2020.

[59] P. S. Heyns, R. Deetlefs, A. Oberholster, T. Botha, P. S. Els, and D. Diamond, "Computer vision for rail surface defect detection," in *Proc. 16th Int. Conf. Condition Monit. Asset Manage. (CM)*, 2019.

[60] X. Ni, Z. Ma, J. Liu, B. Shi, and H. Liu, "Attention network for rail surface defect detection via consistency of intersection-over-union (IoU)-guided center-point estimation," *IEEE Trans. Ind. Informat.*, vol. 18, no. 3, pp. 1694–1705, Mar. 2022.

[61] Y. Wu, Y. Qin, Y. Qian, F. Guo, Z. Wang, and L. Jia, "Hybrid deep learning architecture for rail surface segmentation and surface defect detection," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 37, no. 2, pp. 227–244, Feb. 2022.

[62] I. Aydin, E. Akin, and M. Karakose, "Defect classification based on deep features for railway tracks in sustainable transportation," *Appl. Soft Comput.*, vol. 111, Nov. 2021, Art. no. 107706.

[63] X. Zhang, K. Wang, Y. Wang, Y. Shen, and H. Hu, "An improved method of rail health monitoring based on CNN and multiple acoustic emission events," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf. (IMTC)*, May 2017, pp. 1–6.

[64] S. Iyer, T. Velmurugan, A. Gandomi, V. N. Mohammed, K. Saravanan, and S. Nandakumar, "Structural health monitoring of railway tracks using IoT-based multi-robot system," *Neural Comput. Appl.*, vol. 33, pp. 1–19, Sep. 2020.

[65] Y.-W. Lin, C.-C. Hsieh, W.-H. Huang, S.-L. Hsieh, and W.-H. Hung, "Railway track fasteners fault detection using deep learning," in *Proc. IEEE Eurasia Conf. IoT, Commun. Eng. (ECICE)*, Oct. 2019, pp. 187–190.

[66] P. Chen, Y. Wu, Y. Qin, H. Yang, and Y. Huang, "Rail fastener defect inspection based on UAV images: A comparative study," in *Proc. Int. Conf. Elect. Inf. Technol. Rail Transp.* Singapore: Springer, 2020, pp. 685–694.

[67] X. Wei, Z. Yang, Y. Liu, D. Wei, L. Jia, and Y. Li, "Railway track fastener defect detection based on image processing and deep learning techniques: A comparative study," *Eng. Appl. Artif. Intell.*, vol. 80, pp. 66–81, Apr. 2019.

[68] Y. Zhan, X. Dai, E. Yang, and K. C. P. Wang, "Convolutional neural network for detecting railway fastener defects using a developed 3D laser system," *Int. J. Rail Transp.*, vol. 9, no. 5, pp. 424–444, 2021.

[69] C.-C. Hsieh, Y.-W. Lin, L.-H. Tsai, W.-H. Huang, S.-L. Hsieh, and W.-H. Hung, "Offline deep-learning-based defective track fastener detection and inspection system," *Sensors Mater.*, vol. 32, no. 10, pp. 3429–3442, 2020.

[70] Y. Zhou, X. Li, and H. Chen, "Railway fastener defect detection based on deep convolutional networks," in *Proc. 11th Int. Conf. Graph. Image Process. (ICGIP)*, Jan. 2020, Art. no. 113732.

[71] D. Yao, Q. Sun, J. Yang, H. Liu, and J. Zhang, "Railway fastener fault diagnosis based on generative adversarial network and residual network model," *Shock Vib.*, vol. 2020, pp. 1–15, Nov. 2020.

[72] S. P. Orlov, R. V. Girin, and A. V. Piletskaya, "Intelligent information processing system for monitoring rail tracks," in *Proc. 3rd Int. Conf. Control Tech. Syst. (CTS)*, Oct. 2019, pp. 233–236.

[73] Q. Li and S. Ren, "A real-time visual inspection system for discrete surface defects of rail heads," *IEEE Trans. Instrum. Meas.*, vol. 61, no. 8, pp. 2189–2199, Aug. 2012.

[74] J. Gan, Q. Li, J. Wang, and H. Yu, "A hierarchical extractor-based visual rail surface inspection system," *IEEE Sensors J.*, vol. 17, no. 23, pp. 7935–7944, Dec. 2017.

