Proceedings of the Seventh International Conference on
Railway Technology: Research, Development and Maintenance
Edited by: J. Pombo
Civil-Comp Conferences, Volume 15, Paper 7.6
Civil-Comp Press, Edinburgh, United Kingdom, 2026
ISSN: 2753-3239, doi: 10.4203/ccc.15.7.6

On the Importance of Pre-Training and Spatial Context for Corrugation Detection from In-Service Metro Trains

F. Toschi, M. Carman, A. Facchinetti and C. Somaschini

Politecnico di Milano, Italy

Abstract

We adapt a pre-training framework for corrugation detection using axle-box acceleration signals from commercial trains on the Milan metro network, where the speed, load, and wheel state vary continuously across different runs. Three novel aspects of our approach are the use of: (i) spectrogram pre-processing to account for speed variations, (ii) data augmentation techniques that are designed specifically for spectrograms, and, (iii) a Transformer-based architecture for aggregating contextual features of the track. We validate the approach using a previously developed pipeline achieving a recall of 97.3% on severe corrugation, confirming applicability to the difficult setting of the Milan metro line. Furthermore, when we move to an operationally realistic setting, raising the minimum corrugation detection threshold to 40 micrometers, and extending the feature representation to include a sequence of consecutive spectrograms (covering 6 meters of track), the pipeline achieves an area-under-the-curve of 0.978 and a recall of 94.3% with a false-positive rate of only 1.4% on healthy conditions. Zero-shot evaluation on a held-out section yields a mean Spearman correlation of 0.62, confirming the model generalises well to unseen track geometry.

Keywords: corrugation, spectrograms, pretraining, sequences, transformer, generalization

1 Introduction

Rail corrugation is one of the most critical degradation phenomena affecting railway infrastructure, manifesting as a quasi-periodic irregularity on the running surface of the rail, typically with wavelengths in the range of 30–300 mm [1]. First systematically classified by Grassie and Kalousek, who identified six distinct corrugation types according to their wavelength-fixing and damage mechanisms, the defect arises from complex rail–wheel interaction dynamics and grows progressively under repeated loading. Its presence induces resonance excitation across the track structure and rolling stock, transmitting vibration and structure-borne noise to passengers and nearby buildings. The consequences of such vibrations can be severe enough that affected lines typically require corrective rail grinding every few months, at substantial operational cost. Unlike geometric defects, which produce broadband, spatially stable signatures, corrugation is particularly difficult to detect automatically: its quasi-periodic nature generates narrowband spectral peaks whose frequency shifts with train speed, while the signal undergoes significant attenuation between the wheel–rail contact point and the sensors used in monitoring systems.

Many studies on deep learning (DL)-based rail corrugation typically introduce a new methodology specific to a proprietary dataset for that specific paper. As a result, we rarely see independent replications, where a methodology from one study is re-implemented and re-evaluated on a fundamentally different dataset collected in another railway network. Additionally, many studies present results only in terms of generalisation over time on a single set of track sections, which may hide positional overfitting issues, meaning where the model learns to associate track locations with labels, rather than learning defect-indicative features in the signal. Both of these are critical gaps which must be addressed in order to provide stronger safety guarantees when deploying such models in practice.

Over the past years, machine learning (ML) and DL have been widely adopted for railway health monitoring with a number of different formulations. Some works address this problem with vision-based techniques [2,3], using images acquired by dedicated inspection systems, where convolutional types of architectures have achieved high accuracy in detecting cracks, missing fasteners or other superficial damages. These methods, however, require relatively complex imaging hardware and controlled acquisition conditions, and are often less applicable in practice than approaches based on vibration measurements [4, 5]. These signals are collected via drive-by measurements of dedicated or commercial vehicles. The resulting data can then be used for a wide variety of tasks, ranging from coarse geometric deviations of the rail [6, 7] to short-wavelength, quasi-periodic forms of corrugation [8–10] or in general, concentrated defects in the infrastructure [11, 12]. Within this vibration-based setting, some works focus on comparatively simple and interpretable pipelines, combining physics-based features with classical ML models like support vector machines (SVMs), at times assisted by simulated data to compensate for limitations in data collection campaigns and the rarity of specific defects [11, 13]. Other studies adopt end-to-end DL architectures, including convolutional neural networks (CNNs) operating on raw signals [6] or spectrograms [8], as well as sequence-based models like long short-term memory networks (LSTMs) [6], gated recurrent units (GRUs) [14], or transformers [7]. This diversity of formulation, coupled with the strict limitations on proprietary data sharing, makes it extremely difficult to assess to what extent specific design choices actually confer robustness and generalisation capabilities.

In this work, we build on the pretrain–finetune pipeline of [8] by re-implementing it on a dataset collected from commercial trains running multiple times daily on the Milan metro network, and by designing experiments that explicitly probe cross-track generalisation under real-world operating variability.

