A Novel Approach to Pavement Crack Classification using Joint-Embedding Predictive Architectures

Soroush Amiri[a], Amir Golroo, Fereidoon Moghadas Nejad[a], Mehdi Rasti[b], Mehdi Monemi[b], Hannaneh Dehghan Tezerjani[a], Jouni Salo[c]

[a] Department of Civil and Environmental Engineering, Amirkabir University of Technology, 424 Hafez Ave., 15875-4413, Tehran, Iran

[b] Centre for Wireless Communications, Faculty of Information Technology and Electrical Engineering, University of Oulu, Erkki Koiso-Kanttilan katu 3, P.O. Box

4500, FI-90014, Oulu, Finland

[C] Unikie, Tammasaarenkatu 3, 00180, Helsinki, Finland

Abstract

One of the important issues in automating infrastructure maintenance is accurate and timely pavement distress detection. In this study, an efficient pavement distress classification model is proposed based on the hypothesis that self-supervised pre-training with the Joint Embedded Predictive Architecture (JEPA) can learn more abstract features as opposed to standard YOLOv11 methods. The proposed model specifically investigates the effectiveness of Global Average Pooling (GAP) in comparison with Class Token (CLS) as the standard approach. The proposed model is tested on standard dataset with high resolutions as well as in the target domain. The experimental evaluations on standard dataset showed the effectiveness of the proposed architecture of JEPA(GAP) with an F1-score of 99.50%, which is considerably better than the standard method in the detection of complex alligator cracks. The proposed model's adaptability is also examined through experimental evaluations involving real-world datasets of 640 weather camera images with large domain gaps. By utilizing a Partial Freezing Strategy in JEPA, the model has shown a vast improvement in terms of F1-score of 73.68% on alligator cracks compared to YOLOv11's 50.00% with an improvement of over 23%. The results validate that JEPA self-supervised learning coupled with effective aggregation of spatial features is far more effective in forming a sound base for generalization and adaptability in noisy real-world settings than the previous approaches in developing intelligent pavement management systems.

Keywords: Pavement Distress Classification, Self-Supervised Learning (SSL), JEPA, YOLO, Supervised Learning

1. Introduction

Pavement condition and performance play a critical role in transportation safety, serviceability, and long-term sustainability. Common distress types such as linear cracking, alligator cracking, potholes, and raveling degrade ride quality and structural integrity, accelerate pavement deterioration, and increase accident risk as well as maintenance costs [25]. Conventional pavement management systems (PMS), which rely largely on manual visual inspections or specialized survey vehicles, are labor-intensive, costly, and subject to operator bias [25, 32]. These limitations have driven a growing shift toward automated, data-driven approaches capable of delivering scalable and consistent pavement condition assessment.

Early attempts at automation were primarily based on unsupervised image analysis techniques, including clustering, thresholding, and rule-based morphological post-processing. While such approaches demonstrated initial feasibility, for example, unsupervised segmentation pipelines based on K-means clustering combined with Otsu thresholding and morphological post-processing achieved moderate precision and recall, they proved sensitive to noise, illumination variation, and surface texture [29]. As a result, their robustness was insufficient for large-scale, real-world deployment.

The rapid development of deep learning has fundamentally transformed automated pavement distress analysis. Supervised convolutional neural networks (CNNs), particularly object detection frameworks such as Faster R-CNN and the YOLO family, have enabled accurate and near real-time detection of pavement defects under complex imaging conditions [8, 23, 33]. Despite recent architectural refinements, many supervised deep learning approaches, including YOLO-based and CNN-based models, still suffer from heavy annotation requirements and limited robustness under diverse imaging conditions [10, 36].

These challenges have motivated increasing interest in SSL for visual representation learning from unlabeled data [7, 18]. More recently, predictive self-supervised frameworks such as JEPA have emerged as promising alternatives to contrastive and reconstruction-based approaches [3]. By learning to predict latent representations across views without explicit pixel reconstruction or negative samples, JEPA offers a potentially more semantic and

label-efficient learning paradigm. While JEPA has demonstrated strong performance on natural image benchmarks, its applicability to pavement imagery and infrastructure condition assessment remains largely unexplored.

The availability of benchmark datasets has further shaped progress in this field. The RDD2022 dataset provided geographically diverse annotations for cracks and potholes, facilitating comparative evaluation of supervised models [1]. More recently, the Attain dataset (2025) expanded the scale and diversity of labeled pavement distress imagery, reporting strong performance of YOLO-based baselines in multi-class settings [34]. Nevertheless, even large-scale datasets continue to exhibit class imbalance, environmental bias, and regional constraints, reinforcing the need for approaches that can effectively leverage unlabeled data [1, 34].

2. General Concept

JEPA represents a predictive latent self-supervised learning paradigm that fundamentally departs from pixel-level reconstruction and contrastive learning approaches [3]. Rather than reconstructing raw image content or enforcing explicit pairwise separation between samples, JEPA learns to predict the latent representation of a target view from a given context view using asymmetric encoders and a dedicated predictor network.

This design is motivated by the insight that high-level semantic understanding does not require explicit pixel reconstruction. By operating entirely in latent space, JEPA encourages the model to focus on abstract, task-relevant features while suppressing sensitivity to low-level appearance variations. In contrast to contrastive frameworks, JEPA eliminates the need for negative samples and reduces dependence on heavy data augmentation, while architectural asymmetry implicitly prevents representation collapse [3].

Recent JEPA variants, including I-JEPA and C-JEPA, have demonstrated competitive performance on large-scale visual benchmarks, producing stable and diverse representations suitable for downstream tasks. Theoretical analysis further suggests that JEPA implicitly incorporates regularization effects comparable to contrastive learning, while remaining computationally efficient [27].

From an application perspective, predictive latent modeling is well aligned with visual inspection tasks characterized by structural patterns and limited pixel-level discriminability, such as pavement crack analysis. Crack patterns are characterized more by their geometric continuity and relational structure

than by precise pixel intensities. Latent prediction enables the model to capture such structural regularities without being constrained by pixel-wise reconstruction accuracy. Moreover, JEPA naturally supports learning from large volumes of unlabeled imagery, significantly reducing reliance on costly manual crack annotation and making it suitable for large-scale infrastructure monitoring scenarios.

2.1. Addressing Spatial Uncertainty with Stochastic Positional Modeling

A key extension of the JEPA framework is the introduction of stochastic positional embeddings, as proposed in Predicting Masked Tokens in Stochastic Locations [5]. Unlike deterministic positional encoding, stochastic positional embeddings JEPA (StoP-JEPA) models spatial uncertainty by sampling token positions during training, thereby relaxing rigid spatial assumptions.

This mechanism is highly relevant for pavement crack imagery. Crack locations are inherently unpredictable, and their spatial configuration may vary significantly even within images captured from the same road segment. By explicitly accounting for positional uncertainty, StoP-JEPA enables the model to learn representations that are more tolerant to spatial variation and partial observability.Furthermore, stochastic positional modeling mitigates the sensitivity of masked prediction to exact token placement, which is a known limitation of traditional masked image modeling approaches. This property aligns well with the irregular and fragmented nature of crack patterns and supports more stable representation learning under real-world conditions.

3. Literature Review

3.1. Non-YOLO supervised deep learning methods for pavement distress analysis

Supervised deep learning approaches dominate automated pavement evaluation and are commonly formulated as image or patch classification, object detection, or semantic segmentation tasks [13, 23, 36]. Two-stage detectors such as Faster R-CNN provide accurate localization and geometry estimation but are often limited by inference latency in real-time applications [23]. In parallel, customized CNN architectures have been proposed for crack classification and segmentation, occasionally outperforming general-purpose detectors under constrained conditions [6]. However, these approaches remain

heavily dependent on labeled data and often exhibit limited robustness under varying data conditions.

3.2. YOLO-based approaches for pavement distress analysis

Single-stage YOLO architectures are widely adopted baselines for pavement distress analysis due to their favorable balance between accuracy and computational efficiency [15, 26]. Early applications of YOLOv3 and YOLOv4 demonstrated performance comparable to two-stage detectors with significantly reduced inference time [15, 26]. Subsequent variants introduced architectural refinements to address domain-specific challenges. Models such as YOLOv7-RDD reduce inference time while maintaining competitive accuracy [30], whereas YOLO-RAPD improves detection through multilayer feature fusion and enhanced feature extraction, and YOLO-RD increases sensitivity to small-scale and fine cracks via wavelet transform convolution [39, 40].