[75] P. Rizzo, "Sensing solutions for assessing and monitoring railroad tracks," in *Sensor Technologies for Civil Infrastructures*. Sawston, U.K.: Woodhead Publishing, 2014, pp. 497–524.

[76] J. Wang, X.-Z. Liu, and Y.-Q. Ni, "A Bayesian probabilistic approach for acoustic emission-based rail condition assessment," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 33, no. 1, pp. 21–34, Jan. 2018.

[77] L. Zhou, X.-Z. Liu, and Y.-Q. Ni, "Contemporary inspection and monitoring for high-speed rail system," in *High-Speed Rail*. London, U.K.: IntechOpen, 2018. [Online]. Available: https://www.intechopen.com/chapters/64211, doi: 10.5772/intechopen.81159.

[78] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," 2016, *arXiv:1611.04076*.

[79] D. Li, K. S. C. Kuang, and C. G. Koh, "Rail crack monitoring based on tsallis synchrosqueezed wavelet entropy of acoustic emission signals: A field study," *Struct. Health Monitor.*, vol. 17, no. 6, pp. 1410–1424, Nov. 2018.

[80] K. Kuang, D. Li, and C. G. Koh, "Acoustic emission source location and noise cancellation for crack detection in rail head," *Smart Struct. Syst.*, vol. 18, p. 1063, Jul. 2016.

[81] D. Li, K. S. C. Kuang, and C. G. Koh, "Fatigue crack sizing in rail steel using crack closure-induced acoustic emission waves," *Meas. Sci. Technol.*, vol. 28, no. 6, Jun. 2017, Art. no. 065601.

[82] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 734–750.

[83] M. Niu, K. Song, L. Huang, Q. Wang, Y. Yan, and Q. Meng, "Unsupervised saliency detection of rail surface defects using stereoscopic images," *IEEE Trans. Industr. Inform.*, vol. 17, no. 3, pp. 2271–2281, 2020.

[84] D. Tabernik, S. Šela, J. Skvarč, and D. Skočaj, "Segmentation-based deep-learning approach for surface-defect detection," *J. Intell. Manuf.*, vol. 31, no. 3, pp. 759–776, 2020.

[85] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and $<$ 0.5 MB model size," 2016, *arXiv:1602.07360*.

[86] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[87] Y. Huang, C. Qiu, and K. Yuan, "Surface defect saliency of magnetic tile," *Vis. Comput.*, vol. 36, no. 1, pp. 85–96, 2020.

[88] D. Tang, B. Qin, X. Feng, and T. Liu, "Effective LSTMs for target-dependent sentiment classification," 2015, *arXiv:1512.01100*.

[89] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2005, pp. 524–531.

[90] H. Kato and T. Harada, "Image reconstruction from bag-of-visual-words," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 955–962.

[91] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2006, pp. 2169–2178.

[92] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 776–780.

[93] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[94] S. Li, X. Zhao, and G. Zhou, "Automatic pixel-level multiple damage detection of concrete structure using fully convolutional network," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 34, no. 7, pp. 616–634, Jul. 2019.

[95] G. Kang, S. Gao, L. Yu, and D. Zhang, "Deep architecture for high-speed railway insulator surface defect detection: Denoising autoencoder with multitask learning," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 8, pp. 2679–2690, Aug. 2019.

[96] Z. Gu, Y. Wang, X. Xue, S. Wang, Y. Cheng, X. Du, and P. Dai, "Railway insulator defect detection with deep convolutional neural networks," in *Proc. 12th Int. Conf. Digit. Image Process. (ICDIP)*, Jun. 2020, Art. no. 1151903.

[97] D. Zhang, S. Gao, L. Yu, G. Kang, X. Wei, and D. Zhan, "DefGAN: Defect detection GANs with latent space pitting for high-speed railway insulator," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–10, 2021.

[98] G. Kang, S. Gao, L. Yu, D. Zhang, X. Wei, and D. Zhan, "Contact wire support defect detection using deep Bayesian segmentation neural networks and prior geometric knowledge," *IEEE Access*, vol. 7, pp. 173366–173376, 2019.

[99] J. Chen, Z. Liu, H. Wang, A. Núñez, and Z. Han, "Automatic defect detection of fasteners on the catenary support device using deep convolutional neural network," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 2, pp. 257–269, Feb. 2017.