We first demonstrate that the core results of [8] are replicable under considerably harder conditions: varying speed, passenger load, and wheel out-of-roundness, at larger scale and with simpler model architectures. Where [8] reports accuracies of 94–100% on controlled data, we obtain 81.5% on moderate-to-severe corrugation cases (50–$100\mu m$) and 97.3% on severe cases ($> 100\mu m$), confirming the approach transfers to less controlled operational settings.

Beyond replication, we introduce two extensions: a higher classification threshold ($40\mu m$ versus the $20\mu m$ proposed in [8]) and sequence modelling-based formulations. Choosing a higher classification threshold makes the problem slightly harder, with accuracy on severe examples slightly reducing to 89.7% with the same pipeline as in the $20\mu m$ threshold experiment. However, moving to a sequence-based modelling approach yields the best results of this study, improving Recall on severe cases from 69.7% to 77.8% and reaching 94.4% accuracy on the $> 100\mu m$ class.

Finally, we evaluate this approach in a zero-shot transfer setting on a completely held-out track section. The model correctly ranks the segments most affected by corrugation, with Spearman's r ranging from 0.32 to 0.73 (mean 0.62 across all time periods considered). Pointwise accuracy on severe cases drops to 10–30%, which should be interpreted also in the context of the class distribution discussed in Section 2.2.

2 Methods

This section first describes the reference pipeline of [8] and the assumptions embedded in its design, then introduces the dataset of Milan's metro network and the adaptations motivated by its more demanding operating conditions. Finally, it describes the two phases of the pipeline (i.e. contrastive pre-training and supervised fine-tuning)

2.1 Reference Unsupervised Contrastive Pipeline

Wang et al. [8] propose a two-stage framework for corrugation detection using car-body vertical acceleration signals collected from a commercial B-type train on Beijing Subway Line 6. In the first stage, raw acceleration signals are segmented into 0.5 s windows and transformed into time-frequency spectrograms using the Synchrosqueezed Wave Packet Transform (SSWPT), a sharpened variant of the continuous wavelet transform that compresses energy along the frequency axis to produce higher-resolution instantaneous frequency estimates [15]. The spectrograms are then used to train a ResNet-18 encoder via Momentum Contrast (MoCo) [16], a contrastive self-supervised framework that learns visual representations without labels by training the encoder to produce similar embeddings (i.e. representations) for two augmented views of the same spectrogram and dissimilar embeddings for views from different spectrograms. In the second stage, the pre-trained encoder is frozen and a linear classification head is fine-tuned on a small set of labelled samples (407 in the original study), supplemented by K-Nearest Neighbours(KNN)-generated pseudo-labels. The binary classification threshold is set at 20 μm peak-to-peak Value (PPV) of the 30–100 mm corrugation wavelength band.

A key assumption enabling this design is operating consistency: both the unlabelled pretraining data and the labelled fine-tuning data are collected under stable conditions on the same line, at a near-constant speed of 51 km/h, with no reported wheel out-of-roundness and with ground-truth labels obtained from a dedicated corrugation analysis trolley (CAT) run on the same track. Generalisation is evaluated on a 1.7 km section of the same network, traversed at the same speed and with the same sensor placement. These conditions limit the confounders between train runs and allow the model to learn corrugation-indicative features without being exposed to significant variations in speed, load, or geometry.

2.2 ATM Metro Dataset

The dataset used in this study consists of axle-box acceleration (ABA) signals collected by four accelerometers mounted on the front and rear axle boxes of both the left and right wheels of a commercial train operating on the M1 line of the Milan metro network. Unlike the setup of [8], no dedicated measurement campaign of accelerations was conducted: the train runs scheduled commercial service multiple times daily, so the recordings span the full range of operational variability: varying speed, passenger load, temperature, and unknown wheel out-of-roundness at any given pass. Signals were acquired at 1kHz and stored as time-domain recordings spatially aligned to set positions.

Three metro sections with heterogeneous geometry are included in the study. These sections are known from historic reads of corrugation levels to develop such defect periodically and therefore provide a realistic distribution of both corrugated and non-corrugated samples. Ground-truth corrugation amplitude was obtained from Mermec measurement trolley campaigns, which produce a continuous PPV profile at 0.3 m spatial resolution for different wavelength bands from 10–30 mm up to 1000-3000 mm. Spatial alignment between vibration recordings and Mermec profiles was achieved via cross-correlation on the curvature information of the signal, and resulted in shifts between 5 to 10 meters depending on the single run and measurement campaign. Following alignment, we apply the Synchrosqueezed Wavelet Transform (SSWT) to all four sensor channels, producing a spectrogram that covers 270 log-spaced wavelength bins spanning the corrugation relevant band of this study (100–300mm).