Despite these refinements, YOLO-based methods continue to rely heavily on large annotated datasets and remain sensitive to variations in pavement appearance and imaging conditions [10, 42]. Although recent variants have improved robustness under challenging scenarios such as UAV-based inspection and adverse weather [12, 31, 38, 41, 43], annotation dependency and consistent performance across diverse data conditions remain open challenges.

3.3. Self-supervised learning in computer vision and pavement evaluation

Self-supervised learning seeks to learn meaningful representations from unlabeled data by defining surrogate training objectives [16]. In computer vision, SSL has been shown to reduce annotation requirements while improving robustness and transferability across tasks and acquisition settings [16, 35]. In the context of pavement engineering, SSL-based approaches have demonstrated improved crack segmentation performance and enhanced resilience to illumination variations and surface heterogeneity [9, 35].

Contrastive learning methods such as SimCLR and MoCo maximize agreement between different augmented views of the same image while enforcing separation from other samples [7, 18]. Although effective, these methods typically require large batch sizes or memory banks, increasing computational cost. Alternative strategies, including momentum encoders and non-contrastive approaches, have alleviated some of these constraints; however, they still incur substantial computational and training overhead, which can limit their practicality for large-scale infrastructure monitoring [14, 20, 22].

Masked image modeling (MIM) approaches, including MAE and its extensions, reconstruct masked image regions to learn transferable features [4, 17, 44]. While effective in capturing local structure, MIM-based methods may be less discriminative at the semantic level, motivating hybrid approaches that combine reconstruction and contrastive objectives [2, 19].

3.4. Joint Embedding Predictive Architecture

JEPA constitutes a class of predictive self-supervised methods that have recently gained attention as alternatives to contrastive and reconstruction-based frameworks [3, 28]. Initial studies demonstrated that JEPA can achieve competitive representation quality on large-scale vision benchmarks while reducing reliance on negative samples and heavy data augmentation. Subsequent work, including I-JEPA and C-JEPA, further analyzed the empirical performance and theoretical properties of predictive latent modeling, establishing connections to contrastive regularization [3, 27].

Subsequent variants have extended JEPA to convolutional backbones and non-image modalities, including CNN-JEPA and T-JEPA [21, 37]. The introduction of StoP-JEPA further enhanced JEPA-based masked modeling by addressing spatial uncertainty in visual data [5]. Although JEPA has shown promise in non-natural image domains such as SAR imagery [24], its potential for pavement condition assessment has not yet been systematically investigated.

Despite progress in supervised pavement distress analysis, existing approaches face high annotation costs and limited robustness to variations in pavement appearance and acquisition conditions. While SSL methods have begun to address these challenges, most prior work has focused on contrastive or reconstruction-based paradigms, and predictive latent models such as JEPA remain largely unexplored in pavement-related applications.

4. Objective and Scope

Pavement crack analysis is formulated in this study as an image-level classification task. The scope encompasses not only performance evaluation on standard high-resolution datasets but extends critically to assessing domain adaptability in real-world surveillance scenarios. Specifically, the study targets the challenging domain of fixed weather camera imagery, which introduces significant domain gaps characterized by low resolution, heterogeneous viewpoints, and environmental noise.

Motivated by the limitations of standard supervised models to generalize in such noisy environments, this work explores the feasibility of Joint Embedded Predictive Architecture for pavement distress classification. A core objective is, therefore, to find the best feature aggregation mechanism that serves this particular task; hence, this study will systematically investigate the effectiveness of GAP against the traditional Class Token approach. Performance is stringently benchmarked against a state-of-the-art YOLOv11 supervised baseline to quantify gains in robustness and transferability.
The main contributions are summarized as follows:

- Development of a pavement distress classification framework leveraging a self-supervised JEPA-based encoder.
- A systematic investigation into feature aggregation strategies within the JEPA framework, proposing an enhanced architecture capable of capturing intricate structural dependencies in pavement imagery more effectively than standard token-based approaches.
- Development of an appropriate benchmarking protocol that assesses the applicability of the learned features in the target domains, as well as realizing the fine-tuning strategy for handling domain transition in realistic surveillance conditions.
- A comprehensive comparative analysis quantifying the superior robustness of proposed model against state-of-the-art completely supervised baseline (YOLOv11), specifically in scenarios characterizing low resolution and high environmental noise.

5. Methodology

In this section, the proposed methodological paradigm for pavement distress image classification will be outlined. Notably, the proposed strategy establishes itself on the basis of a self-supervised paradigm which relies on the application of JEPA methodology for the extraction of semantic cues on pavement distress images. This paradigm will follow clearly defined stages, beginning initially with the detailing of the proposed architectural design, taking particular notice of an ablation experiment analyzing the applicability of either GAP or CLS approaches for semantic feature extraction as applicable for tasks revolving around texture classification. Secondly, it will

elaborate on the complete training procedure, as well as the implementation of a partial freezing strategy to adapt the pre-trained encoder to the target domain through supervised downstream training.

5.1. Data preparation and preprocessing

The data used in this study consist of 2000 high dimension images with dimensions of 1868×4000 pixels divided equally into two sets, 1000 images for linear cracks and 1000 images for alligator cracks. of this number of data for each group, 800 were used for training, 100 for validation, and 100 for testing. The images were all taken under standard conditions at the same angle and distance for minimizing variability due to perspective and scale(standard dataset). Also, in the preprocessing, the size of the input images was changed to the standard size of the JEPA model, which is 224×224, and the pixel values were normalized using the mean and standard deviation of the ImageNet dataset. A second dataset was also obtained for domain adaptation experiments from roadside weather cameras. Differently from the common dataset, these wide-angle photos comprised noticeable environmental noises and non-pavement regions. To counter this challenge, a cropping step was applied to deliberately identify and remove regions of great irrelevance to pavement classification namely, background regions outside road surfaces. After extracting these photos, they were also subjected to similar processing steps as those for the standard dataset rescaling to a resolution of 224×224 pixels and image normalization.

5.2. JEPA model architecture and pre-training process

Joint-Embedding Predictive Architecture captures high-level semantic features of pavement distress without requiring labor-intensive pixel-level annotations. We integrated the StoP-JEPA variant into our framework. Standard JEPA architectures assume deterministic spatial relationships. This assumption does not hold for pavement cracks. Irregular boundaries and ambiguous edges introduce inherent spatial uncertainty. StoP-JEPA addresses this limitation by introducing stochasticity into positional embeddings. The model learns robust structural representations. It avoids overfitting to precise high-frequency spatial details.

5.2.1. JEPA Architecture: Components and Pretraining Objective

Figure 1 shows the JEPA framework which will provide the encoder component in our proposed pipeline, as will be explained later.It has three main

parts, a Context Encoder f_θ, a Target Encoder $f'_{\theta'}$, and a Predictor g_ϕ. First, the input image gets divided into non-overlapping patches. A masking strategy picks a context block x (visible patches) and a target block y (patches whose embeddings are to be predicted). The context encoder f_θ produces a latent representation from the context ($s_x = f_\theta(x)$). The target encoder $f_{\theta'}$ does the same for the target view ($s_y = f_{\theta'}(y)$). A known problem is representation collapse, where encoders output trivial, constant features. To stop this, the target encoder's weights θ' aren't trained with gradient descent. They're updated using an Exponential Moving Average (EMA) of the context encoder's weights θ:

$$\theta' \leftarrow \tau\theta' + (1-\tau)\theta \tag{1}$$

where $\tau \in [0, 1)$ is a momentum coefficient, which usually increases on a schedule during training.In a conventional I-JEPA, the predictor is deterministic. It takes the context representation s_x and a target position mask p_y to predict the target embedding $\hat{s}_y = g_\phi(s_x, p_y)$. The objective is to minimize the prediction error in the latent space without reconstructing raw pixels. The standard loss function measures the distance between the predicted and actual target representations:

$$\mathcal{L}_{std} = \frac{1}{M}\sum_{i=1}^{M} \|g_\phi(s_x[i], p_y[i]) - s_y[i]\|_2^2 \tag{2}$$

where M is the batch size.

