Robustness Evaluation of a JEPA-Based Model and YOLO for Pavement Distress Classification under Noise Corruptions

Soroush Amiri[a], Amir Golroo, Fereidoon Moghadas Nejad[a], Mehdi Rasti[b], Mehdi Monemi[b], Hannaneh Dehghan Tezerjani[a]
[a]Department of Civil and Environmental Engineering, Amirkabir University of Technology, Tehran, Iran
[b]Centre for Wireless Communications, University of Oulu, Oulu, Finland

Abstract

While automated pavement distress classification is critical for Intelligent Transportation Systems (ITS), conventional evaluations on pristine datasets often mask deep learning vulnerabilities to real-world environmental degradation. This study evaluates architectural robustness by contrasting a supervised Convolutional Neural Network (YOLOv11m-cls) with a self-supervised Vision Transformer framework (StoP-JEPA). Models were subjected to extreme Out-of-Distribution (OOD) environmental stress tests, including Defocus Blur, Gaussian Sensor Noise, and Salt-and-Pepper impulse corruption. Furthermore, a Balanced Performance Index (BPI) is introduced to quantify diagnostic symmetry between fundamentally distinct crack topologies. Despite near-perfect baseline accuracy, stress testing revealed distinct architectural inductive biases. YOLO demonstrated inherent robustness to optical blur via local convolutional low-pass filtering, whereas the StoP-JEPA configuration utilizing Global Average Pooling (GAP) excelled under Gaussian noise through statistical smoothing. Crucially, under extreme impulse noise, GAP suffered catastrophic asymmetric collapse, dropping to a BPI of 7.06% ($\pm$ 4.67%). Replacing GAP with an attention-driven [CLS] token effectively mitigated this vulnerability. By selectively routing semantic context from uncorrupted patches, the [CLS] mechanism maintained a statistically stable BPI of 61.71% ($\pm$ 3.94%). Ultimately, for practical ITS deployment, StoP-JEPA provides the most stable diagnostic equilibrium under severe Gaussian noise, while YOLO remains superior under optical blur, highlighting the necessity of aligning architectural priors with specific deployment environments.

Keywords: Image Corruptions, Joint Embedding Predictive Architecture (JEPA), Noise Robustness, Pavement Distress Classification, Robustness, Self-Supervised Learning (SSL), YOLO

1. Introduction

Pavement condition information is a significant component in Pavement Management Systems, because it allows timely maintenance to preserve the serviceability of road networks, which are critical for mobility and economic activities. Pavement damages such as cracks, potholes, and rutting reduce road performance, and if ignored, accelerate structural deterioration and increase safety risks [6, 13]. Traditional methods of pavement condition monitoring rely mainly on manual inspections, which are costly, time-consuming, and rely on individual judgment. These limitations have motivated the application of automated methods based on machine vision and deep learning. Currently, deep learning has become the dominant approach in pavement damage analysis, with existing studies mainly classified into two categories: object recognition and semantic segmentation [1, 15].

1.1. State-of-the-Art Road Damage Detection Methods

Deep learning has facilitated automated monitoring in civil infrastructure systems. The YOLO family of single-stage object detectors offers an effective trade-off between detection accuracy and computational efficiency, demonstrating strong performance in applications such as pavement damage detection [7, 9, 12], UAV-based inspection [14], and tunnel defect analysis [16]. In parallel, the Joint-Embedding Predictive Architecture (JEPA) has emerged as a self-supervised framework that learns transferable visual representations from unlabeled data. Incorporating self-supervised signals into the training process enhances model robustness and improves its ability to estimate uncertainty compared with purely supervised approaches [2, 3, 5, 10].

1.2. Image Degradations and Their Impact on Vision-Based Detection Robustness

In real-world deployments, visual fidelity is often affected by a combination of physical, environmental, and system-related factors. Sensor noise and lens distortion during image acquisition, compression artifacts and reduced resolution within the processing pipeline, as well as illumination and atmospheric variations such as shadows, reflections, fog, rain, and dust can all obscure fine surface details and degrade the accuracy of computer vision models. Moreover, camera or object motion, calibration errors, and domain shifts across different environments further weaken the performance of deep learning models. Collectively, these factors result in common corruptions

such as Gaussian noise, salt-and-pepper noise, and various forms of blur, whose identification and modeling are essential for evaluating the robustness of vision-based systems.

Evaluations incorporating real sensor and image-pipeline corruptions have consistently demonstrated a decline in detection accuracy [4, 8]. Empirical evaluations in object detection also report that mAP (mean Average Precision) decreases continuously with increasing Gaussian noise standard deviation and impulse noise probability, as well as in the presence of blur and illumination shifts; these findings motivate explicit robustness assessment in civil infrastructure imagery [8, 11].

1.3. Motivation

However, direct comparisons between YOLO and JEPA under such corruptions remain limited for pavement imagery. Given that self-supervised frameworks such as JEPA are designed to learn invariances to visual perturbations during pretraining, it is hypothesized that they may exhibit inherently greater robustness to input corruptions than purely supervised detectors such as YOLO. However, this hypothesis has not yet been empirically validated in the context of pavement distress detection. In order to bridge this identified gap, this research aims to carry out a thorough robustness analysis for deep learning architectures in the field of Intelligent Transportation Systems (ITS) under degraded visual conditions. The key contributions of this research are highlighted below:

- **Development of a JEPA-based Classification Framework:** To this end, a novel framework for automated pavement distress classification using a self-supervised Joint-Embedding Predictive Architecture backbone is proposed. The framework bypasses the need for entirely on-sight trained models by exploiting rich structural priors with global self-attention and extracting highly resilient semantic features in the presence of complex crack topologies.

- **Comprehensive Robustness Benchmarking:** This study conducts a comprehensive, zero-shot robustness evaluation, directly contrasting the representational resilience of a purely supervised convolutional baseline (YOLOv11m-cls) against the self-supervised StoP-JEPA framework. The architectures are subjected to mathematically defined environmental stress tests specifically Optical Defocus, Gaussian Sensor

Noise, and Salt-and-Pepper Impulse Corruption to expose their intrinsic architectural inductive biases.

- **Identification of Asymmetric Pooling Collapse and Mitigation:** We empirically demonstrate that standard unweighted Global Average Pooling (GAP) in Vision Transformers suffers a catastrophic, class-specific representation collapse under extreme unstructured noise. Consequently, we validate that substituting GAP with an attention-driven, single-patch aggregation strategy (the [CLS] configuration) effectively dynamically filters poisoned tokens, thereby maintaining the integrity of the learned representations under severe noise.
- **Formulation of the Balanced Performance Index (BPI):** To address the positive bias and limitations of standard overall accuracy metrics under ideal conditions, we formulate the BPI. By computing the harmonic mean of class-wise F1-scores, this proposed evaluation metric mathematically penalizes asymmetric diagnostic failures. This provides a rigorous quantification of the model's operational equilibrium and classification reliability under extreme levels of information loss.