The dataset covers the period from January to October 2024. A temporal split is adopted: training data are drawn from runs completed on or before 30 June 2024, validation data from 1 July to 30 September 2024, and the test set from runs in October 2024. This split reflects the realistic deployment scenario in which a model trained during one maintenance cycle is applied to future data. However, the resulting class distribution is heavily imbalanced: over the test set, the ratio of non-corrugated (Level I, PPV $< 40\,\mu$m) to corrugated (Level II, PPV $\geq 40\,\mu$m) samples is approximately 55:1.

A further difficulty distinguishing this dataset from the controlled conditions of [8] is the low and variable correlation between raw vibration energy in the corrugation wavelength band and the Mermec PPV label. Wang et al. [8] report Spearman rank correlations between signal energy and ground-truth corrugation amplitude in the range 0.3–0.9 (mean above 0.5) on their single-line, constant-speed dataset. On our dataset, we compute the equivalent correlation using the total spectrogram energy (integrated over all wavelength bins). The results, illustrated in Figure 1, reveal a mean Spearman r of only 0.25 (left rail) and 0.15 (right rail), with values ranging from a maximum of 0.60 down to near zero depending on the period. The high variability is most pronounced for section 18, where the correlation drops from $r = 0.60$ in April 2024 to $r = 0.09$ in July 2024. This instability means that a simple energy-threshold detector would produce highly inconsistent results across maintenance cycles, and motivates the use of a representation-learning approach that can separate corrugation-indicative spectral patterns from confounding vibration sources.

2.3 Contrastive Pretraining

The encoder is pre-trained on the full unlabelled dataset of spectrograms using the MoCo framework [16], following the same contrastive formulation as [8]. Two augmented views of the same spectrogram are treated as a positive pair; views from different windows are treated as negative pairs. Augmentations are spectral-domain operations: random energy scaling, random frequency-bin masking, random spatial-column masking, additive Gaussian noise, and random sensor-channel dropout. This is a different choice from [8], which applies generic image-domain augmentations (brightness/contrast jitter, random cropping, horizontal flipping, black and white filter) to the spectrogram treated as an image.

By contrast, our augmentations each correspond to a plausible perturbation of the physical measurement: energy scaling reflects speed and load variation, frequency

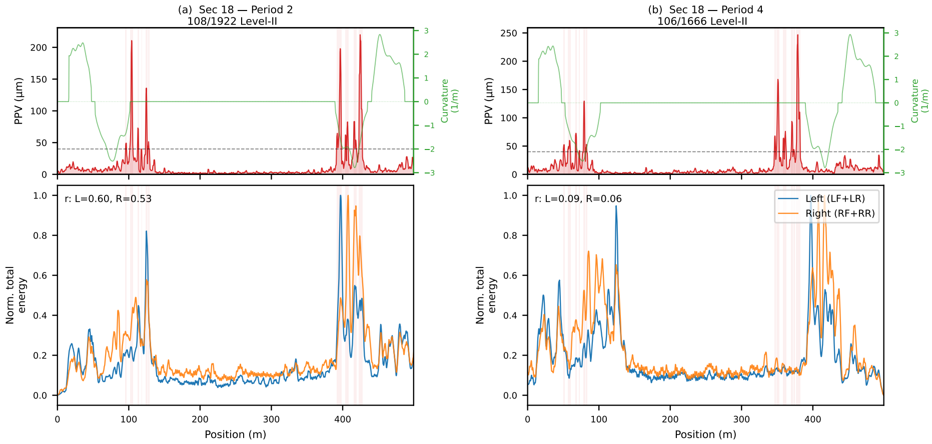

Figure 1: Normalised total spectrogram energy for left and right rail sensors vs. Mermec PPV profile along section 18. (a) April 2024 Mermec campaign: Spearman $r = 0.60$ (left), 0.53 (right); corrugated patches are clearly reflected in the energy signal. (b) July 2024 campaign: $r = 0.09$ (left), 0.06 (right); Shaded regions indicate windows with PPV $\geq 40\,\mu$m.

masking simulates partial loss of a corrugation wavelength component, spatial masking mimics brief signal dropouts, noise reflects sensor and quantisation uncertainty, and channel dropout simulates a degraded or missing accelerometer. The encoder is encouraged to learn representations that are invariant to realistic measurement variability, which we expect to produce more transferable features.