While the standard deterministic approach is effective for semantic segmentation of large objects, it faces challenges when the masked region content involves high uncertainty, such as the exact trajectory of a fine pavement crack. To address this, we integrate a stochastic predictor (StoP-JEPA) into the framework. Unlike the standard fixed positional embeddings, our predictor g_ϕ utilizes stochastic positional embeddings where the masked patch position is modeled as a distribution. This allows the network to predict the target representation based on both the context s_x and a stochastic latent variable z:

$$\hat{s}_y = g_\phi(s_x, z) \tag{3}$$

The stochastic nature of this encourages the model to learn robust, high-level semantic features, rather than overfitting particular high-frequency details. Our particular pre-training objective thus minimizes distance in this stochastic space. Furthermore, to enhance robustness against outliers in pavement

data, we employ the Smooth L_1 Loss instead of the traditional L_2 distance. The final loss function is defined as:

$$\mathcal{L} = \frac{1}{M}\sum_{i=1}^{M} \text{Smooth}_{L1}(\hat{s}_y[i], s_y[i]) = \frac{1}{M}\sum_{i=1}^{M} \text{Smooth}_{L1}(g_\phi(s_x[i]; z[i]), s_y[i]) \tag{4}$$

Minimizing this objective enables the context encoder and predictor to effectively capture the structural semantics of pavement distress.

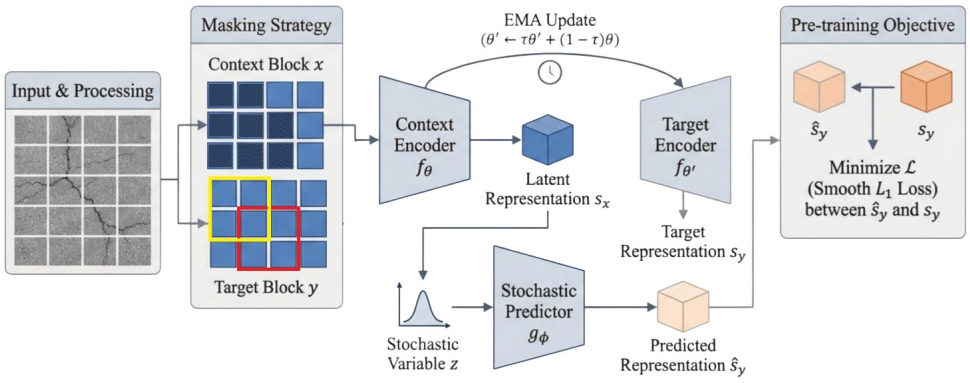

Figure 1: Schematic illustration of the proposed self-supervised pre-training framework integrating StoP-JEPA.

5.2.2. Implementation Details

The proposed model has been developed using the PyTorch library. The input images were resized to 224×224 and divided into 16×16 tokens. The multi-block masking method was used for the selection of the context and target blocks. The model has been trained using the AdamW optimizer. The weight decay regularization has been applied during the training. The learning rate has been implemented according to the WarmupCosineDecay method. The learning rate was increased linearly during the warmup phase. Then the cosine decay function has been applied. In order to prevent the problem of gradient explosion and promote smooth convergence of the training, the gradient clipping function was used for the backpropagation of the

encoder parameters. In addition, to stabilize the update of the target encoder, the momentum coefficient (τ) for the EMA update rule was adopted to follow the linear schedule from 0.996 to 1.0. The pre-training stage involved 100 epochs with a batch size of 64. Lastly, the obtained context encoder f_θ was fixed and used as the feature extractor for the classification task.

5.3. Feature Aggregation and Classification Head Architecture

In particular, we have adapted the pre-trained JEPA encoder to the specific task of pavement crack classification by introducing a dedicated classification head: with it, the latent representations obtained by the encoder are mapped to the target class probabilities. The architecture of this downstream adaptation module consists of the following keystone elements: a feature aggregation mechanism and a projection head.

5.3.1. Feature Aggregation Strategies

For finding the appropriate mechanism to transfer the learned features from the JEPA encoder to the downstream task of classification, we implemented two separate techniques for aggregating the extracted features from the sequence of embedding patches generated by the JEPA context encoder into a global representation for classification. The architectural overview and the distinction between these two adaptation approaches are visually illustrated in Figure 2.

Strategy I: Class Token (`[CLS]`) Adaptation.

The first strategy involves the standard Vision Transformer fine-tuning process. In this scenario, we add a learnable [CLS] token to the input sequence in the encoder. The purpose of this token, that interacts with the embeds of patches by self-attention layers, is to model the global information. During the fine-tuning process, the final hidden representation of this token from the last layer L ($z_{\text{CLS}}^{(L)}$) is used for classification:

$$y = W \cdot \text{BatchNorm}(z_{\text{CLS}}^{(L)}) + b \tag{5}$$

Since this token is introduced only during the fine-tuning stage (as it is absent in JEPA pre-training), this strategy places an additional burden on the model to establish the attention-based aggregation mechanism from scratch.

Strategy II: Global Average Pooling (GAP).

The second strategy utilizes the spatial feature maps obtained from the pre-trained backbone. Pavement distress features, such as alligator cracking, usually distribute across the entire image instead of concentrating in limited areas. Global average pooling (GAP) is applied for this reason. The global representation forms through computation of the arithmetic mean over all patch embeddings from the final encoder layer:

$$z_{\text{global}} = \frac{1}{N} \sum_{i=1}^{N} z_i^{(L)} \tag{6}$$

The vector proceeds through batch normalization and a linear projection layer. This method depends on the pre-trained patch embeddings. Initialization of additional tokens becomes unnecessary.

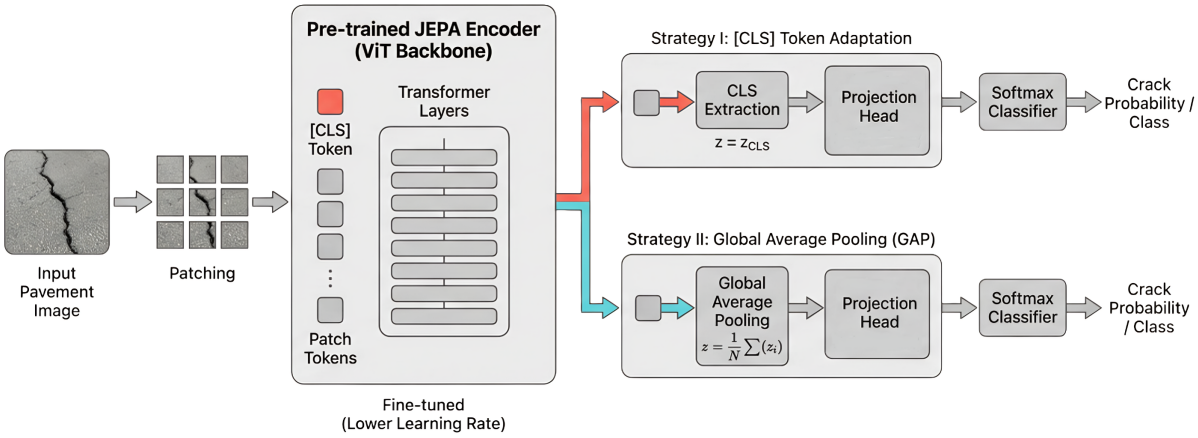

Figure 2: Overview of the [CLS] token and GAP strategies for downstream classification.

5.3.2. Fine-tuning and Optimization Strategy

Training high-capacity Transformer-based models calls for unique optimization approaches based on the source and target domains. In our implementation, the task required a two-step optimization approach in validating the encoder's ability in the source domain first, followed by its adaptation to the road weather cameras target domain.

Phase I: Validation on Clean Source Data (Two-Stage Fine-tuning).

First, to preliminarily verify the quality of the representations learned by the JEPA backbone, we fine-tuned the model on the classification task using the same dataset as that used during the pre-training. As these images had clear features according to the pre-training distribution, we adopted a two-stage optimization Strategy with high rigor over 100 epochs to strive for the best performance. In stage one (Epochs 1–15), the encoder was frozen to serve as a fixed feature extractor. Only the classification head was trained with a learning rate of $\eta = 1 \times 10^{-3}$ to allow the classifier boundaries to adapt to the feature space without degrading the backbone weights. In second stage (Epochs 16–100), the encoder was unfrozen, and the entire model was fine-tuned in an end-to-end manner. A discriminative learning rate was applied, where the encoder was updated with a lower rate ($\eta = 1 \times 10^{-5}$) to preserve pre-trained knowledge, while the head continued to refine using the labeled data.

Phase II: Adaptation to New Domain Data (Partial Freezing Strategy).