2. Methodology

This section outlines the proposed methodological paradigm for evaluating the robustness of deep learning architectures in pavement distress classification. The core objective is to benchmark a purely supervised local feature detector (YOLOv11m-cls) against a classification model leveraging the self-supervised JEPA under mathematically defined environmental and sensor corruptions. The methodology progresses through data preparation, architectural formulation, a rigorous two-stage fine-tuning process, and concludes with the introduction of robustness evaluation metrics.

2.1. Dataset Description and Preprocessing

The dataset utilized in this study comprises 2,000 high-resolution pavement images (1868×4000 pixels), uniformly distributed across two primary distress categories: linear cracks (1,000 images) and alligator

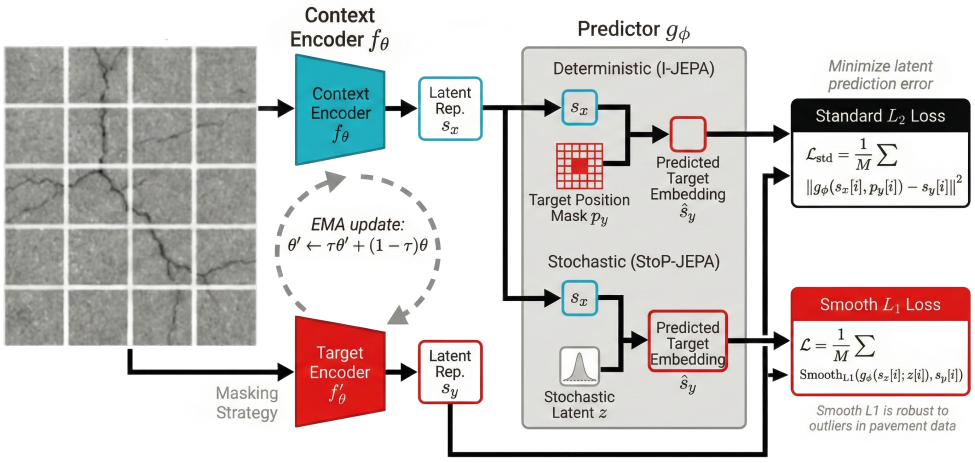


Figure 1: Schematic overview of the proposed JEPA framework for pavement distress classification.

cracks (1,000 images). To ensure a rigorous and unbiased evaluation, each category was partitioned into strictly disjoint subsets: 800 images for training, 100 for validation, and 100 for testing. To minimize spatial variability and structural distortions stemming from perspective and scale inconsistencies, all images were captured under standardized conditions with fixed camera angles and focal distances.

During the preprocessing phase, the high-dimensional input images were uniformly resized to 224×224 pixels to align with the standard spatial input requirements of the Vision Transformer (JEPA) backbone. Furthermore, the pixel intensities were normalized using the channel-wise mean and standard deviation of the ImageNet dataset. This normalization step is critical to accelerate gradient convergence and ensure statistical stability across the feature extraction layers.

2.2. Stochastic Joint-Embedding Predictive Architecture (StoP-JEPA)

To capture high-level semantic features of pavement distress without relying on pixel-level reconstruction, we integrate the Stochastic Joint-Embedding Predictive Architecture (StoP-JEPA) into our framework, as illustrated in Fig. 1. The architecture operates entirely in the latent

space and consists of three primary modules: a context encoder f_θ, a target encoder $f_{\theta'}$, and a predictor g_ϕ.

Given an input image, it is partitioned into a sequence of non-overlapping patches. A masking strategy is then applied to select a context block x and a target block y. The context encoder maps the visible context patches into a latent representation $s_x = f_\theta(x)$. Simultaneously, the target encoder processes the target patches to generate the ground-truth latent embedding $s_y = f_{\theta'}(y)$. To prevent representation collapse a critical failure mode where encoders produce trivial, constant outputs regardless of the input a stop-gradient operation is applied to the target encoder. The target encoder's weights θ' are not updated through backpropagation. Rather, they are updated solely by means of an Exponential Moving Average (EMA) of the context encoder's weights θ:

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

where τ is a momentum coefficient that gradually increases during training.

Unlike the deterministic spatial relationships assumed by standard JEPA, pavement cracks exhibit complex, irregular structures with highly uncertain edge information. To address this limitation, our framework leverages StoP-JEPA to explicitly account for edge uncertainty and non-deterministic spatial distributions. This method replaces standard positional embeddings with stochastic positional embeddings. Specifically, the masked patch position is modeled as a distribution, introducing a stochastic latent variable z to the prediction process to yield the predicted target representation $\hat{s}_y$:

$$\hat{s}_y = g_\phi(s_x; z) \tag{2}$$

Finally, to enhance robust feature extraction and mitigate the impact of severe outliers intrinsic to real-world pavement imaging, we employ the Smooth L_1 Loss instead of the traditional L_2 distance. The pre-training objective minimizes this prediction error in the stochastic latent space for a batch of size M, where i denotes the index of each sample:

$$\mathcal{L} = \frac{1}{M}\sum_{i=1}^{M} \mathrm{Smooth}_{L_1}(g_\phi(s_x^{(i)}; z^{(i)}), s_y^{(i)}) \tag{3}$$

2.3. Feature Aggregation Strategies: Implicit Anchor vs. GAP

To adapt the pre-trained JEPA encoder to the downstream task of pavement distress classification, the sequence of patch embeddings from the final layer of the encoder must be consolidated into a single global representation. For this study, we propose and evaluate two distinct feature aggregation schemes in terms of their structural robustness to environmental noise.

Strategy I: Implicit Spatial Anchor (Single-Patch Routing) Unlike standard Vision Transformer protocols that introduce an artificial, prepended classification token, this strategy directly utilizes the latent representation of the first spatial patch (Index 0) as an implicit anchor for classification. This design choice is strategically motivated by two architectural imperatives. First, it avoids the cold start problem of introducing a newly initialized, random token into a mature, pre-trained self-supervised backbone. second, it maintains strict mathematical compatibility with the 2D spatial coordinate system required by the StoP framework, which a coordinate-free global token cannot naturally accommodate. Due to the dense, global self-attention mechanisms operating throughout the deep JEPA encoder, the receptive field of any individual patch spans the entire image by the final layer. Consequently, the representation of the first patch is no longer strictly localized; rather, it inherently captures contextual, image-level structural information.

During the downstream adaptation, the network optimizes its attention weights to dynamically route diagnostic features from across the sequence into this specific spatial anchor. Therefore, the final hidden state of this patch at the last encoder layer L, denoted as $z_0^{(L)}$, is extracted as the unified representation. The final class probability is determined through a linear projection head defined by:

$$y = \mathbf{W} \cdot \text{BatchNorm}(z_0^{(L)}) + \mathbf{b} \tag{4}$$

where $\mathbf{W}$ and $\mathbf{b}$ are the learned weight matrix and bias vector of the classification head, respectively.For the remainder of this paper, to maintain notational consistency with standard Vision Transformer literature, we refer to this implicit anchor configuration simply as the `[CLS]` strategy.