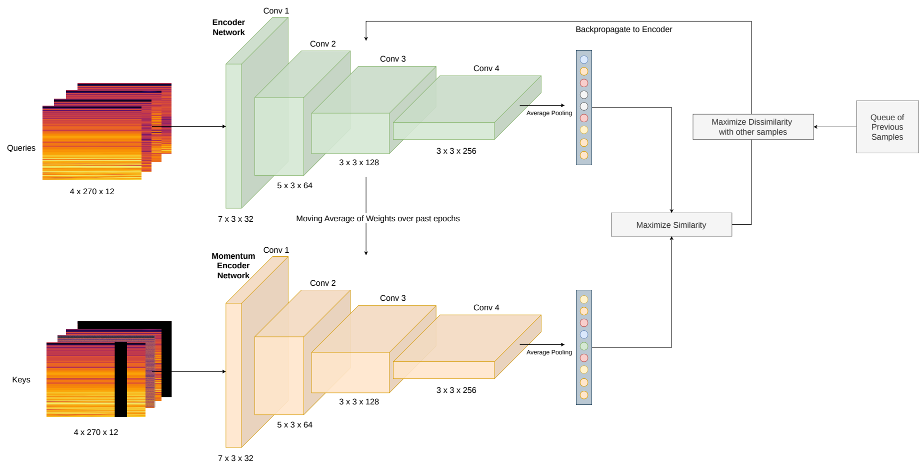

Figure 2: Diagram of the contrastive pre-training pipeline adopted in this work

The encoder backbone is a custom four-layer CNN. Convolutions use larger strides along the wavelength axis than along the spatial axis (See Figure 2) to progressively compress the spectral dimension while preserving spatial resolution; global average pooling over the final feature map produces a 256-dimensional embedding. The network contains no skip connections, making it a deliberately simpler encoder than the

ResNet-18 used in [8].

Pre-training is performed on all unlabelled data from the training period (before 30 June 2024) across all three sections. After pre-training, the encoder weights are frozen for all subsequent steps. The pre-training objective encourages the encoder to learn representations that are invariant to the specific augmentations applied which, in the context of corrugation spectrograms, corresponds to learning features that are robust to minor speed fluctuations and sensor noise rather than memorising the specific vibration signature of individual track locations. This is the central reason the pre-training step is important under our operating conditions: without it, a model trained directly on labelled data would face a much smaller training set and would be more susceptible to the positional overfitting described in Section 1.

2.4 Spatial Sequence Fine-Tuning

The most significant change from [8] is the fine-tuning head and the input formulation. Rather than feeding a single spectrogram window to a linear classifier, we construct spatial sequences of $2W + 1 = 41$ consecutive windows ($W = 20$) centred on the query position, giving the model access to $\pm 6\,\text{m}$ of track context. This design is physically motivated: corrugation develops over characteristic spatial scales and rarely occurs in complete isolation. A single 0.3 m window may be ambiguous, meaning that the local vibration could reflect a wheel-flat, a rail joint, or an isolated geometry irregularity. Observing a larger spatial neighbourhood allows the model to distinguish between a localised defect and a genuine quasi-periodic surface defect.

The architecture operates as follows (see Figure 3): each of the windows is encoded independently by the frozen CNN to produce a 256-dimensional feature vector, which is projected linearly to a $d_{\text{model}} = 64$-dimensional token. In addition, the track curvature and a binary switch-presence indicator are injected at this stage. Curvature is included because the lateral wheel–rail contact forces that cause corrugation are geometry-dependent, and its inclusion allows the model to distinguish corrugation patterns that are characteristic of curved track from those seen on straight sections. Switch presence is a binary flag that marks windows which include a switch or crossing, where the vibration signature is dominated by geometric discontinuities rather than surface defects; including it prevents the model from confusing switch-induced transients with corrugation. The sequence is then processed by a Transformer encoder with two layers and four attention heads, and the final representation is passed to a binary classification head.

The classification threshold is set at $40\,\mu$m PPV, compared to the $20\,\mu$m threshold used in [8]. At $20\,\mu$m the boundary between classes falls in a region of moderate to low corrugation where vibration signatures tend to be more noisy. Finally, during fine-tuning, only the Transformer and classification head parameters are updated; the CNN encoder remains frozen. Cross-entropy loss is used for the classification objective. No KNN pseudo-labelling is applied, in contrast to [8]; instead, given the severe class imbalance (55:1), we use an oversampling strategy to make the problem more

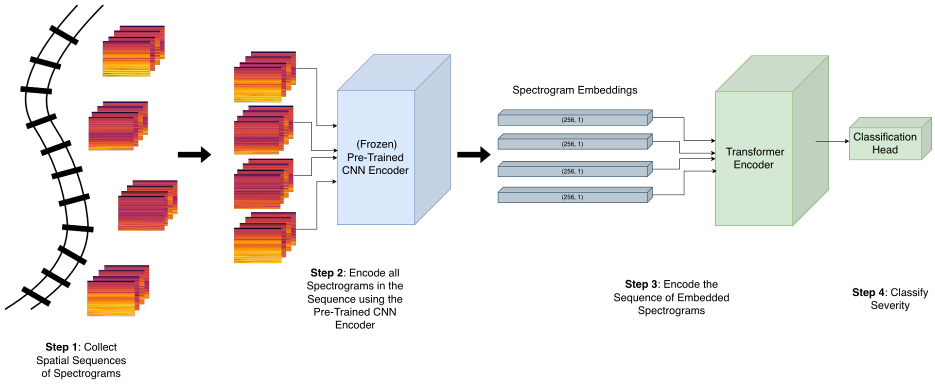

Figure 3: Diagram of the pipeline for the spatial formulation

tractable.