[100] J. Wang, L. Luo, W. Ye, and S. Zhu, "A defect-detection method of split pins in the catenary fastening devices of high-speed railway based on deep learning," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 12, pp. 9517–9525, Dec. 2020.

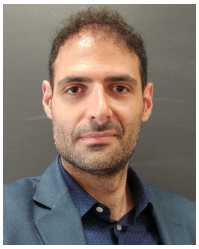
[101] H. Huang, J. Xu, J. Zhang, Q. Wu, and C. Kirsch, "Railway infrastructure defects recognition using fine-grained deep convolutional neural networks," in *Proc. Digit. Image Comput., Techn. Appl. (DICTA)*, 2018, pp. 1–8.

[102] J. Liu, Y. Wu, Y. Qin, H. Xu, and Z. Zhao, "Defect detection for bird-preventing and fasteners on the catenary support device using improved faster R-CNN," in *Proc. 4th Int. Conf. Elect. Inf. Technol. Rail Transp.*, vol. 640. Singapore: Springer, 2020, pp. 695–704.
[103] J. Zhong, Z. Liu, C. Yang, H. Wang, S. Gao, and A. Núñez, "Adversarial reconstruction based on tighter oriented localization for catenary insulator defect detection in high-speed railways," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 2, pp. 1109–1120, Feb. 2022.
[104] Y. Lyu, Z. Han, J. Zhong, C. Li, and Z. Liu, "A GAN-based anomaly detection method for isoelectric line in high-speed railway," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf. (IMTC)*, May 2019, pp. 1–6.
[105] Y. Lyu, Z. Han, J. Zhong, C. Li, and Z. Liu, "A generic anomaly detection of catenary support components based on generative adversarial networks," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 5, pp. 2439–2448, May 2020.
[106] W. Liu, Z. Liu, H. Wang, and Z. Han, "An automated defect detection approach for catenary rod-insulator textured surfaces using unsupervised learning," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 10, pp. 8411–8423, Oct. 2020.
[107] J. Liu, Z. Wang, Y. Wu, Y. Qin, X. Cao, and Y. Huang, "An improved faster R-CNN for UAV-based catenary support device inspection," *Int. J. Softw. Eng. Knowl. Eng.*, vol. 30, no. 7, pp. 941–959, Jul. 2020.
[108] W. Liu, Z. Liu, and A. Núñez, "Virtual reality and convolutional neural networks for railway catenary support components monitoring," in *Proc. IEEE Intell. Transp. Syst. Conf. (ITSC)*, Oct. 2019, pp. 2183–2188.
[109] X. Zhang, Y. Gong, C. Qiao, and W. Jing, "Multiview deep learning based on tensor decomposition and its application in fault detection of overhead contact systems," *Vis. Comput.*, vol. 38, pp. 1–11, Feb. 2021.
[110] Z. Wang, X. Liu, H. Peng, L. Zheng, J. Gao, and Y. Bao, "Railway insulator detection based on adaptive cascaded convolutional neural network," *IEEE Access*, vol. 9, pp. 115676–115686, 2021.
[111] Z. Liu, Y. Lyu, L. Wang, and Z. Han, "Detection approach based on an improved faster RCNN for brace sleeve screws in high-speed railways," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 7, pp. 4395–4403, Jul. 2020.
[112] Z. Liu, J. Zhong, Y. Lyu, K. Liu, Y. Han, L. Wang, and W. Liu, "Location and fault detection of catenary support components based on deep learning," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf. (IMTC)*, May 2018, pp. 1–6.
[113] Z. Liu, L. Wang, C. Li, and Z. Han, "A high-precision loose strands diagnosis approach for isoelectric line in high-speed railway," *IEEE Trans. Ind. Informat.*, vol. 14, no. 3, pp. 1067–1077, Mar. 2018.
[114] Y. Li, D. Wei, X. Wei, K. Wu, Y. Liang, and X. Shen, "Defects detection of catenary suspension device based on image processing and CNN," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Dec. 2019, pp. 1756–1761.
[115] F. Andert, N. Kornfeld, F. Nikodem, H. Li, S. Kluckner, L. Gruber, and C. Kaiser, "Automatic condition monitoring of railway overhead lines from close-range aerial images and video data," in *Proc. Int. Conf. Unmanned Aircr. Syst. (ICUAS)*, Sep. 2020, pp. 1270–1277.
[116] W. Liu, D. Wang, C. Yang, Y. Li, H. Wang, and Z. Liu, "An automatic defect detection method for catenary bracing wire components using deep convolutional neural networks and image processing," in *Proc. Int. Conf. Sens., Meas. Data Anal. Era Artif. Intell. (ICSMD)*, Oct. 2020, pp. 106–111.
[117] X. Wei, S. Jiang, Y. Li, C. Li, L. Jia, and Y. Li, "Defect detection of pantograph slide based on deep learning and image processing technology," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 3, pp. 947–958, Mar. 2020.
[118] S. Jiang, X. Wei, and Z. Yang, "Defect detection of pantograph slider based on improved faster R-CNN," in *Proc. Chin. Control Decis. Conf. (CCDC)*, Jun. 2019, pp. 5278–5283.
[119] G. Karaduman, M. Karakose, and E. Akin, "Deep learning based arc detection in pantograph-catenary systems," in *Proc. 10th Int. Conf. Elect. Electron. Eng.*, 2017, pp. 904–908.
[120] Y. Luo, Q. Yang, and S. Liu, "Novel vision-based abnormal behavior localization of pantograph-catenary for high-speed trains," *IEEE Access*, vol. 7, pp. 180935–180946, 2019.