Strategy II: Global Average Pooling (GAP)
The second strategy uses the spatial feature maps directly. In other words, instead of using a representative patch, the global representation is generated using the arithmetic mean of all the spatial patch embeddings in the final encoder layer L:

$$z_{\text{gap}} = \frac{1}{N} \sum_{i=1}^{N} z_i^{(L)} \tag{5}$$

where $z_i^{(L)}$ is the latent embedding of the i-th spatial patch in the final layer L, while N is the total number of patches in the sequence. Then, the generated global representation is passed through the batch normalization layer and the linear projection layer. Although computationally simple, theoretically speaking, the above averaging approach is more susceptible to the absorption of severe pixel-level corruptions, leading to classifier collapse under severe noise, as will be shown later.

2.3.1. Supervised Fine-Tuning Strategy

In order to classify pavement distress using the pre-trained StoP-JEPA encoder, a linear classification head is appended to the context encoder. The two-stage fine-tuning process is carried out over a total of 100 epochs (15 epochs for the first stage and 85 epochs for the second stage) using the AdamW optimizer with a weight decay of 0.01.

In the first stage, the transformer backbone is kept frozen to function as a fixed feature extractor. Only the newly initialized classification head is trained with an initial learning rate of $\eta = 1 \times 10^{-3}$. This phase is designed to allow the classifier's decision boundaries to rapidly align with the self-supervised feature space without altering or degrading the robust representations acquired during the JEPA pre-training.

In the second stage, the entire model is unfrozen and updated in an end-to-end manner. To prevent catastrophic forgetting and maintain the integrity of the pre-trained structural knowledge, a discriminative learning rate strategy is applied. The encoder parameters are updated with a strictly conservative initial learning rate of $\eta_{enc} = 1 \times 10^{-5}$, while the classification head continues its refinement at a reduced rate of $\eta_{head} = 1 \times 10^{-4}$. To ensure stable and optimal convergence, the

learning rates in both stages are dynamically decayed following a Cosine Annealing schedule.

The objective is to minimize the Label-Smoothed Cross-Entropy loss, denoted as $\mathcal{L}_{LS-CE}$, for a mini-batch of $M = 32$ samples. By softening the hard targets, label smoothing prevents the network from becoming over-confident and further acts as a regularization technique against unstructured noise. This can be expressed as follows:

$$\mathcal{L}_{LS-CE} = -\frac{1}{M}\sum_{i=1}^{M}\sum_{c=1}^{C}\tilde{y}_{i,c}\log(\hat{y}_{i,c}) \tag{6}$$

where $\tilde{y}_{i,c} = y_{i,c}(1-\alpha)+\alpha/C$ is the smoothed target distribution, $y_{i,c}$ is the binary indicator (0 or 1) representing whether class c is the correct ground-truth label, $\alpha = 0.1$ is the smoothing factor, and $\hat{y}_{i,c}$ is the predicted probability.

2.4. Corruption Modeling and Robustness Protocol

To rigorously evaluate the operational reliability of the trained architectures for real-world pavement maintenance programs and ITS, we introduce a controlled robustness assessment protocol. The core of this methodology involves exposing the pristine baseline models to three mathematically defined categories of artificial image degradation. These corruptions specifically simulate the physical, environmental, and hardware-related failures frequently encountered during automated pavement surface imaging. It is crucial to note that these corruptions were applied strictly during the evaluation phase on the clean test dataset. This Out-of-Distribution (OOD) corruption protocol ensures that the evaluation measures the inherent structural resilience of the architectures rather than their ability to memorize augmented noise patterns.

2.4.1. Defocus Blur (Optical Focus Error Simulation)

In practical pavement inspection, optical defocusing frequently occurs due to the high mounting altitude of cameras on inspection vehicles combined with incorrect focal length calibrations. To mathematically approximate this out-of-focus optical degradation, the blurred image

$I_{\mathrm{B}}(x, y)$ is obtained by convolving the original image with a 2D Gaussian kernel $G(x, y)$ of size $k \times k$:

$$I_{\mathrm{B}}(x, y) = I(x, y) * G(x, y) \tag{7}$$

Based on our rigorous stress-testing framework, we applied progressive optical degradation using massive kernel sizes: $k \in \{65, 85, 105\}$. The standard deviation (σ) of the filter in both spatial directions was automatically computed relative to the kernel size to maintain optical consistency, effectively obfuscating the fine-grained edges of linear and alligator cracks at higher intensities.

2.4.2. Salt and Pepper Noise (Transmission Error and Dead Pixel Simulation)

This form of noise represents sudden impulse errors in images due to data transmission losses between the external camera and processing unit, as well as defective pixels in the hardware's sensor array. It creates deterministic random variations where some pixels are fixed at extreme values, either salt (255) or pepper (0). Let H and W denote the height and width of the images, and p denote the probability of nominal corruption. The number of pixels affected for each of these noise types is defined as:

$$N_{\mathrm{salt}} = N_{\mathrm{pepper}} = \lfloor 0.5 \times p \times H \times W \rfloor \tag{8}$$

In this study, we evaluated three levels of salt and pepper noise with severe corruption probabilities: $p \in \{0.30, 0.60, 0.90\}$. However, a key stochastic effect arises at very high nominal intensities. Since the noise coordinates are generated via independent random sampling with replacement from uniform discrete distributions, coordinate collisions become inevitable. Mathematically, the probability that a specific pixel escapes being selected for salt noise after N_{salt} independent draws is given by $(1 - \frac{1}{H \times W})^{N_{\mathrm{salt}}}$, which approximates to $e^{-p/2}$ for large high-resolution pavement images. Consequently, the effective corruption rate is strictly lower than the nominal rate p. This probabilistic coordinate collision explains why even at a nominal intensity of 90%, the structural topology of the pavement is not entirely eradicated, retaining scattered remnants of semantic information.

2.4.3. Gaussian Noise (Low Light Sensor Noise Simulation)

In image acquisition by automated pavement imaging systems, significant electrical noise is introduced to the sensor under low-illumination conditions. Gaussian noise is added to simulate this sensor electrical noise. The noisy pixel $I_{\mathrm{G}}(x, y, c)$ is obtained by adding a continuous random noise matrix to the original image pixel $I(x, y, c)$. Since the noise is applied to a 3-channel image, the operation is defined as:

$$I_{\mathrm{G}}(x, y, c) = \mathrm{clip}(I(x, y, c) + \mathcal{N}(0, \sigma^2), 0, 255) \tag{9}$$

where $c \in \{R, G, B\}$ denotes the color channel. To evaluate the models under progressive sensor degradation, the noise intensity was tested across three severe standard deviations: $\sigma \in \{60, 90, 120\}$ (with a mean $\mu = 0$). The clip function ensures the final pixel intensities remain within the valid 8-bit color space.