3 Results

The results are presented in four parts. As a baseline, re-applying [8] with a frozen MoCo encoder, linear head, 20 µm threshold to our dataset yields 97.3% recall on severe corrugation ($> 100\,\mu$m) and 81.6% on the 50–100 μm stratum, confirming the core detection capability transfers. The remainder of this section probes its robustness: first an ablation on encoder freezing (§3.1), then the spatial context window (§3.2), then the full pipeline on the October test set (§3.3), and finally zero-shot generalisation (§3.4).

3.1 Positional Overfitting and the Role of the Frozen Encoder

The central motivation for freezing the MoCo-pretrained encoder during fine-tuning is to prevent the classification head from exploiting location-specific spectral signatures that happen to correlate with the label in the training period but do not reflect genuine corrugation features. To probe this risk, a plain CNN classifier (single window, no spatial context) was fine-tuned in three configurations: (i) encoder frozen, with geometry features; (ii) encoder unfrozen, with geometry features; and (iii) encoder unfrozen, without geometry features.

Results are summarised in Table 1. The geometry-free unfrozen model provides the clearest positional overfitting signal: it achieves a near-perfect training recall of 100.0 % on severely corrugated windows ($> 100\,\mu$m PPV) and an AUC of 0.999, yet collapses to 50.7 % recall and AUC of 0.905 on the held-out test period. Without any physical anchor, the encoder adjusts its filters to memorise the vibration character of specific track locations visited during training. This is a pattern that does not transfer once the wheel state, temperature, or load distribution shifts in the test period.

Including geometry features (curvature and switch indicator) when fine-tuning an unfrozen encoder partially mitigates the problem: the train-to-test drop in $> 100\,\mu$m recall narrows from $\approx$50 to $\approx$17 percentage points, and test AUC rises to 0.949. However, unfreezing the encoder can worsen the learned representation for a marginal improvement. For this reason, in the next experiments we chose to keep the encoder frozen, maintaining the pre-training procedure completely task-agnostic.

	Training set		Test set	
Configuration	AUC	Recall$_{>100}$	AUC	Recall$_{>100}$
Frozen, with geometry	0.981	98.4 %	0.904	78.3 %
Unfrozen, with geometry	0.996	99.8 %	0.949	82.4 %
Unfrozen, no geometry	0.999	100.0 %	0.905	50.7 %

Table 1: Train and test performance for frozen vs. non-frozen encoder (single window, no spatial context; $40\,\mu$m threshold). Recall$_{>100}$ is the true-positive rate on windows with ground-truth PPV $> 100\,\mu$m.

3.2 Effect of Spatial Context Window

The Transformer fine-tuning stage aggregates a spatial sequence of $2W + 1$ consecutive windows centred on each query position, giving the model access to $\pm W \times 0.3$ m of track context. To identify the optimal context radius, models were trained with $W \in \{5, 10, 20, 30\}$, corresponding to context radii of 1.5, 3.0, 6.0, and 9.0 m respectively, with all other hyperparameters held fixed.

The results are reported in Table 2. AUC and F1 score for the positive class (Level II, $\geq 40\,\mu$m) both peak at $W = 20$. The gains from $W = 5$ to $W = 20$ are substantial: AUC rises from 0.931 to 0.978 and Level II F1 improves by nearly ten percentage points. Performance degrades at $W = 30$, suggesting that windows beyond ± 6 m introduce more uninformative variation than signal. This is corroborated by the ground-truth PPV profiles: across the three training sections and five Mermec measurement campaigns, 85 corrugated patches were identified, with a median length of 1.8 m, a 95th-percentile length of 4.5 m, and a maximum of 6.9 m. More specifically, 98.8 % of all patches fit entirely within a ± 6 m radius, while 100 % fit within ± 9 m. At $W = 30$ every patch is already enclosed but the three additional metres of context on each side consist almost exclusively of uncorrugated track, which appears to dilute the spatial coherence signal rather than add to it. The $W = 20$ model is used in all subsequent analyses.

3.3 Classification Performance on the Test Set

The best model ($W = 20$, frozen encoder, geometry features) is evaluated on the full three-section October 2024 test set comprising 66 816 windows at a class ratio of

W	Context (m)	AUC	F1-macro	F1-Level II
5	± 1.5	0.931	0.696	0.406
10	± 3.0	0.959	0.724	0.460
20	± 6.0	**0.978**	**0.744**	**0.503**
30	± 9.0	0.969	0.700	0.411

Table 2: Effect of spatial context window W on classification performance (test set, 40 μm threshold). Context is the one-sided radius in metres.