To assess the adaptability of the model to the real-world application of infrastructure monitoring, we moved on to the target domain, which was a set of images taken by road weather cameras[11]. The goal was to adapt the pre-trained model to predict pavement conditions in the new domain, which is quite different from the source domain. The set of data specifically tailored for this task was divided into 448 samples for training, 96 samples for validation, and 96 samples for testing. Due to the small size of the dataset and the large domain shift, the Fine-tuning Strategy, which was used in the previous phase, was likely to suffer from overfitting and catastrophic forgetting.This problem was remedied through the implementation of the Partial Freezing Strategy. We conjecture that early Vision Transformer layers focus on domain-independent geometric information such as edges and gradient directions that are domain transferable across road weather cameras images, and that higher layers have learned high-level semantic representations that need adaptation.

On this assumption, we froze the first 10 blocks of the ViT encoder. This made it mandatory for the model to utilize the robust structural information that was extracted during JEPA pre-training. As such, only the last 2 blocks of the ViT encoder, along with the projection head, were made trainable.

For optimization, we utilized the AdamW optimizer with weight decay at 0.1. To ensure equilibrium in model adaptation, we utilized differential learning rates. A conservative value of 5×10^{-5} was used for training the trainable layers in the ViT encoder to ensure that existing knowledge was not lost. On the other hand, an aggressive value of 1×10^{-3} was utilized to train the classification head. The training of the model was intended to be performed to a maximum of 35 epochs. A checkpoint solution was implemented in the model to save the weights of the maximum validation accuracy.

5.4. Benchmarking with YOLO

To analyze the efficiency of the proposed solution against the existing best solutions available today, the performance of the end model was compared to that of a standard YOLO model. To do this, the actual labels were assigned to all images from the dataset using the Roboflow platform. Then the YOLO11 Medium classification model (YOLO11m-cls) was used to train on the same dataset for 100 epochs to notice the outcome of the two solutions.

5.5. Model Performance evaluation criteria

Accordingly, precision, recall, and the F-1 score were given for each class separately in this study, providing a detailed view of the performance of each model. Precision estimates the exactitude in the positive predictions, and it refers to the ratio of true positive samples to the total number of positive predictions. It is calculated as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

Where TP, FP, and FN represent the number of True Positives, False Positives, and False Negatives, respectively.

Recall, or sensitivity, refers to the model's ability to identify all relevant instances within a dataset. It is defined as the ratio of true positive samples to the total number of actual positive samples :

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

The F1-score is the harmonic mean of precision and recall, providing a single metric that balances both concerns, especially in cases of uneven class distribution. It is calculated using the following equation:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

5.6. Model Adaptability to Real-World Surveillance Data

To evaluate the performance of the model on data close to reality and not standard photos taken with the same distance and resolution, the existing data from road weather cameras in Finland was used. About 640 pavement crack images were collected, including 320 linear cracks and 320 alligator cracks. These images are a subset of the weather cameras in Finland, which are specific to roads. Therefore, the specific pavement distress areas in these cameras were cropped uniformly to remove cars, bridges, and other irrelevant items in these images and to focus on the pavement cracks. All images were prepared with a resolution of 240×240 and were used for training and testing in the jepa and yolo models.

6. Results and Analysis

6.1. Training Stability and Convergence Analysis

Prior to describing the quantitive metrics of the classification accuracy, there is a need to examine the training dynamics to ensure the validity of the training phase. The training and validation loss for the baseline YOLOv11 model can be seen in Figure 3, while the training dynamics for the proposed JEPA-based architecture can be observed from Figure 4.

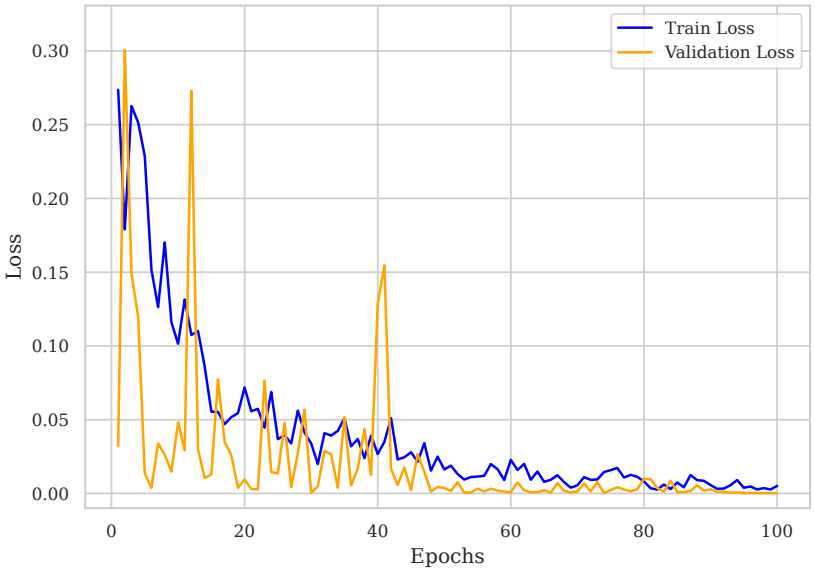

Figure 3: Training dynamics of the YOLOv11 baseline.

It is seen from Figure 3 that the YOLOv11 baseline has noticeable volatility or high-frequency oscillation. Although Model Checkpointing mechanisms usually select the best model regarding the optimal metric, such as highest accuracy or lowest loss, this stochastic behavior creates concerns that the selected checkpoint could correspond to a transient spike rather than a stable optimum. For example, high accuracy in one epoch could be followed by a significant drop in the next, as depicted by the fluctuations. As a result, the model's reliability on unseen data may turn out to be lower compared to some model with a smoother trajectory. This rugged optimization landscape notwithstanding, the general trend confirms that, eventually, the model learned generalizable features without suffering from overfitting.

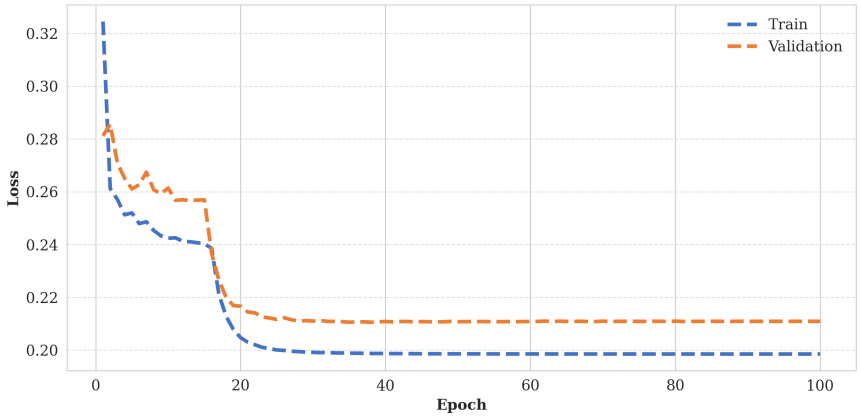

Figure 4: Training dynamics of the JEPA model.

Contrary to this, the JEPA model (Figure 4) exhibits a very smooth and monotonic convergence behavior. One of the very interesting aspects of this convergence behavior is the sharp fall in the loss function in the initial epochs. This suggests that the transfer learning efficiency is very high, implying that the self-supervised pre-training has already identified a feature space very close to the target task, thus reducing any semantic gaps. Moreover, since the training and validation losses remain very close, constant in value, and very small in magnitude, it is very evident that the JEPA model is also very insensitive to the data. Moreover, in contrast to the baseline method, the JEPA method exhibits a very smooth convergence behavior, ensuring that the final performance represents a deterministic training result.

The root cause of the observed disparity in training stability can be traced to the gap in training paradigms. First, the YOLOv11 model is trained in an end-to-end supervised manner, in which the backbone has to learn low-level feature representations and high-level class discriminators simultaneously. This usually leads to conflicts in gradient updates and eventual optimization instability while handling complex crack patterns. In contrast, the proposed JEPA-based approach enjoys a two-stage learning strategy. First, the encoder is pre-trained in a self-supervised learning manner to capture the intrinsic morphology of the pavement distresses without label bias. As such, the downstream classification task starts with a strong, domain-specific feature extractor. It thus turns the optimization problem from a chaotic search to a smooth fine-tuning trajectory.

To further validate these observations regarding optimization stability,

Figure 5 shows the evolution of validation accuracy for both architectures. Figure 5a presents the YOLOv11 model, which, while achieving high levels of accuracy, peaking near 99.5%, is seen to have a highly oscillatory pattern throughout its training process. This verifies the hypothesis made earlier that the supervised baseline is sensitive to batch-wise variations and could be non-robust, leading to potentially inconsistent predictions on the halt of training at an unfavorable epoch.