2.5. Evaluation Metrics

A critical set of evaluation metrics was used to measure the quantitative classification performance and structural strength of the proposed models. The classification performance of each type of pavement distress is primarily presented based on the standard criteria of Precision, Recall, and an F1-score in this study. Let TP, FP, and FN denote the number of True Positives, False Positives, and False Negatives, respectively. These basic metrics are defined mathematically as follows;

Precision: It measures the ratio of correctly predicted positive observations to the total predicted positives:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{10}$$

Recall (Sensitivity): It measures the proportion of correctly predicted positive observations out of all the positives in the dataset:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

Harmonic mean of Precision and Recall. It offers a single combined metric of overall accuracy, which is essential for an infrastructure inspection system where false negatives and false positives cost a lot:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

While standard metrics or simple arithmetic averages provide a general overview, they often mask severe class imbalances, especially under noisy conditions. In real-world pavement maintenance, a model that perfectly detects linear cracks but completely fails to identify critical alligator cracks is operationally unacceptable. To quantitatively measure the classification equilibrium between the distinct distress types, we introduce the Balanced Performance Index (BPI). Mathematically, BPI is formulated as the harmonic mean between the F1-scores of the two constituent classes:

$$\mathrm{BPI} = 2 \times \frac{\mathrm{F1}_{\mathrm{Linear}} \times \mathrm{F1}_{\mathrm{Alligator}}}{\mathrm{F1}_{\mathrm{Linear}} + \mathrm{F1}_{\mathrm{Alligator}}} \tag{13}$$

The underlying rationale behind using the harmonic mean instead of naively computing the averages based on arithmetic measures is its inherent mathematical property to heavily punish wide gaps. If the model is highly accurate for linear cracks but fails catastrophically for alligator cracks in the presence of environmental noise, then the BPI goes down the drain. High BPI, therefore, strictly guarantees symmetrical and reliable diagnostic capability for all pavement distress topology with the selected model.

Statistical Validation via Bootstrap Resampling

Given the relatively constrained size of the test set (100 images per class), relying on a single deterministic point estimate for the aforementioned evaluation metrics could render the results sensitive to sample variance. To rigorously establish statistical significance and compute the 95% Confidence Intervals (CIs) reported in our robustness analysis, a nonparametric bootstrapping procedure was employed during the testing phase.

Specifically, for each network configuration and corruption intensity, the test set predictions were randomly resampled with replacement for $B = 1,000$ iterations. The evaluation metrics (Precision, Recall, F1-score, and BPI) were recalculated for each bootstrap sample. The final reported performance is expressed as the mean of this bootstrapped distribution, accompanied by its standard deviation ($\pm$ Std). This

probabilistic approach guarantees that the observed architectural vulnerabilities and the performance disparities between YOLO and StoP-JEPA are statistically robust and not merely artifacts of a specific test set distribution.

2.6. Baseline Configuration: Purely Supervised vs. Self-Supervised Transfer Learning

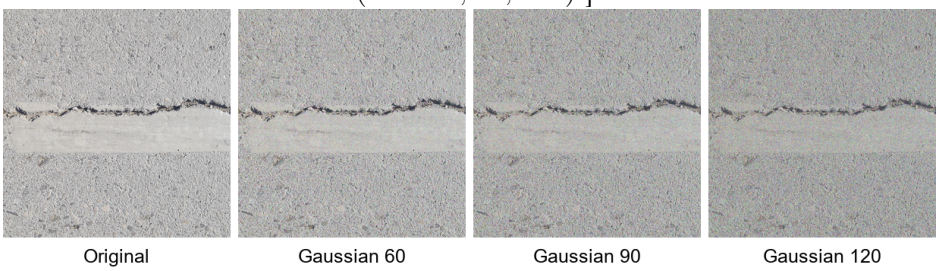
To establish a rigorous baseline representing the traditional purely supervised learning paradigm, the medium-sized classification variant of the YOLO architecture (YOLOv11m-cls) was employed. Unlike the proposed StoP-JEPA framework, which utilizes a two-stage transfer learning approach leveraging a massive self-supervised pre-training phase to build robust domain-specific representations followed by a supervised fine-tuning phase, the YOLOv11m-cls model relies entirely on supervised learning using explicit dataset labels.

The YOLO model was trained for exactly 100 epochs using the identical dataset splits. Crucially, the YOLO series inherently relies on complex, built-in data augmentation pipelines to artificially boost their robustness. However, the primary objective of this comparative analysis is to evaluate the intrinsic structural and representational resilience of these two distinct learning paradigms under environmental stress, rather than the effectiveness of external data augmentation algorithms. Therefore, to isolate the fundamental learning capabilities, all internal synthetic data augmentation hyperparameters within the YOLOv11 training configuration were explicitly disabled. This setup logically guarantees that any performance degradation observed during OoD noise stress tests is solely attributable to the core methodological differences, namely the vulnerability of YOLO's purely supervised local convolutions, versus the robustness of StoP-JEPA's global self-attention mechanism, which is profoundly fortified by self-supervised structural priors.

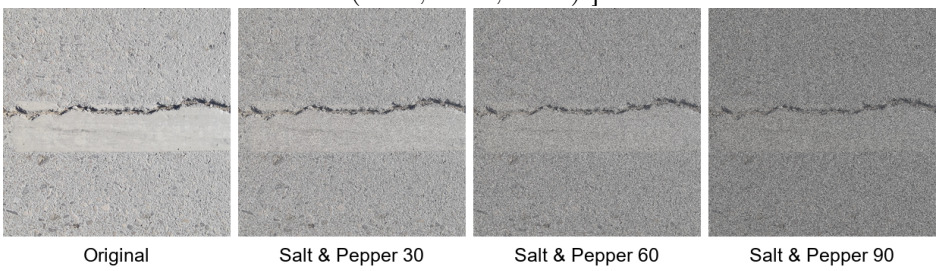
3. Results and Analysis

In this section, we conduct a comprehensive comparative analysis between the purely supervised YOLOv11m-cls baseline and the self-supervised StoP-JEPA framework (utilizing both GAP and [CLS] aggregation

[Examples of Gaussian noise corruption at different intensity levels ($\sigma = 60, 90, 120$).]



[Examples of Salt-and-pepper noise corruption at different intensity levels (30%, 60%, 90%).]



[Examples of Gaussian blur corruption at different kernel sizes ($65 \times 65, 85 \times 85, 105 \times 105$).]