55.3:1 (Level I to Level II). Overall accuracy is 97.27 %, AUC is 0.978, and macro-averaged F1 is 74.4 %. Because raw accuracy is dominated by the majority negative class, Table 3 disaggregates performance by PPV stratum to expose the detection quality across the corrugation severity spectrum.

PPV range (μm)	n	Accuracy	Positive-class rate predicted
< 20	62 972	98.6 %	1.4 % (False positives)
20–50	2 931	74.4 %	28.2 %
50–100	453	69.1 %	69.1 % (recall)
> 100	460	94.4 %	94.3 % (recall)
Overall	66 816	97.3 %	—

Table 3: Stratified test-set performance of the best model ($W = 20$). For the pure-positive strata ($\geq 50\,\mu$m) accuracy equals recall. For $< 20\,\mu$m it equals specificity. The 20–50 μm stratum is mixed (Level I and Level II windows both present).

Several patterns in Table 3 merit attention. At the severe end of the spectrum ($> 100\,\mu$m), the model achieves 94.3 % recall, meaning that nearly all windows with actionable corrugation are correctly flagged; the false-positive rate on healthy track sections ($< 20\,\mu$m) is only 1.4 %. This combination is operationally favourable: the maintenance team would receive very few spurious alerts while recovering almost all genuinely severe sections.

The 50–100 μm stratum is the most challenging, with a recall of 69.1 %. This stratum sits just above the classification threshold and its vibration signature overlaps substantially with that of moderate-severity windows.

The 20–50 μm stratum contains a mix of Level I (20–40 μm) and Level II (40–50 μm) windows; 9.3 % of its entries are truly positive and 28.2 % are predicted positive, indicating that the model is overestimating severity in this range. These boundary-stratum errors are the dominant contributor to the macro-F1 gap (50.3 % for Level II F1 vs. 97.3 % accuracy), and are harder to reduce without a regression formulation.

A regression variant of the same model confirms this pattern: it correctly localises corrugation peaks but predicts conservatively, with a median output of $67.7, \mu$m on the $> 100, \mu$m stratum (Figure 4).

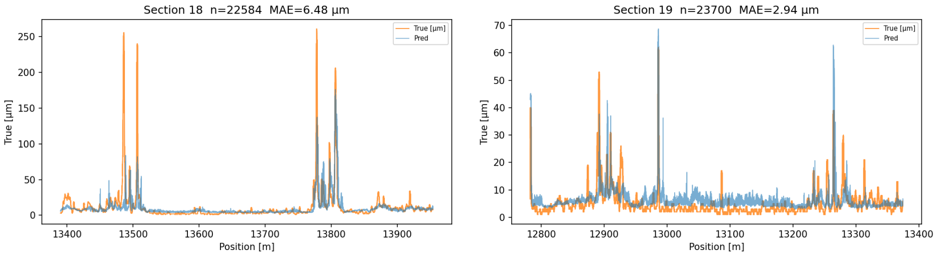

Figure 4: Results of the Spatial Transformer in a Regression setting. The model makes conservative predictions for Section 18 which contains more severe examples of corrugation, while its predictions match more closely for Section 19, where corrugation is milder.

3.4 Zero-Shot Generalisation to Held-Out Sections

To assess whether the model has learned corrugation features that transfer to unseen track geometry, it is evaluated in a zero-shot setting on four different periods of a section that was not included in any phase of training or validation.

Results are presented in Table 4. The aggregate Spearman correlation over all four periods is 0.618, indicating a moderately strong monotone relationship between the model's output and actual corrugation severity despite the model having no knowledge of these locations during training. Performance varies across periods: in April, the model achieves a Spearman correlation of 0.722 and a top-10 % recall of 54.1 %, while runs from October have the lowest correlation (0.447) and a top-10 % recall of 28.0 %.

Period	n	Spearman r	Top-5 %	Top-10 %
April 2024	15 743	0.722	0.414	0.541
June 2024	22 490	0.663	0.189	0.412
August 2024	33 735	0.509	0.243	0.318
October 2024	11 245	0.447	0.229	0.280
Aggregate	83 213	0.618	0.295	0.419

Table 4: Zero-shot rank-based evaluation on four different periods of a held-out section. Top-k % recall is the fraction of the truly most-corrugated k % of windows that the model ranks in its top k %.

The zero-shot results establish that the model's output is not merely a memorised map of previously visited locations but carries information about corrugation severity that is transferable to new geometry. This can be seen even more clearly in Figure 5, where we run the same model in a regression setting on the same track.