On the other hand, the accuracy curve of the model based on the JEPA algorithm in Figure 5b illustrates steadfast stability. The graph starts with a steep climb towards an accuracy close to perfection in a few iterations and then sustains a stable, non-vibrational plateau for the remaining period of the training phase. The lack of volatility in the graph of the validation metric provides substantial evidence for the efficiency of the self-supervised pre-training procedure. This reveals the fact that the JEPA encoder has been capable of learning consistent features, unaffected by the noise of the input data, with predictable performance in classification.

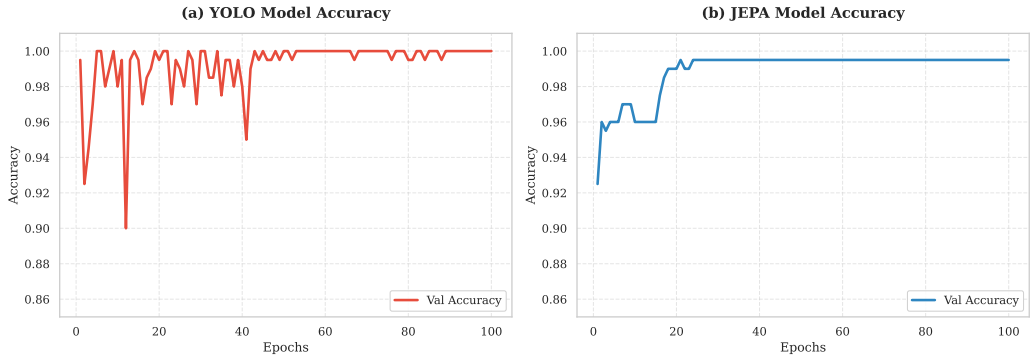

Figure 5: Validation accuracy progression over epochs for (a) the supervised YOLO baseline and (b) the proposed JEPA architecture.

6.2. Performance Comparison of YOLO and JEPA

In this section we compare our JEPA based proposed method to YOLOv11 on the standard dataset. Proposed model is tested in using the standard classification head (JEPA(cls))and the GAP technique (JEPA(GAP)). The performance is then measured in terms of precision, recall, and the F1-score in the context of linear and alligator cracks.

Table 1 Comparative Performance Results of Models for the Linear Crack Class

Linear					
YOLO		JEPA(cls)		JEPA(GAP)	
Precision	Recall	Precision	Recall	Precision	Recall
97.06	99.00	100.00	98.00	99.01	100.00
f1=98.02		f1=98.99		f1=99.50	

Table 2 Comparative Performance Results of Models for the Alligator Crack Class

Alligator					
YOLO		JEPA(cls)		JEPA(GAP)	
Precision	Recall	Precision	Recall	Precision	Recall
98.98	97.00	98.00	100.00	100.00	99.00
f1=97.98		f1=98.99		f1=99.50	

Table 1 illustrates the results for the Linear crack category. It is seen that all models exhibit high performance. YOLO performs outstandingly well with an excellent value for the F1 score of 98.02% and an excellent value for the recall of 99.00%. However, the proposed models outperform the baseline. JEPA(cls) has a perfect precision of 100.00%, meaning there are zero false positives, albeit at the cost of the model's recall of 98.00%. JEPA(GAP) is the most well-rounded and best performer that yielded an F1 score of 99.50% with excellent precision of 99.01% and an excellent recall of 100.00%. Moving on to the Alligator crack class in Table 2, the superiority of the proposed method is further emphasized. Alligator cracks are known to have a highly complex pattern of irregular shapes, whose detection is much more challenging than the basic linear shapes in the simple crack class. Although the YOLO model performs very well with an F1-score of 97.98%, the models based on the JEPA are exceptionally stable handling the complexi- ties. The JEPA(cls) model performs flawlessly with a perfect recall of 100.00%, making sure that not a single complex pattern is missed. Most importantly, the JEPA(GAP) model obtains a flawless F1-score of 99.50% along with a perfect precision of 100.00%. It is observed that the GAP

mechanism correctly combines the spatial features from the JEPA encoder to differentiate the complex road distress patterns from the background perfectly without any false alarm. The experiments have asserted that, although YOLO is a competent candidate, the proposed JEPA(GAP) architecture is a more authentic and precise solution for the complex road distress detection problems. Thus, for the implementation of the intelligent Pavement Management System, the JEPA(GAP) model proves to be the best choice. Its highest precision of 100.00% to wards the detection of the alligator cracks is most useful for the budget allocation process, as it avoids the allocation of precious structural repair materials towards false positives. On the other hand, the highest value of recall of 100.00% towards the detection of the linear cracks is most useful for the maintenance process, as it avoids the occurrence of the structural damages by sealing the detected cracks at their initial stages.

Figure 6 presents the comparative performance of the YOLO, JEPA(cls), and JEPA(GAP) models across both Linear and Alligator crack classes.

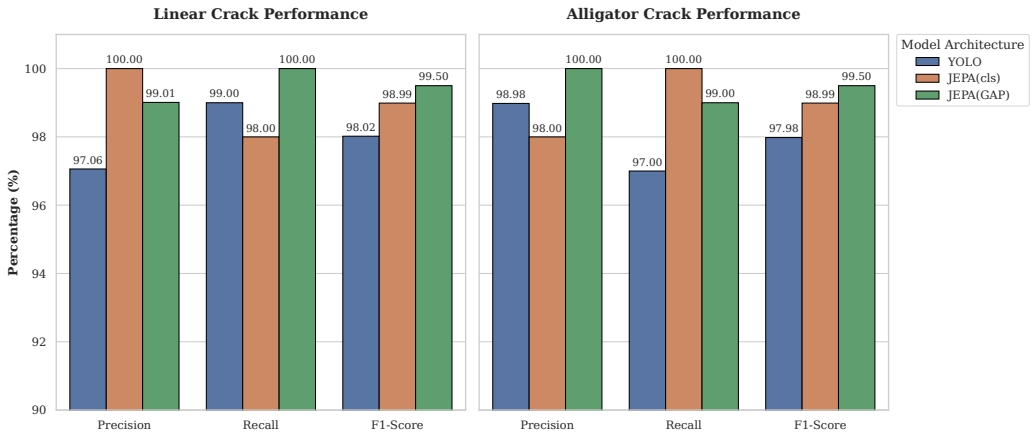

Figure 6 Comparative performance metrics of the YOLO, JEPA(cls), and JEPA(GAP) models across both Linear and Alligator crack classes.

6.3. Qualitative Analysis and Visualization

Although the quantitative results shown in the preceding section demonstrate the effectiveness of the proposed model in pavement distress classification, it is important to emphasize that a complete understanding of deep neural networks' decision-making processes, described as black boxes, requires a

more in-depth study. High model accuracy does not always translate to effective feature-learning, since the model could potentially learn features from either crack patterns or background noise. Therefore, this section aims to provide a qualitative evaluation of the features learned from the pre-trained encoder (f_θ), trying to validate that a meaningful and well-structured feature space has, in fact, been engineered by employing the self-supervised JEPA technique. To do so, a diverse evaluation of the latent space along multiple aspects evaluating the intrinsic separability of data clusters through t-SNE evaluation, ensuring that the network omits unwanted region activation through feature activation maps, and evaluating geometric preservation through image reconstruction analysis will be pursued. Collectively, these visual evaluations demystify the model's internal representations, confirming its interpretability and reliability for real-world deployment.

6.3.1. Latent Space Visualization (t-SNE)

For the qualitative analysis of the learned representations via the pre-trained encoder (f_θ) and the data distribution in the feature space, the t-Distributed Stochastic Neighbor Embedding technique was used. The t-SNE algorithm is a non-linear manifold learning algorithm used for the reduction of the data dimensionality, which has applications in data visualization. This algorithm maintains the local properties of the data, as opposed to linear techniques like Principle Component Analysis (PCA); it does this by ensuring data points which are close together in the original high-dimensional feature space have a high likelihood of being close together in the lower-dimensional feature space, which has two dimensions. This process involves the transformation Euclidean distances amongst data points to conditional probabilities, which are further minimized based on the difference in the KL divergence amongst the probability distributions of the lower-dimensional and the original high-dimensional data spaces.To create the feature visualization map using the t-SNE algorithm, the feature vectors extracted via the last layer of the ViT-Base model were processed by the algorithm after the GAP process on the patch representations. It is important to note that this feature extraction process was conducted entirely without the influence of class labels.