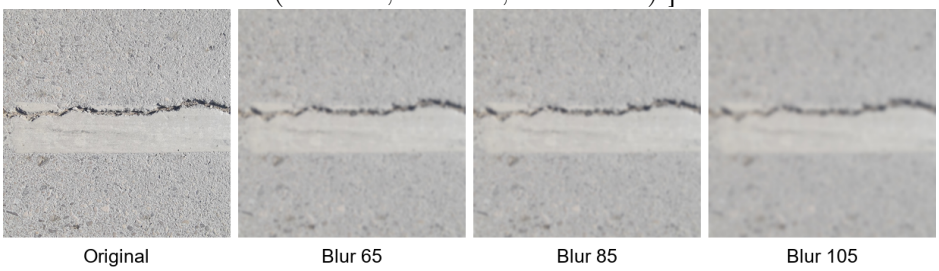


Figure 2: Visual examples of the three corruption types applied to a sample pavement crack image from the test set. Each row shows the original image and the corresponding corruptions at three increasing intensity levels: (a) Gaussian noise, (b) Salt-and-pepper noise, and (c) Gaussian blur.

strategies) for the automated classification of pavement distress morphologies, specifically Linear and Alligator cracking. The detailed empirical outcomes, evaluated across multiple intensities of simulated environmental corruptions, are systematically presented in Table I (Alligator cracks) and Table II (Linear cracks). Furthermore, Table III introduces the BPI to quantitatively assess the diagnostic symmetry between

the classes. As the initial pristine-data results indicate, both architectural paradigms achieve exceptional, near-identical baseline proficiency. However, JEPA architecture demonstrates a superior initial feature extraction capability, particularly indicating a stronger inherent alignment with the complex, multi-directional topological features of alligator cracks from the outset. The subsequent subsections meticulously dissect these empirical behaviors: evaluating baseline capabilities, analyzing architectural vulnerabilities against three distinct mathematical noise paradigms, and quantifying overall operational reliability.

3.1. Baseline Performance on Pristine Data

Before exposing the models to environmental stress tests, it is critical to establish their baseline classification capabilities under optimal, noise-free conditions. The test set containing pristine pavement images was utilized to benchmark the supervised YOLOv11m-cls baseline against the proposed StoP-JEPA framework. The empirical results demonstrate that all models achieve exceptional baseline accuracy, reaffirming the fundamental viability of deep learning for automated pavement distress classification.

Specifically, as detailed in Table 3, the YOLOv11m-cls model yielded a robust Balanced Performance Index (BPI) of 98.46%, driven by highly symmetrical F1-scores of 98.46% for alligator cracks and 98.47% for linear cracks. The StoP-JEPA framework, empowered by its domain-specific self-supervised pre-training, produced similarly exceptional results. The GAP configuration achieved a BPI of 99.02% (with F1-scores of 99.01% and 99.02% for alligator and linear cracks, respectively), while the [CLS] configuration followed closely with a BPI of 98.98%.

While the absolute BPI scores of the StoP-JEPA models are nominally higher, the overlapping 95% confidence intervals (visualized in Figure 3) indicate statistical parity between the two distinct learning paradigms under ideal conditions. Therefore, rather than demonstrating an absolute initial superiority, the tighter variance exhibited by the StoP-JEPA models (with lower bounds of 97.50% and 97.48%) provides early evidence of stable feature representations.

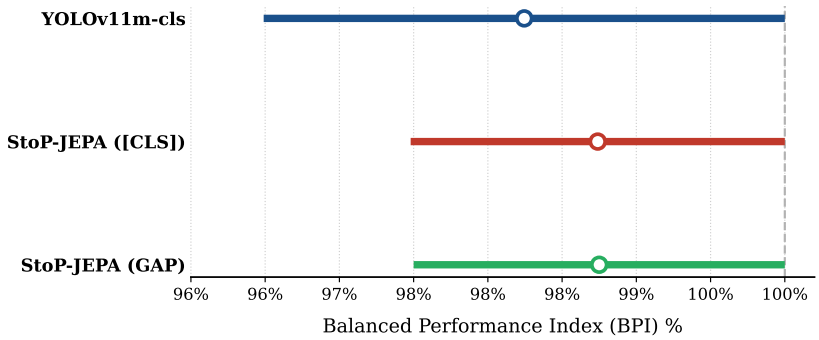


Figure 3: The 95% Confidence Intervals of the Balanced Performance Index for the baseline models on the pristine test dataset.

3.2. Robustness Analysis under Environmental Corruptions

To assess the intrinsic structural resilience of the architectures, the models were subjected to three mathematically defined stress tests. The innate behaviors and vulnerability profiles observed under these extreme OoD conditions can be directly inferred from the underlying architectural paradigms of CNNs and ViTs.

– **Defocus Blur:** Under progressive optical blurring, a drastic low-pass filtering operation that destroys high-frequency spatial gradients, the purely supervised YOLOv11m-cls baseline proved remarkably resilient. At the highest blur setting ($k = 105$), YOLO delivered an F1-score of 94.54% for alligator cracks and 94.37% for linear cracks. By contrast, both StoP-JEPA variants recorded a much more perceptible degradation, dropping to approximately 85.28%–85.69% for alligator cracks and 88.46%–88.86% for linear cracks. This empirical divergence is theoretically sound. YOLO's primary dependence on hierarchical, local convolutional receptive fields heavily biases the model to learn low-frequency spatial hierarchies, making it naturally predisposed to blur resistance. Conversely, the JEPA-based architecture relies heavily on the tokenization of discrete image patches. Severe blurring erases the sharp structural distinctions between these non-overlapping patches, partially crippling the global self-attention mechanism's ability to compute meaningful spatial correlations.

– **Gaussian Noise:** Tests involving extreme Gaussian noise ($\sigma = 120$) revealed the superior predictive stability of the self-supervised StoP-JEPA framework. The JEPA (GAP) configuration, in particular, maintained exceptional F1-scores of 94.79% for alligator cracks and 95.08% for linear cracks, outperforming the YOLO baseline which dropped to 93.44% and 93.34%, respectively. The mathematical justification lies in the structural synergy between patch-based tokenization and the GAP strategy. Unlike YOLO's local convolutions, which are highly sensitive to pixel-level high-frequency variations, JEPA initially projects macro-level pixel patches into single latent tokens. Provided that Gaussian noise comprises independent random variables with a zero mean ($\mu = 0$), this initial spatial aggregation naturally averages out pixel-level thermal noise within each patch. Subsequently, the unweighted arithmetic mean operation of the GAP layer over the entire sequence of tokenized patches acts as a secondary global statistical smoothing filter. According to the Law of Large Numbers, this hierarchical averaging drastically reduces the variance of the stochastic noise. Crucially, this theoretical variance reduction is consistent with the empirical bootstrapping results: the StoP-JEPA (GAP) model exhibits marginally tighter standard deviations compared to YOLO, suggesting that the framework preserves the underlying semantic representations with slightly greater predictive stability.