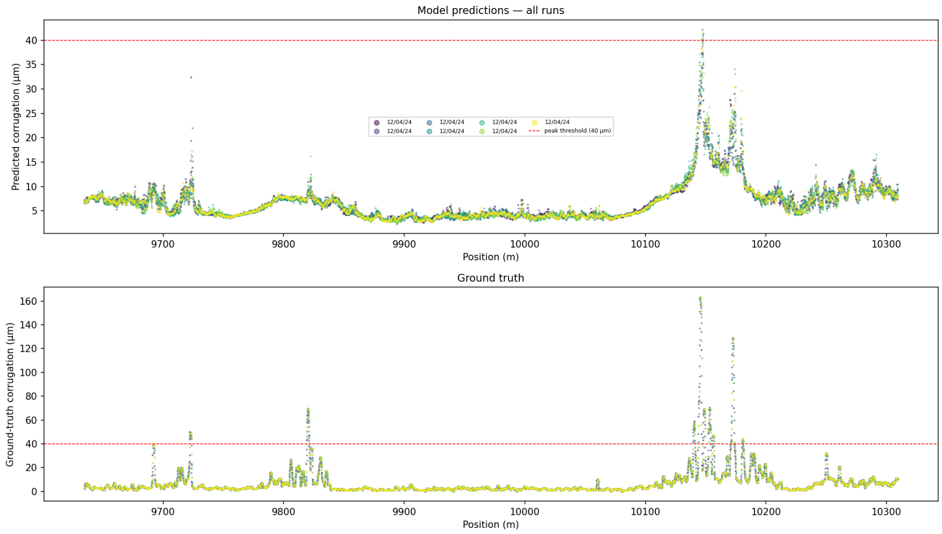

Figure 5: Results of the Zero-shot experiment in a regression setting. The model correctly tracks the regions of space where corrugation is more likely to appear, but it fails in predicting the true severity. Top: Model predictions, Bottom: Ground truth for corrugation

That said, the top-k % recall figures reveal the practical limitation: even in the best-performing section, the model correctly identifies roughly half of the most corrugated 10 % of the track in a zero-shot setting, and aggregate performance is closer to 42 %. This gap between in-distribution classification (AUC 0.978) and zero-shot ranking (aggregate Spearman 0.618) highlights the importance of including at least a small number of labelled windows from any new section before deploying the model operationally, namely a few-shot adaptation step.

4 Conclusions

This paper demonstrates that the two-stage self-supervised framework of Wang et al. [8], consisting of contrastive pretraining followed by supervised fine-tuning, transfers to a substantially harder operating setting: axle-box acceleration signals from a commercial train in scheduled service on the Milan metro network, where speed, passenger load, and wheel out-of-roundness vary continuously across runs. Three adaptations were required to make the transfer viable. First, signals were processed on a fixed spatial grid rather than a fixed time window, ensuring that spectrograms cover identical track lengths regardless of instantaneous speed. Second, physically-grounded spectral augmentations were introduced to replace image-domain transforms. Third, the single-window classification head was replaced by a Transformer that aggregates a spatial sequence of ± 6 m around each query position, augmented with track curvature and switch-presence features.

The results establish three findings of practical relevance. *Positional overfitting* is a genuine risk in this setting: a model that fine-tunes the encoder without geometry features achieves near-perfect recall on severe corrugation during training but recovers only 50.7 % on the temporal test split, because it learns location-specific vibration signatures. Keeping the encoder frozen eliminates this risk while preserving the task-agnostic representations built during pretraining. *Spatial context* is necessary and sufficient at the scale of ± 6 m: this radius encloses 98.8 % of all corrugated patches in the dataset by length, and extending the context further to ± 9 m yields slightly lower performance because the additional windows contain mostly uncorrugated track. *Detection quality* on the held-out test set (October 2024, 66 816 windows, 55:1 class imbalance) reaches AUC 0.978 and a recall of 94.3 % on the most operationally significant stratum ($> 100\,\mu$m PPV), with a false-positive rate of only 1.4 %. Zero-shot evaluation on a completely unseen section yields aggregate Spearman rank correlations of 0.62 between predicted severity and Mermec PPV, confirming that the model has learned corrugation-indicative features.

The primary limitation of this work is the training data, we use: three metro sections, one train type, one ten-month temporal window and labels from the left rail only. Generalisation to sections with substantially different rail profiles remains an open question, as evidenced by the variability in zero-shot Spearman correlations across the held-out section (0.45 to 0.72). Future work should investigate few-shot adaptation, to fine-tune only the Transformer head on a small number of labelled windows from a new section avoiding a full retraining cycle. Extending the evaluation across multiple metro networks and incorporating wheel-condition data are further directions toward a robust, network-agnostic corrugation detection system.

Acknowledgements

The authors would like to thank Azienda Trasporti Milanesi S.p.A. for providing the data for this study.