Figure 7 indicates how test sample data is scattered in the two-dimensional feature space, with sample data points color-coded with corresponding ground-truth types of cracks. From the analysis of the diagram, high separability among classes is observed since Linear and Alligator crack samples distribute

in two distinct regions with significant separation by an appropriate data-sparse margin. This clear separation indicates that the JEPA model is capable of learning an optimal discriminative representation for each distress type, given that clear separation among classes under minimal data is achievable. On the other hand, the diagram indicates an appropriate semantic clustering achievement since the points corresponding to the different crack types have been self-organized into specific regions in the feature space without employing manual labels during the pre-training phase. Moreover, the high compactness obtained among each group of points indicates that the model considers different points with the same crack type but with negligible orientation or width variations as similar semantic high-level features.

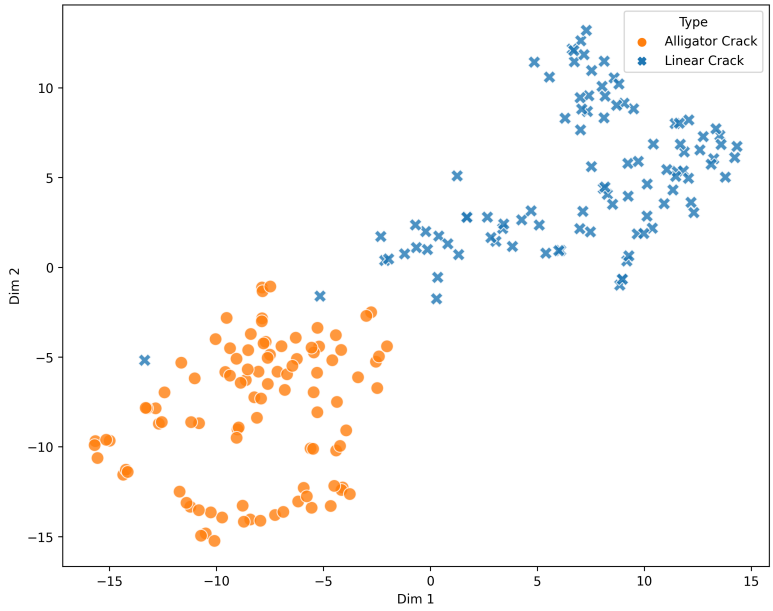

Figure 7 t-SNE visualization of the learned feature space.

6.3.2. Feature Activation Analysis

To better comprehend the concept of space attention focus embodied in the pre-trained encoder $f\theta$, as well as to make sure that the information about physical distress is not encoded in irrelevant details of the background, Feature Activation Maps were created. These maps were produced by calculating the L2-norm value for each patch in the final Transformer layer's output feature vector, representing the magnitude of neural activity at each

spatial location. Figure 8 presents a qualitative comparison comprising the original image, the activation map, and their overlay.A very interesting observation is the distinct feature contrast between the pavement surface and the distress trajectory. The encoder exhibits high feature magnitude (indicated by warmer colors) on the aggregate-rich asphalt texture, while the crack formations are consistently represented as regions of significantly lower activation. This suggests that the model effectively distinguishes distresses as structural voids or anomalies within the continuous pavement texture. It is quite impressive to note that this precise localization is achieved without seeing pixel segmentation masks at the training stage. The fact that the encoder can implicitly localize the distress through this negative contrast confirms that it has learned to prioritize semantic structural information, thereby establishing a robust foundation for the downstream classification task.

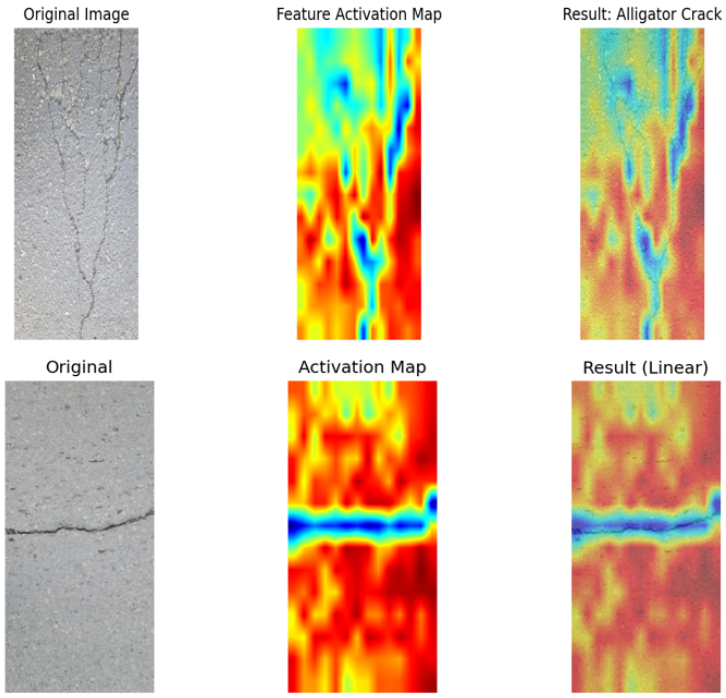

Figure 8 Qualitative visualization of Feature Activation Maps.

6.3.3. Feature Reconstruction Analysis

To assess the fidelity and semantic content of the learned feature representations, a lightweight decoder network, implemented as a Reconstruction Head, was trained for reconstructing the original input images using nothing but the frozen feature representations from the JEPA encoder. The decoder network achieved learned results within 10 epochs, demonstrating satisfactory results, thus indicating that the spatial and structural information was well-organized in the latent feature space of the JEPA encoder.Figure 9 exemplifies how, despite a short training period and a simple structure of the decoder network, basic characteristics and complex geometric patterns of cracks are preserved. Moreover, it can be observed that the model efficiently eliminates high-frequency textures of the environment, like detailed aggregates of asphalt, while maintaining characteristics of distress patterns. It is evident that there is a selective reconstruction, indicating that the JEPA model has learned to distinguish between feature signals and environment noise, thus performing effective semantic compression.

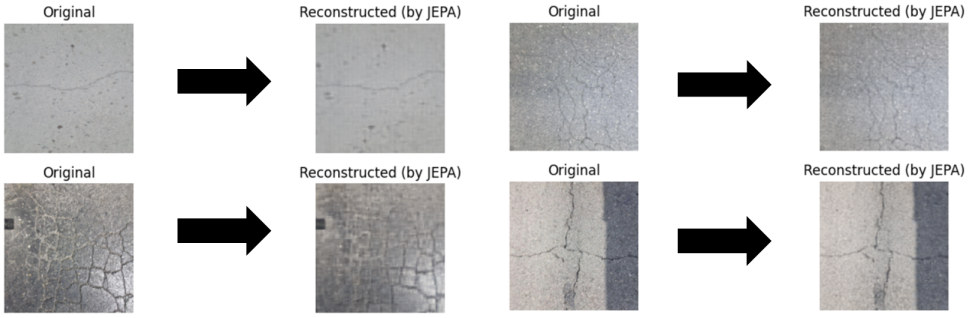

Figure 9 Qualitative comparison of original pavement images and their reconstructions by the JEPA model.

6.4. Evaluation of the Adaptability of the Models

This section tackles a key question for real-world use. Which architecture generalizes better on actual datasets from the field, especially when the data looks nothing like the training set and labeled samples are scarce? To find this out, a domain adaptation experiment was conducted on only 640 data (real-world images) captured from road weather cameras, as described in detail in Section 5.6.

As illustrated in Figure 10, there is a significant visual disparity between the source domain (standard data) and the target domain (weather cam-

era data). The target images present several challenges, including viewpoint variations, low resolution, lens obstructions , and varying lighting conditions, which collectively increase the complexity of the task. Therefore in both cases, fine-tuning was conducted using the target dataset. For the YOLOv11 baseline model fine-tuning, the conventional fine-tuning strategy was adopted. In the case of the JEPA model, the strategy used was the Partial Freezing strategy (Phase II), in which the initial 10 blocks of the ViT Encoder were frozen to conserve the strong domain-irrelevant geometric information extracted in the pre-training stage. Only the last 2 blocks and the projection head were fine-tuned using differential learning rates.

Figure 10 Illustration of the domain shift between datasets. (a) The source domain (Standard Dataset) containing high-fidelity crack samples. (b) The target domain (Weather Camera Dataset).

Table 3 Performance of models in identifying linear cracks in domain change scenarios.