– **Salt-and-Pepper Noise:** The most revealing architectural vulnerabilities surfaced under extreme impulse noise. At a nominal 90% Salt-and-Pepper corruption, the local convolutional gradients of YOLO suffered heavily, dropping to F1-scores of 60.11% (alligator) and 76.06% (linear). However, a catastrophic classifier collapse occurred within the JEPA (GAP) configuration for linear cracks, plummeting to an abysmal F1-score of 3.82%. As hypothesized in Section 2.6, the unweighted arithmetic averaging of the GAP mechanism forces the model to absorb entirely destroyed, meaningless patches into the global representation, causing a total representation collapse. In stark contrast, the JEPA ([CLS]) configuration, leveraging the dedicated [CLS] token, survived the 90% impulse noise with significantly higher structural integrity, maintaining a 72.80% F1-score for alligator cracks and 53.69% for

linear cracks. Through the dynamic global self-attention layers, the [CLS] token successfully learned to assign near-zero attention weights to the entirely corrupted patches, acting as a highly selective stochastic filter. It dynamically aggregated context exclusively from the scattered surviving uncorrupted patches, thereby experimentally validating the theoretical advantage of attention-based spatial anchoring against severe, unstructured pixel-level corruptions.

3.3. Architectural Inductive Bias and Morphological Recall Asymmetry

A more granular assessment of the performance indices under extreme impulse noise reveals a profound architectural divergence regarding morphological sensitivity. Specifically, under 90% Salt-and-Pepper corruption, a stark asymmetry emerges in the *Recall* metric across the two distinct distress morphologies: the YOLO baseline exhibits a severe bias towards retaining linear cracks, whereas the StoP-JEPA framework is heavily biased towards retaining alligator cracks.

Empirically, YOLO delivers an extraordinary Recall of 95.10% ($\pm$ 2.11) for linear cracks, yet plummets to 45.25% ($\pm$ 5.04) for alligator cracks. Conversely, the StoP-JEPA ([CLS]) configuration demonstrates the exact opposite behavior, maintaining a robust Recall of 91.96% ($\pm$ 2.74) for alligator cracks while suffering a significant drop to 39.80% ($\pm$ 4.78) for the linear class. This asymmetric phenomenon can be fundamentally explained by the intrinsic inductive biases of their respective architectures.

The YOLO architecture is founded on local convolutional receptive fields, which are mathematically optimized to function as localized edge and gradient detectors. Linear cracks inherently possess strong, continuous, and unidirectional local gradients. Even when an image is severely perforated by impulse noise, if a minimal sequence of contiguous pixels belonging to a linear crack survives within a localized kernel window, the convolutional filters remain activated. Consequently, under severe structural pressure, YOLO relies heavily on an inbuilt bias toward geometric linearity, granting the Linear class an exceptionally high Recall.

In stark contrast, alligator cracks lack a single dominant linear edge; rather, they form a widely distributed, multi-directional, and interconnected textural network spanning extensive areas of the pavement. The StoP-JEPA architecture, driven by global self-attention mechanisms, does not explicitly rely on local linear edges. Instead, it aggregates contextual information through a non-local, holistic approach. Under catastrophic pixel destruction, the self-attention mechanism is capable of cross-referencing and accumulating scattered semantic cues from distant surviving patches of the alligator crack network. Therefore, the self-attention mechanism is intrinsically biased toward the holistic contextual reasoning of textural distress patterns, leading to the near-perfect Recall for the Alligator class even at maximal image degradation.

Crucially, this inherent architectural bias also manifests statistically as a severe Precision-Recall trade-off under extreme noise. When visual evidence is largely destroyed, the models default to their structural priors, leading to an over-prediction of their favored morphology. For instance, YOLO's extreme Recall (95.10%) for linear cracks is accompanied by a diminished Precision of 63.48%, as the blinded model assumes most surviving noise artifacts are localized edges. Similarly, StoP-JEPA's high Recall (91.96%) for alligator cracks comes with a Precision drop to 60.36%, as the self-attention mechanism over-aggregates scattered noise into textural patterns. This confirms that the observed Recall asymmetry is actively driven by the models over-predicting the distress type most aligned with their underlying mathematical formulation.

3.4. Diagnostic Symmetry and Balanced Performance Index Assessment

Real-world pavement diagnostics require symmetry and the ability to analyze damage in a holistic manner, as opposed to relying on isolated metrics that might mask specific architectural weak points. An operationally viable ITS cannot champion the detection of one distress morphology at the catastrophic expense of its counterpart. To quantitatively assess this vital classification equilibrium, the BPI, calculated as the harmonic mean of class-wise F1-scores, was cumulatively tracked

Table 1: Robustness Evaluation Results for the **Alligator** Crack Class under Various Environmental Corruptions (Bootstrapping Mean ± Std).

Corruption	Level	YOLOv11m-cls			StoP-JEPA (GAP)			StoP-JEPA (CLS)		
		Pre (%)	Rec (%)	F1 (%)	Pre (%)	Rec (%)	F1 (%)	Pre (%)	Rec (%)	F1 (%)
BLUR	0 (Clean)	98.93 ± 1.06	98.01 ± 1.42	98.46 ± 0.88	99.02 ± 0.96	99.02 ± 0.97	99.01 ± 0.68	98.01 ± 1.39	100.00 ± 0.00	98.99 ± 0.71
	65	93.11 ± 2.49	97.08 ± 1.76	95.02 ± 1.56	100.00 ± 0.00	81.94 ± 3.75	90.02 ± 2.27	98.75 ± 1.23	80.18 ± 3.86	88.45 ± 2.44
	85	93.40 ± 2.47	97.03 ± 1.71	95.16 ± 1.52	100.00 ± 0.00	82.05 ± 3.92	90.09 ± 2.38	98.75 ± 1.22	80.08 ± 3.92	88.39 ± 2.50
	105	93.16 ± 2.53	96.01 ± 2.01	94.54 ± 1.72	100.00 ± 0.00	75.07 ± 4.46	85.69 ± 2.92	98.76 ± 1.25	75.15 ± 4.31	85.28 ± 2.87
SALT & PEPPER	0 (Clean)	98.93 ± 1.06	98.01 ± 1.42	98.46 ± 0.88	99.02 ± 0.96	99.02 ± 0.97	99.01 ± 0.68	98.01 ± 1.39	100.00 ± 0.00	98.99 ± 0.71
	30%	100.00 ± 0.00	86.01 ± 3.47	92.44 ± 2.02	89.32 ± 2.83	100.00 ± 0.00	94.34 ± 1.58	90.10 ± 2.86	100.00 ± 0.00	94.77 ± 1.59
	60%	100.00 ± 0.00	62.63 ± 4.85	76.91 ± 3.69	63.53 ± 3.80	100.00 ± 0.00	77.63 ± 2.85	82.25 ± 3.65	95.96 ± 2.06	88.53 ± 2.35
	90%	90.18 ± 4.26	45.25 ± 5.04	60.11 ± 4.83	50.39 ± 3.53	100.00 ± 0.00	66.94 ± 3.14	60.36 ± 3.98	91.96 ± 2.74	72.80 ± 3.19
GAUSSIAN	0 (Clean)	98.93 ± 1.06	98.01 ± 1.42	98.46 ± 0.88	99.02 ± 0.96	99.02 ± 0.97	99.01 ± 0.68	98.01 ± 1.39	100.00 ± 0.00	98.99 ± 0.71
	$\sigma = 60$	96.11 ± 1.93	100.00 ± 0.00	98.01 ± 1.01	97.94 ± 1.41	98.03 ± 1.40	97.97 ± 1.02	98.07 ± 1.35	100.00 ± 0.00	99.02 ± 0.69
	$\sigma = 90$	92.38 ± 2.68	97.09 ± 1.60	94.65 ± 1.63	98.02 ± 1.40	98.00 ± 1.41	98.00 ± 1.01	96.16 ± 1.85	100.00 ± 0.00	98.03 ± 0.97
	$\sigma = 120$	93.01 ± 2.57	93.94 ± 2.36	93.44 ± 1.79	97.90 ± 1.45	91.92 ± 2.73	94.79 ± 1.61	94.93 ± 2.26	92.94 ± 2.66	93.90 ± 1.85