References

[1] S. Grassie, J. Kalousek, ”Rail Corrugation: Characteristics, Causes and Treatments”, Proceedings of The Institution of Mechanical Engineers Part F-journal of Rail and Rapid Transit - PROC INST MECH ENG F-J RAIL R, 207, 57-68, 1993, DOI:10.1243/PIME_PROC_1993_207_227_02.

[2] D. Zheng, L. Li, S. Zheng, X. Chai, S. Zhao, Q. Tong, J. Wang, L. Guo, ”A Defect Detection Method for Rail Surface and Fasteners Based on Deep Convolutional Neural Network”, Computational Intelligence and Neuroscience, 2021, 2565500, 2021, DOI:https://doi.org/10.1155/2021/2565500

[3] M. Wang, K. Li, X. Zhu, Y. Zhao, ”Detection of Surface Defects on Railway Tracks Based on Deep Learning”, IEEE Access, 10, 126451-126465, 2022, DOI:10.1109/ACCESS.2022.3224594.

[4] M. Di Summa, M.E. Griseta, N. Mosca, C. Patruno, M. Nitti, V. Renò, E. Stella, ”A Review on Deep Learning Techniques for Railway Infrastructure Monitoring”, IEEE Access, 11, 114638-114661, 2023, DOI:10.1109/ACCESS.2023.3309814.

[5] K. Eltouny, M. Gomaa, X. Liang, ”Unsupervised Learning Methods for Data-Driven Vibration-Based Structural Health Monitoring: A Review”, Sensors (Basel, Switzerland), 23, 3290, 2023, DOI:10.3390/s23063290.

[6] X. Wang, Y. Bai, X. Liu, ”Prediction of railroad track geometry change using a hybrid CNN-LSTM spatial-temporal model”, Advanced Engineering Informatics, 58, 102235, 2023, DOI:10.1016/j.aei.2023.102235.

[7] D.C. Larese, A.B. Cerrada, G.D. Tomei, A. Guerrero-López, P.M. Olmos, M.J.G. García, ”Transformer Vibration Forecasting for Advancing Rail Safety and Maintenance 4.0”, 2025, DOI:10.48550/arXiv.2501.11730.

[8] Y. Wang, H. Xiao, M.M. Nadakatti, Z. Zhang, Y. Chi, X. Liu, ”A metro rail corrugation detection framework based on car body vibration signals and unsupervised learning”, Engineering Applications of Artificial Intelligence, 153, 110976, 2025, DOI:10.1016/j.engappai.2025.110976.

[9] M. Haghbin, J. Chiachío, S. Muñoz, J.L. Escalona Franco, A.J. Guillén, A. Crespo Marquez, S. Cantero-Chinchilla, ”Predicting Rail Corrugation Based on Convolutional Neural Networks Using Vehicle's Acceleration Measurements”, Sensors (Basel, Switzerland), 24, 4627, 2024, DOI:10.3390/s24144627.

[10] W. Hassanieh, A. Chehade, A. Facchinetti, M. Carman, M. Bocciolone, C. Somaschini, ”Leveraging machine learning to predict rail corrugation level from axle-box acceleration measurements on commercial vehicles”, International Journal of Rail Transportation, 12, 604-625, 2024, DOI:10.1080/23248378.2023.2220112.

[11] W. Phusakulkajorn, Y. Zeng, Z. Li, A. Núñez, ”Unsupervised Representation Learning for Monitoring Rail Infrastructures With High-Frequency Moving Vibration Sensors”, IEEE Transactions on Intelligent Transportation Systems, 26, 12746-12760, 2025, DOI:10.1109/TITS.2025.3557712.

[12] J. Silva-Rodríguez, P. Salvador, V. Naranjo, R. Insa, ”Supervised contrastive learning-guided prototypes on axle-box accelerations for railway crossing inspections”, Expert Systems with Applications, 207, 117946, 2022, DOI:10.1016/j.eswa.2022.117946.

[13] N. Traquinho, C. Vale, D. Ribeiro, A. Meixedo, P. Montenegro, A. Mosleh, R. Calçada, ”Damage Identification for Railway Tracks Using Onboard Monitoring Systems in In-Service Vehicles and Data Science”, Machines, 11, 981, 2023, DOI:10.3390/machines11100981.

[14] J. Liu, W. Yang, J. He, Z. Wang, L. Jia, C. Zhang, W. Yang, ”Intelligent prediction of rail corrugation evolution trend based on self-attention bidirectional TCN and GRU”, Intelligence & Robotics, 4, 318-38, 2024, DOI:10.20517/ir.2024.20.

[15] H. Yang, ”Synchrosqueezed wave packet transforms and diffeomorphism based spectral analysis for 1D general mode decompositions”, Applied and Computational Harmonic Analysis, 39, 33-66, 2015,

DOI:https://doi.org/10.1016/j.acha.2014.08.004.
[16] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, ”Momentum Contrast for Unsupervised Visual Representation Learning”, 2020.