[121] Y. Shen, Z. Liu, and L. Chang, "A pantograph horn detection method based on deep learning network," in *Proc. IEEE 3rd Optoelectron. Global Conf. (OGC)*, Sep. 2018, pp. 85–89.
[122] X. Yang, N. Zhou, Y. Liu, W. Quan, X. Lu, and W. Zhang, "Online pantograph-catenary contact point detection in complicated background based on multiple strategies," *IEEE Access*, vol. 8, pp. 220394–220407, 2020.
[123] W. Lin, G. Peng, M. Wu, Y. Lin, and L. Jin, "A fault detection method of high speed train pantograph based on deep learning," in *Proc. 8th Int. Conf. Condition Monitor. Diagnosis (CMD)*, Oct. 2020, pp. 254–257.
[124] D. Zhang, S. Gao, L. Yu, G. Kang, D. Zhan, and X. Wei, "A robust pantograph–catenary interaction condition monitoring method based on deep convolutional network," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 5, pp. 1920–1929, May 2020.
[125] P. Tan, X. Li, Z. Wu, J. Ding, J. Ma, Y. Chen, Y. Fang, and Y. Ning, "Multialgorithm fusion image processing for high speed railway dropper failure–defect detection," *IEEE Trans. Syst., Man, Cybern. Syst*, vol. 51, no. 7, pp. 4466–4478, Jul. 2021.
[126] J. Cui, Y. Wu, Y. Qin, and R. Hou, "Defect detection for catenary sling based on image processing and deep learning method," in *Proc. 4th Int. Conf. Elect. Inf. Technol. Rail Transp. (EITRT)*, vol. 640. Singapore: Springer, 2020, pp. 675–683.
[127] C. Huang and Y. Zeng, "The fault diagnosis of catenary system based on the deep learning method in the railway industry," in *Proc. 5th Int. Conf. Multimedia Image Process.*, Jan. 2020, pp. 135–140.
[128] Y. Chen, B. Song, Y. Zeng, X. Du, and M. Guizani, "A deep learning-based approach for fault diagnosis of current-carrying ring in catenary system," *Neural Comput. Appl.*, pp. 1–13, Jul. 2021.
[129] Q. Guo, L. Liu, W. Xu, Y. Gong, X. Zhang, and W. Jing, "An improved faster R-CNN for high-speed railway dropper detection," *IEEE Access*, vol. 8, pp. 105622–105633, 2020.
[130] M. Everingham and J. Winn, "The PASCAL visual object classes challenge 2012 (VOC2012) development kit," *Pattern Anal., Stat. Model. Comput. Learn.*, vol. 8, p. 5, 2011.
[131] M. Everingham, A. W. Zisserman, K. I. Christopher, L. Van Gool, M. Allan, C. M, Bishop, O. Chapelle, N. Dalal, T. Deselaers, and G. Dorkó, "The PASCAL visual object classes challenge 2007 (VOC2007) results," Tech. Rep., 2007.
[132] W. Liu, Z. Liu, A. Núñez, L. Wang, K. Liu, Y. Lyu, and H. Wang, "Multi-objective performance evaluation of the detection of catenary support components using DCNNs," *IFAC-PapersOnLine*, vol. 51, no. 9, pp. 98–105, 2018.
[133] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.
[134] P. Tan, X.-F. Li, J.-M. Xu, J.-E. Ma, F.-J. Wang, J. Ding, Y.-T. Fang, and Y. Ning, "Catenary insulator defect detection based on contour features and gray similarity matching," *J. Zhejiang Univ.-Sci. A*, vol. 21, no. 1, pp. 64–73, Jan. 2020.
[135] X. Gibert, V. M. Patel, and R. Chellappa, "Deep multitask learning for railway track inspection," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 1, pp. 153–164, Jan. 2016.
[136] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, Aug. 2013.
[137] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 2017–2025.
[138] Y. Yu, M. Wang, Z. Wang, and P. Zhou, "Surface defect detection of hight-speed railway hub based on improved YOLOv3 algorithm," in *Proc. IEEE 4th Adv. Inf. Manage., Communicates, Electron. Autom. Control Conf. (IMCEC)*, vol. 4, Jun. 2021, pp. 1386–1390.
[139] Y. Zhan, K. Linb, H. Zhan, Y. Guo, and G. Sun, "A unified framework for fault detection of freight train images under complex environment," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 1348–1352.
[140] Y. Zhang, M. Liu, Y. Chen, H. Zhang, and Y. Guo, "Real-time vision-based system of fault detection for freight trains," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 7, pp. 5274–5284, Jul. 2020.
[141] L. Xiao, B. Wu, Y. Hu, and J. Liu, "A hierarchical features-based model for freight train defect inspection," *IEEE Sensors J.*, vol. 20, no. 5, pp. 2671–2678, Mar. 2019.
[142] J. Sun, Y. Xie, and X. Cheng, "A fast bolt-loosening detection method of running train's key components based on binocular vision," *IEEE Access*, vol. 7, pp. 32227–32239, 2019.
[143] Y. Yang, Y. Hu, L. Chen, X. Liu, N. Qin, and Z. Liu, "Defect detection of axle box cover device fixing bolts in metro based on convolutional neural network," in *Proc. 39th Chin. Control Conf. (CCC)*, Jul. 2020, pp. 7504–7509.
[144] C. Chen, K. Li, C. Zhongyao, F. Piccialli, S. C. H. Hoi, and Z. Zeng, "A hybrid deep learning based framework for component defect detection of moving trains," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 4, pp. 3268–3280, Apr. 2022.