Linear			
YOLO		JEPA(GAP)	
Precision	Recall	Precision	Recall
58.33	87.50	73.47	75.00
F1-score = 70.00		F1-score = 74.23	

For linear cracks as can be seen from Table 3 and visually summarized in Figure 11 (a), both models suffered from the domain gap, while the

JEPA(GAP) model performs much more stably. With a Recall of 87.50%, YOLO was found to be much more overconfident with a Precision of 58.33% and gave much false-positives in the noisy environment. On the other hand, the JEPA model performance was decent with a Precision of 73.47% and a Recall of 75.00%. Therefore, a 74.23% F1-score is got by the JEPA network, which is higher than the result in YOLO of 70.00%. That means the pre-trained features in JEPA are much less sensitive to environmental noise such as shadows or road markings, which always confuses the standard supervised model in linear detection.

Table 4 Performance of models in identifying alligator cracks in the domain change scenario

Alligator			
YOLO		JEPA (GAP)	
Precision	Recall	Precision	Recall
75.00	37.50	74.47	72.92
F1-score = 50.00		F1-score = 73.68	

The performance gap increases substantially in the Alligator category, which involves complex patterns, as illustrated in Figure 11 (b) and Table 4. The performance of the YOLO model has deteriorated severely, obtaining only a Recall of 37.50% and an F1-score of 50.00%. This drastic fall shows that the baseline model was unsuccessful in transferring its expertise in extracting discriminative features to the complex patterns of the new domain. On the other hand, the performance of the JEPA(GAP) was nothing short of impressive, gaining an F1-score of 73.68% that enhanced by a staggering 23% compared to the baseline model. With a balanced Precision of 74.47% and a Recall of 72.92%, the JEPA model was able to prove that the semantic features acquired in the self-supervised pre-training phase are significantly transferable. The model appropriately utilized the Partial Freezing scheme to modify the domain-independent semantic understanding to the new domain without getting affected by the catastrophic forgetting problem that severely affected the performance of the YOLO baseline model.

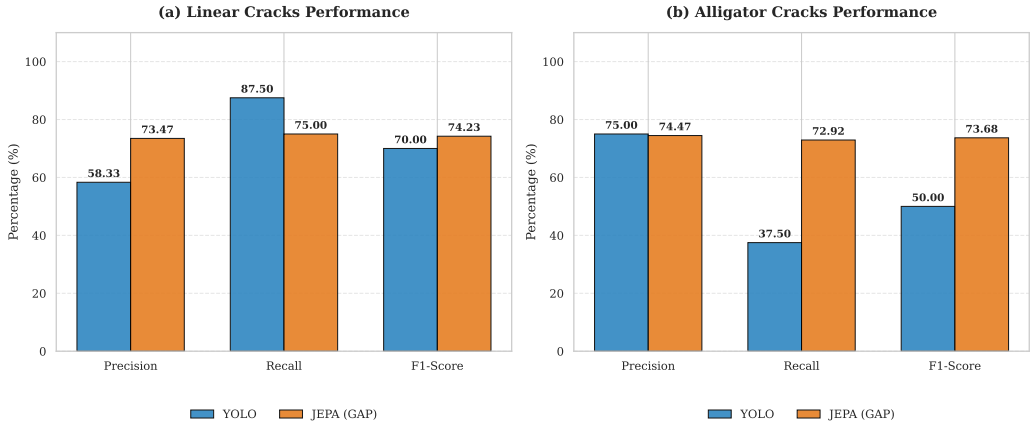

Figure 11 Comparative performance analysis of YOLO and JEPA(GAP) models in the target domain environment. The grouped bar charts illustrate Precision, Recall, and F1-Score metrics for (a) Linear cracks and (b) Alligator cracks.

7. Limitations

Although the architecture based on JEPA shows better adaptation capabilities in the experiments presented here, there are a number of limitations that keep the results close to the current scope and point out important directions for future investigation. The key limitations originate from technical design choices about input resolution, diversity in distress categories, dataset size for domain adaptation, and preprocessing pipeline. All these have to be pointed out as important framing of the results and roadmapping of future research.

The main limitation from the technical perspective originates from the resolution of the inputs. High resolution images were resized to 224×224 pixels to make them compatible with the existing Vision Transformer model. This results in the loss of minute details in the images, making it difficult to identify micro-cracks in the early stage of distress. Future research should investigate advanced input processing strategies that allow the model to leverage high-fidelity information without being constrained by the fixed input size of standard vision transformers.

Another important limitation has to do with the number of distress categories considered. For this experiment, only two categories of distress are considered: Linear Cracks and Alligator Cracks. Though this narrowed the comparison, in a practical pavement management system, more categories

of distress need to be identified. Moreover, the experiment has not implemented severity levels, but these levels have to be identified as low, medium, or high in order for scheduling of maintenance to occur. Future work has to concentrate on adding more distress categories in the annotation of the dataset by incorporating the multi-task learning head.

Concerning the assessment of the results regarding the adaptation of the domains, although there is a superiority of the results for the JEPA approach in terms of performance, the sample set for the target domain only consists of 640 images. While there is an apparent consistency for the findings regarding the performance assessment, the results' statistical integrity is limited due to the sample size. To ensure the soundness of the results through statistics, assessment tests for larger samples are necessary. The sample size for the target set can include images covered by different environmental settings.

Finally, it should be noted that in the present workflow, the pavement surface extraction is done manually in order to address domain changes. However, since this is done manually, it makes the workflow not completely automatic. Hence, in future versions of the framework, modules for automatic region of interest extraction need to be implemented in order to make the workflow completely automatic.

It is important to note that addressing these specific limitations fell outside the defined scope of the current study, as the study was mainly focused on the validation of the primary adaptation process of the JEPA system architecture. Thus, the identified gaps have not been investigated in the context of the existing study but can be considered promising topics for carrying out future studies aimed at improving the applicability of the system.

8. Conclusions

The aim of the study was to provide a solid foundation for a pavement distress classification model that relies on the assumption that a self-supervised pre-training model using the JEPA architecture which extracts more abstract and semantic features than the traditional supervised learning models, such as YOLOv11. The empirical findings support the assumption and can be outlined in the following three main points:

(1) superiority over standard data and practical implications: Comparative analysis based on standard datasets proved the overall superiority of

the proposed JEPA(GAP) method over both the YOLOv11 benchmark and the standard JEPA(cls) configuration.

In the Linear cracks category, YOLO was amazingly successful (Recall: 99.00%, F1-score: 98.02%), and the JEPA(GAP) model proved the best all-rounder in this category because it was able to achieve a Recall of 100% and an F1-score of 99.50% respectively. From a practical perspective, having a 100% recall is critical in the aspect of preventive maintenance because it ensures that the sealing process has been started at the earliest point.

In the Alligator cracks category,the strength of the proposed technique in dealing with complex topologies was further obvious in this case. Even though YOLO had a decent F1-score of 97.98%, the JEPA(GAP) model had a perfect Precision of 100.00% and an F1-score of 99.50%. The absence of false positives means that this model can distinguish between complex distress signals and background noise exceptionally well, and this is very important for effective spending of funds, as it avoids wasting costly resource on false signals.

(2) Efficacy of Global Average Pooling: The comparison between JEPA(cls) and JEPA(GAP) has demonstrated the architectural relevance of the pooling strategy. Indeed, although JEPA(cls) is very precise, the performance of the JEPA(GAP) model is always more balanced and robust. This is because, for tasks such as pavement distress detection, which are focused on textures, for which the semantic content is widespread in the entire image, the spatial aggregation of the features done by the GAP is more informative than the CLS feature.

(3) Strong adaptation to domain differences: The experiment of domain adaptation using real-world images further confirmed the robustness. Even though there were large domain differences, the JEPA (GAP), with a partial freezing strategy, showed an admirable adaptability. Compared to the unstable results and low F1-score of the baseline YOLO model, which struggled with complex textures in Alligator cracks images (F1-score: 50.00%), the results for the JEPA model were very encouraging, with a high F1-score of 73.68%. This proves the excellent transferability of the semantic features obtained in the pre-

training phase.

In conclusion, what this study proposes is that the future of intelligent infrastructure inspections will be found in architectural paradigms that extends from simple pattern recognition to contextually informed insights. Future work will include proving out this approach on full-scale real-world data and extending this robust framework to more complex dense prediction tasks, such as semantic segmentation.

Conflict of Interest

The authors declare no competing interests.

Declaration of Funding

The authors declare no funding was received.

References

[1] Arya, D., Maeda, H., Ghosh, S.K., Toshniwal, D., Sekimoto, Y., 2024. Rdd2022: A multi-national image dataset for automatic road damage detection. Geoscience Data Journal 11, 846–862.