Table 2: Robustness Evaluation Results for the **Linear** Crack Class under Various Environmental Corruptions (Bootstrapping Mean ± Std).

Corruption	Level	YOLOv11m-cls			StoP-JEPA (GAP)			StoP-JEPA (CLS)		
		Pre (%)	Rec (%)	F1 (%)	Pre (%)	Rec (%)	F1 (%)	Pre (%)	Rec (%)	F1 (%)
BLUR	0 (Clean)	98.01 ± 1.43	98.94 ± 1.04	98.47 ± 0.88	99.02 ± 0.97	99.03 ± 0.95	99.02 ± 0.68	100.00 ± 0.00	97.98 ± 1.41	98.97 ± 0.72
	65	96.94 ± 1.88	92.81 ± 2.63	94.80 ± 1.68	84.71 ± 3.19	100.00 ± 0.00	91.69 ± 1.88	83.22 ± 3.47	98.98 ± 1.01	90.38 ± 2.13
	85	96.91 ± 1.79	93.12 ± 2.60	94.95 ± 1.60	84.82 ± 3.38	100.00 ± 0.00	91.75 ± 1.99	83.24 ± 3.42	98.98 ± 1.00	90.39 ± 2.11
	105	95.90 ± 1.99	92.94 ± 2.58	94.37 ± 1.70	80.02 ± 3.62	100.00 ± 0.00	88.86 ± 2.24	79.98 ± 3.50	99.05 ± 0.95	88.46 ± 2.21
SALT & PEPPER	0 (Clean)	98.01 ± 1.43	98.94 ± 1.04	98.47 ± 0.88	99.02 ± 0.97	99.03 ± 0.95	99.02 ± 0.68	100.00 ± 0.00	97.98 ± 1.41	98.97 ± 0.72
	30%	87.76 ± 3.07	100.00 ± 0.00	93.45 ± 1.75	100.00 ± 0.00	88.08 ± 3.10	93.63 ± 1.76	100.00 ± 0.00	89.06 ± 3.13	94.18 ± 1.76
	60%	72.75 ± 3.73	100.00 ± 0.00	84.17 ± 2.51	100.00 ± 0.00	42.89 ± 5.00	59.86 ± 4.94	95.15 ± 2.45	79.29 ± 4.18	86.43 ± 2.76
	90%	63.48 ± 4.05	95.10 ± 2.11	76.06 ± 3.08	86.30 ± 34.38	1.97 ± 1.40	3.82 ± 2.66	83.26 ± 5.41	39.80 ± 4.78	53.69 ± 4.88
GAUSSIAN	0 (Clean)	98.01 ± 1.43	98.94 ± 1.04	98.47 ± 0.88	99.02 ± 0.97	99.03 ± 0.95	99.02 ± 0.68	100.00 ± 0.00	97.98 ± 1.41	98.97 ± 0.72
	$\sigma = 60$	100.00 ± 0.00	95.95 ± 1.97	97.92 ± 1.03	98.05 ± 1.38	97.96 ± 1.40	98.00 ± 1.01	100.00 ± 0.00	98.04 ± 1.37	99.00 ± 0.70
	$\sigma = 90$	96.93 ± 1.73	91.99 ± 2.77	94.37 ± 1.71	98.00 ± 1.44	98.01 ± 1.41	97.99 ± 1.03	100.00 ± 0.00	96.03 ± 1.90	97.96 ± 0.99
	$\sigma = 120$	93.86 ± 2.40	92.90 ± 2.64	93.34 ± 1.86	92.35 ± 2.52	98.01 ± 1.37	95.08 ± 1.48	93.07 ± 2.58	95.01 ± 2.23	94.00 ± 1.79

across the most extreme environmental stress tests, as visually summarized with 95% Confidence Intervals in Figure 4.

- **Pristine Conditions:** As established in the baseline analysis, both architectures initially exhibited near-perfect diagnostic symmetry. With reference BPI scores firmly exceeding 98%, the complete absence of inherent morphological bias prior to degradation serves as an ideal, neutral baseline for evaluating the subsequent stress tests.
- **Gaussian Noise:** Under severe noise ($\sigma = 120$), the performance degradation remained highly symmetric across all models. The JEPA (GAP) configuration emerged as the most stable architecture, recording a BPI of 94.93% (± 1.51), driven by beautifully balanced F1-scores for alligator (94.79%) and linear

cracks (95.08%). This empirical balance firmly demonstrates that unweighted global averaging acts as an inherently unbiased statistical smoother against zero-mean stochastic noise, preserving the diagnostic equilibrium. YOLO also maintained its symmetry, albeit at a marginally lower BPI of 93.39% (± 1.77).

- **Defocus Blur:** Under extreme optical degradation ($k = 105$), the models manifested a different performance tier, yet the symmetry of degradation was preserved. YOLO heavily leveraged its architectural merit to secure a staggering BPI of 94.45% (± 1.66), guaranteeing the structural topology of both crack types equally. In contrast, while the StoP-JEPA configurations lost precision under extreme blur, they registered a symmetric drop, yielding BPI scores of 87.23% (± 2.45) for GAP and 86.83% (± 2.40) for [CLS]. Because the global self-attention mechanism failed equally for both crack types, the harmonic mean did not impose a disproportionate penalty, indicating that the architecture degrades gracefully rather than selectively under optical blur.
- **Salt-and-Pepper Noise:** The true discriminative power of the BPI metric surfaced under extreme impulse noise (90% intensity), where severe asymmetric collapse occurred. The JEPA (GAP) configuration was critically penalized under random pixel annihilation. while its F1-score for alligator cracks stood at 66.94%, its performance for linear cracks collapsed miserably to 3.82%. Because the BPI relies on the harmonic mean, it heavily penalizes such extreme disparity, resulting in a disastrously low BPI of 7.06% (± 4.67). This clearly indicates that unweighted global pooling struggles to maintain a symmetric topological understanding under unstructured spatial loss.