[145] T. Ye, Z. Zhang, X. Zhang, Y. Chen, and F. Zhou, "Fault detection of railway freight cars mechanical components based on multi-feature fusion convolutional neural network," *Int. J. Mach. Learn. Cybern.*, vol. 12, no. 6, pp. 1789–1801, Jun. 2021.
[146] Y. Zhang, M. Liu, Y. Yang, Y. Guo, and H. Zhang, "A unified light framework for real-time fault detection of freight train images," *IEEE Trans. Ind. Informat.*, vol. 17, no. 11, pp. 7423–7432, Nov. 2021.
[147] K. K. Mohan, C. R. Prasad, and P. Kishore, "YOLO v2 with bifold skip: A deep learning model for video based real time train bogie part identification and defect detection," *J. Eng. Sci. Technol.*, vol. 16, no. 3, pp. 2166–2190, 2021.
[148] J. Sun and Z. Xiao, "Potential fault region detection in TFDS images based on convolutional neural network," *Proc. SPIE*, vol. 10157, Oct. 2016, Art. no. 101571L.
[149] B. Zhao, M.-R. Dai, P. Li, and X.-N. Ma, "Data mining in railway defect image based on object detection technology," in *Proc. Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2019, pp. 814–819.
[150] B. Zhao, M. Dai, P. Li, R. Xue, and X. Ma, "Defect detection method for electric multiple units key components based on deep learning," *IEEE Access*, vol. 8, pp. 136808–136818, 2020.
[151] R. S. Pahwa, V. R. Chandrasekhar, J. Chao, J. Paul, Y. Li, M. T. L. Nwe, S. Xie, A. James, A. Ambikapathi, and Z. Zeng, "FaultNet: Faulty rail-valves detection using deep learning and computer vision," in *Proc. IEEE Intell. Transp. Syst. Conf. (ITSC)*, Oct. 2019, pp. 559–566.
[152] Y. Zhang, "Hierarchical feature matching of fault images in TFDS based on improved Markov random field and exact height function," M.S. thesis, Dept. Mech. Eng., Hubei Univ. Technol., Wuhan, China, 2017.
[153] G. Sun, Y. Zhang, H. Tang, H. Zhang, M. Liu, and D. Zhao, "Railway equipment detection using exact height function shape descriptor based on fast adaptive Markov random field," *Opt. Eng.*, vol. 57, no. 5, 2018, Art. no. 053114.
[154] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, Oct. 2010.
[155] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.
[156] K. Song and Y. Yan, "A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects," *Appl. Surf. Sci.*, vol. 285, no. 21, pp. 858–864, Nov. 2013.
[157] Z. Yang, X. Gao, and H. Xia, "An improved faulting detection algorithm for subway tunnel segment," in *Proc. Int. Conf. Wireless Commun., Signal Process. Netw. (WiSPNET)*, Mar. 2016, pp. 1710–1716.
[158] H.-W. Huang, Q.-T. Li, and D.-M. Zhang, "Deep learning based image recognition for crack and leakage defects of metro shield tunnel," *Tunnelling Underground Space Technol.*, vol. 77, pp. 166–176, Jul. 2018.
[159] Y. Xue and Y. Li, "A fast detection method via region-based fully convolutional neural networks for shield tunnel lining defects," *Comput. Aided Civil Infrastruct. Eng.*, vol. 33, no. 8, pp. 638–654, Aug. 2018.
[160] Q. Song, Y. Wu, X. Xin, L. Yang, M. Yang, H. Chen, C. Liu, M. Hu, X. Chai, and J. Li, "Real-time tunnel crack analysis system via deep learning," *IEEE Access*, vol. 7, pp. 64186–64197, 2019.