[2] Assran, M., Caron, M., Misra, I., Bojanowski, P., Bordes, F., Vincent, P., Joulin, A., Rabbat, M., Ballas, N., 2022. Masked siamese networks for label-efficient learning, in: European conference on computer vision, Springer. pp. 456–473.

[3] Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N., 2023. Self-supervised learning from images with a joint-embedding predictive architecture, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15619–15629.

[4] Bao, H., Dong, L., Piao, S., Wei, F., 2021. Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 .

[5] Bar, A., Bordes, F., Shocher, A., Assran, M., Vincent, P., Ballas, N., Darrell, T., Globerson, A., LeCun, Y., 2024. Predicting masked tokens in stochastic locations improves masked image modeling. URL: https://openreview.net/forum?id=jLnygpRFYm.

[6] Bouhsissin, S., Assemlali, H., Sael, N., 2025. Enhancing road safety: A convolutional neural network based approach for road damage detection. Machine Learning with Applications , 100668.

[7] Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PmLR. pp. 1597–1607.

[8] Du, Y., Pan, N., Xu, Z., Deng, F., Shen, Y., Kang, H., 2021. Pavement distress detection and classification based on yolo network. International Journal of Pavement Engineering 22, 1659–1672.

[9] Dutta, P., Kanti Podder, K., Zhang, J., Hecht, C., Swarna, S., Bhavsar, P., 2024. A self-supervised learning approach to road anomaly detection using masked autoencoders, in: International Conference on Transportation and Development 2024, pp. 536–547.

[10] El-Din Hemdan, E., Al-Atroush, M., 2025. A review study of intelligent road crack detection: Algorithms and systems. International Journal of Pavement Research and Technology , 1–31.

[11] Fintraffic, 2025. Liikennetilanne - kelikamerat (road weather cameras). URL: https://www.liikennetilanne.fi/kelikamerat. accessed: 2025.

[12] Geng, H., Liu, Z., Wang, Y., Fang, L., 2025. Sdfc-yolo: A yolo-based model with selective dynamic feature compensation for pavement distress detection. IEEE Transactions on Intelligent Transportation Systems .

[13] Gopalakrishnan, K., 2018. Deep learning in data-driven pavement image analysis and automated distress detection: A review. Data 3, 28.

[14] Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar,

M., et al., 2020. Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems 33, 21271–21284.

[15] Guerrieri, M., Parla, G., Khanmohamadi, M., Neduzha, L., 2024. Asphalt pavement damage detection through deep learning technique and cost-effective equipment: A case study in urban roads crossed by tramway lines. Infrastructures 9, 34.

[16] Gui, J., Chen, T., Zhang, J., Cao, Q., Sun, Z., Luo, H., Tao, D., 2024. A survey on self-supervised learning: Algorithms, applications, and future trends. IEEE Transactions on Pattern Analysis and Machine Intelligence 46, 9052–9071.

[17] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 16000–16009.

[18] He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9729–9738.

[19] Hondru, V., Croitoru, F.A., Minaee, S., Ionescu, R.T., Sebe, N., 2025. Masked image modeling: A survey. International Journal of Computer Vision , 1–47.

[20] Hu, H., Wang, X., Zhang, Y., Chen, Q., Guan, Q., 2024. A comprehensive survey on contrastive learning. Neurocomputing 610, 128645.

[21] Kalapos, A., Gyires-Tóth, B., 2024. Cnn-jepa: Self-supervised pretraining convolutional neural networks using joint embedding predictive architecture, in: 2024 International Conference on Machine Learning and Applications (ICMLA), IEEE. pp. 1111–1114.

[22] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D., 2020. Supervised contrastive learning. Advances in neural information processing systems 33, 18661–18673.

[23] Kothai, R., Prabakaran, N., Murthy, Y.S., Cenkeramaddi, L.R., Kakani, V., 2024. Pavement distress detection, classification and analysis using machine learning algorithms: a survey. IEEE Access .

[24] Li, W., Yang, W., Liu, T., Hou, Y., Li, Y., Liu, Z., Liu, Y., Liu, L., 2024. Predicting gradient is better: Exploring self-supervised learning for sar atr with a joint-embedding predictive architecture. ISPRS Journal of Photogrammetry and Remote Sensing 218, 326–338.

[25] Lv, Z., Hao, Z., Zhu, Y., Lu, C., 2025. A review on automated detection and identification algorithms for highway pavement distress. Applied Sciences 15, 6112.

[26] Manjusha, M., Sunitha, V., 2025. Optimizing yolo models for high-accuracy automated detection and classification of road surface distresses. Innovative Infrastructure Solutions 10, 381.

[27] Mo, S., Tong, S., 2024. Connecting joint-embedding predictive architecture with contrastive self-supervised learning. Advances in neural information processing systems 37, 2348–2377.

[28] Monemi, M., Chinipardaz, M., Rasti, M., Bennis, M., Latva-aho, M., 2025. Tutorial on joint embedding predictive architectures (jepa): Foundations, applications, and future directions. TechRxiv preprint. URL: https://doi.org/10.36227/techrxiv.176469421.19270944/v2, doi:10.36227/techrxiv.176469421.19270944/v2. preprint submitted December 10, 2025.

[29] Mubashshira, S., Azam, M.M., Ahsan, S.M.M., 2020. An unsupervised approach for road surface crack detection, in: 2020 IEEE Region 10 Symposium (TENSYMP), IEEE. pp. 1596–1599.

[30] Ning, Z., Wang, H., Li, S., Xu, Z., 2024. Yolov7-rdd: A lightweight efficient pavement distress detection model. IEEE Transactions on Intelligent Transportation Systems 25, 6994–7003.

[31] Qiu, Q., Lau, D., 2023. Real-time detection of cracks in tiled sidewalks using yolo-based method applied to unmanned aerial vehicle (uav) images. Automation in Construction 147, 104745.

[32] Ragnoli, A., De Blasiis, M.R., Di Benedetto, A., 2018. Pavement distress detection methods: A review. Infrastructures 3, 58.

[33] Ren, S., He, K., Girshick, R., Sun, J., 2016. Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence 39, 1137–1149.

[34] Rezaeimanesh, M., Golroo, A., Fahmani, M.S., Amani, M.J., Hasanitabar, F., Entezari, M.S., Karimi, S., 2025. Attain: Inclusive annotated pavement distress types and severity dataset. Data in Brief , 111715.

[35] Song, Q., Yao, W., Tian, H., Guo, Y., Muniyandi, R.C., An, Y., 2024. Two-stage framework with improved u-net based on self-supervised contrastive learning for pavement crack segmentation. Expert Systems with Applications 238, 122406.

[36] Tamagusko, T., Gomes Correia, M., Ferreira, A., 2024. Machine learning applications in road pavement management: A review, challenges and future directions. Infrastructures 9.

[37] Thimonier, H., Costa, J.L.D.M., Popineau, F., Rimmel, A., Doan, B.L., 2024. T-jepa: Augmentation-free self-supervised learning for tabular data. arXiv preprint arXiv:2410.05016 .

[38] Wang, S., Chen, X., Dong, Q., 2023. Detection of asphalt pavement cracks based on vision transformer improved yolo v5. Journal of Transportation Engineering, Part B: Pavements 149, 04023004.

[39] Wang, W., Yu, X., Jing, B., Tang, Z., Zhang, W., Wang, S., Xiao, Y., Li, S., Yang, L., 2025. Yolo-rd: A road damage detection method for effective pavement maintenance. Sensors 25, 1442.

[40] Wei, Y., Peng, Y., Cheng, H., Wang, D., Zhang, A.A., 2025. Yolo-rapd: Enhanced yolov8s-based automated detection of road assets and pavement distress. Journal of Computing in Civil Engineering 39, 04025092.

[41] Xu, X., Tao, L., Zou, L., Qin, H., Deng, Z., Zheng, F., 2025. An enhanced yolov11-based approach for pavement distress detection via multi-scal feature fusion and adaptive learning, in: 2025 10th International Conference on Information Science, Computer Technology and Transportation (ISCTT), IEEE. pp. 121–125.

[42] Zhang, H., Dong, Y., Hou, Y., Cheng, X., Xie, P., Di, K., 2025. Research on asphalt pavement surface distress detection technology coupling deep learning and object detection algorithms. Infrastructures 10, 72.

[43] Zhang, Y., Lu, Y., Huo, Z., Li, J., Sun, Y., Huang, H., 2024. Ussc-yolo: Enhanced multi-scale road crack object detection algorithm for uav image. Sensors (Basel, Switzerland) 24, 5586.

[44] Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T., 2021. ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832 .