[161] X. Xu and H. Yang, "Vision measurement of tunnel structures with robust modelling and deep learning algorithms," *Sensors*, vol. 20, no. 17, p. 4945, Sep. 2020.
[162] D. Li, Q. Xie, X. Gong, Z. Yu, J. Xu, Y. Sun, and J. Wang, "Automatic defect detection of metro tunnel surfaces using a vision-based inspection system," *Adv. Eng. Informat.*, vol. 47, Jan. 2021, Art. no. 101206.
[163] G. XingQi, L. Quan, Z. MeiLing, and J. HuiFeng, "Analysis and test of concrete surface crack of railway bridge based on deep learning," in *Proc. IEEE 5th Inf. Technol. Mechatronics Eng. Conf. (ITOEC)*, Jun. 2020, pp. 437–442.
[164] X. Guo, X. Wei, M. Guo, X. Wei, L. Gao, and W. Xing, "Anomaly detection of trackside equipment based on semi-supervised and multi-domain learning," in *Proc. 15th IEEE Int. Conf. Signal Process. (ICSP)*, Dec. 2020, pp. 268–273.
[165] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4293–4302.
[166] I. Jung, J. Son, M. Baek, and B. Han, "Real-time MDNet," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 83–98.
[167] Z. Wu, T. Lu, Y. Zhang, B. Wang, and X. Zhao, "Crack detecting by recursive attention U-Net," in *Proc. 3rd Int. Conf. Robot., Control Autom. Eng. (RCAE)*, Nov. 2020, pp. 103–107.
[168] L. Zhang, F. Yang, Y. Daniel Zhang, and Y. J. Zhu, "Road crack detection using deep convolutional neural network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 3708–3712.
[169] F. Yang, L. Zhang, S. Yu, D. V. Prokhorov, X. Mei, and H. Ling, "Feature pyramid and hierarchical boosting network for pavement crack detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 4, pp. 1525–1535, Apr. 2020.
[170] M. Eisenbach, R. Stricker, D. Seichter, K. Amende, K. Debes, M. Sesselmann, D. Ebersbach, U. Stoeckert, and H.-M. Gross, "How to get pavement distress detection ready for deep learning? A systematic approach," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 2039–2047.
[171] Y. Shi, L. Cui, Z. Qi, F. Meng, and Z. Chen, "Automatic road crack detection using random structured forests," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 12, pp. 3434–3445, Dec. 2016.
[172] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 261–318, 2020.
[173] S. Sony, K. Dunphy, A. Sadhu, and M. Capretz, "A systematic review of convolutional neural network-based structural condition assessment techniques," *Eng. Struct.*, vol. 226, Jan. 2021, Art. no. 111347.
[174] W. Z. Khan, E. Ahmed, S. Hakak, I. Yaqoob, and A. Ahmed, "Edge computing: A survey," *Future Gener. Comput. Syst.*, vol. 97, pp. 219–235, Aug. 2019.
[175] G. Garibotto *et al.*, "White paper on industrial applications of computer vision and pattern recognition," in *Image Analysis and Processing—ICIAP*. Berlin, Germany: Springer, 2013, pp. 721–730.
[176] F. Flammini, C. Pragliola, and G. Smarra, "Railway infrastructure monitoring by drones," in *Proc. Int. Conf. Electr. Syst. Aircr., Railway, Ship Propuls. Road Vehicles Int. Transp. Electrific. Conf. (ESARS-ITEC)*, Nov. 2016, pp. 1–6.

**LORENZO DE DONATO** received the bachelor's and master's *(cum laude)* degrees in computer engineering from the University of Naples Federico II, Italy, in 2016 and 2020, respectively, where he is currently pursuing the Ph.D. degree in information technology and electrical engineering (ITEE). His main research interests include artificial intelligence, specifically deep learning and computer vision, its applicability to the railway sector, and the study of critical railway systems.

**FRANCESCO FLAMMINI** (Senior Member, IEEE) has worked in private and public companies on transportation and infrastructure safety and security projects, from 2003 to 2017. Since 2018, he has been working as an Associate Professor at Linnaeus University, Sweden, where he has chaired the cyber-physical systems (CPS) environment. He is currently a Full Professor of computer science at Mälardalen University, Sweden. He serves as the Technical Manager for the RAILS Research Project.



**STEFANO MARRONE** is currently a Research Fellow and an Assistant Professor at the University of Naples Federico II. More recently, he has been working on ethics, fairness, and privacy in artificial intelligence. His background also covers AI for embedded systems design. His research interests include pattern recognition and computer vision, with applications ranging from biomedical image processing to remote sensing and image/video forensics.

**CLAUDIO MAZZARIELLO** received the Ph.D. degree in computer and automation engineering from the University of Napoli, Italy. He is currently a Senior Innovation Engineer at the Digital and Data Driven Innovation Team, Hitachi Rail STS. He leads an initiative for the definition of research and innovation topics within Shift2Rail for asset management and is an Advisory Board Member of Shift2Rail funded projects. His research interests include AI in computer network security and critical infrastructure protection, especially in the railway domain.

**ROBERTO NARDONE** received the master's degree in computer engineering and the Ph.D. degree in computer and automation engineering from the University of Naples Federico II, Italy, in 2009 and 2013, respectively. He is currently an Assistant Professor at the University of Naples "Parthenope," Italy. His research interests are in the area of V&V of critical systems and include quantitative evaluation of non-functional properties by means of formal methods and model-driven techniques.

**CARLO SANSONE** (Member, IEEE) is currently a Full Professor of computer engineering at the University of Naples Federico II. From an applicative point of view, his main contributions were in the fields of biomedical image analysis, biometrics, and image forensics. He has coordinated several projects, mainly in the areas of biomedical images interpretation and network intrusion detection. His basic research interests include image analysis, pattern recognition, and machine and deep learning. He was a co-editor of three special issues and three books.

**VALERIA VITTORINI** is currently an Associate Professor of computer engineering at the University of Naples Federico II. She has coauthored more than 100 papers in the areas of verification and validation of critical systems, physical protection of critical infrastructure, formal modeling, and model-driven engineering. She is recently investigating the application of AI to the transportation sector and in particular to railway systems. She serves as the Coordinator for the H2020 Shift2Rail Project Roadmaps for A.I. integration in the raiL Sector (RAILS).