  In stark contrast, the YOLO baseline maintained commendable symmetry (BPI: 67.07% ± 3.80). Most importantly, the selective, attention-driven [CLS] token effectively mitigated the asymmetric collapse experienced by GAP. It achieved F1-scores of 72.80% and 53.69%, resulting in a highly competitive BPI of 61.71% (± 3.94).

This all-encompassing BPI analysis provides compelling quantitative evidence that while Gaussian noise and optical blur equivalently produce symmetric performance degradation, drastic impulse noise trig-

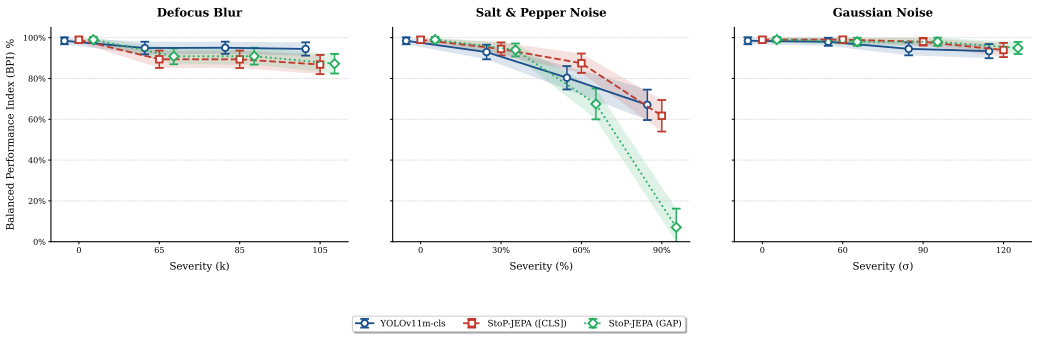


Figure 4: Comprehensive 95% Confidence Interval (CI) analysis of the Balanced Performance Index (BPI) across progressive intensities of Defocus Blur, Salt & Pepper, and Gaussian Noise. The error bars and shaded regions represent the 95% CI.

Table 3: Detailed Balanced Performance Index (BPI) Analysis (Bootstrapping Mean ± Std).

Corruption	Level	YOLOv11m-cls			StoP-JEPA (GAP)			StoP-JEPA (CLS)		
		Alligator	Linear	BPI	Alligator	Linear	BPI	Alligator	Linear	BPI
Clean	0	98.46 ± 0.88	98.47 ± 0.88	98.46 ± 0.87	99.01 ± 0.68	99.02 ± 0.68	99.02 ± 0.67	98.99 ± 0.71	98.97 ± 0.72	98.98 ± 0.71
Blur	65	95.02 ± 1.56	94.80 ± 1.68	94.91 ± 1.58	90.02 ± 2.27	91.69 ± 1.88	90.84 ± 1.98	88.45 ± 2.44	90.38 ± 2.13	89.40 ± 2.17
	85	95.16 ± 1.52	94.95 ± 1.60	95.05 ± 1.52	90.09 ± 2.38	91.75 ± 1.99	90.91 ± 2.11	88.39 ± 2.50	90.39 ± 2.11	89.37 ± 2.19
	105	94.54 ± 1.72	94.37 ± 1.70	94.45 ± 1.66	85.69 ± 2.92	88.86 ± 2.24	87.23 ± 2.45	85.28 ± 2.87	88.46 ± 2.21	86.83 ± 2.40
Salt & Pepper	30%	92.44 ± 2.02	93.45 ± 1.75	92.94 ± 1.82	94.34 ± 1.58	93.63 ± 1.76	93.98 ± 1.61	94.77 ± 1.59	94.18 ± 1.76	94.47 ± 1.63
	60%	76.91 ± 3.69	84.17 ± 2.51	80.35 ± 2.92	77.63 ± 2.85	59.86 ± 4.94	67.51 ± 3.85	88.53 ± 2.35	86.43 ± 2.76	87.46 ± 2.43
	90%	60.11 ± 4.83	76.06 ± 3.08	67.07 ± 3.80	66.94 ± 3.14	3.82 ± 2.66	7.06 ± 4.67	72.80 ± 3.19	53.69 ± 4.88	61.71 ± 3.94
Gaussian	$\sigma = 60$	98.01 ± 1.01	97.92 ± 1.03	97.97 ± 1.01	97.97 ± 1.02	98.00 ± 1.01	97.98 ± 1.01	99.02 ± 0.69	99.00 ± 0.70	99.01 ± 0.69
	$\sigma = 90$	94.65 ± 1.63	94.37 ± 1.71	94.51 ± 1.62	98.00 ± 1.01	97.99 ± 1.03	98.00 ± 1.01	98.03 ± 0.97	97.96 ± 0.99	98.00 ± 0.97
	$\sigma = 120$	93.44 ± 1.79	93.34 ± 1.86	93.39 ± 1.77	94.79 ± 1.61	95.08 ± 1.48	94.93 ± 1.51	93.90 ± 1.85	94.00 ± 1.79	93.95 ± 1.77

gers a catastrophic, contrapuntal asymmetry in standard global pooling practices. The proposed [CLS] attention mechanism is shown to empirically mitigate this vulnerability, providing a more robust and reliable diagnostic equilibrium under highly destructive circumstances.

4. Limitations and Future Work

While this study provides compelling quantitative evidence regarding the structural resilience and inductive biases of fundamentally distinct vision paradigms, several limitations must be acknowledged to contextualize the findings and guide future research.

First, the morphological scope of the dataset is constrained to two primary distress categories. linear and alligator cracks. While these represent fundamental and highly contrasting structural topologies (localized edges vs. networked textures), the generalizability of the observed architectural behaviors to other complex pavement distresses such as raveling, block cracking, or potholes warrants further empirical validation.

Second, the robustness assessment was intentionally focused on three specific synthetic corruptions namely Gaussian noise, Salt-and-Pepper noise, and Defocus Blur. Although these effectively simulate critical real-world sensor anomalies, they do not encompass the full spectrum of environmental challenges encountered by vehicle-mounted In ITS.

Consequently, future research must extend this evaluation framework to encompass dynamic and meteorological perturbations. Crucial unexamined factors include motion blur induced by high-speed data acquisition, surface occlusions, and adverse weather conditions such as rain or surface moisture, which profoundly alter the reflective properties of the pavement.

5. Conclusions

This study conducted a rigorous robustness evaluation of automated pavement distress classification models, contrasting a supervised baseline (YOLOv11m-cls) with a self-supervised Vision Transformer framework (StoP-JEPA). To move beyond the overestimation of performance on pristine datasets, the architectures were subjected to extreme environmental stress tests, including Defocus Blur, Gaussian Sensor Noise, and Salt-and-Pepper Impulse Corruptions. Furthermore, the Balanced Performance Index (BPI) was introduced to mathematically quantify the diagnostic equilibrium between fundamentally distinct distress topologies (linear versus alligator cracks).

Our empirical analysis provides compelling evidence that under severe visual degradation, classification performance is heavily dictated by the underlying architectural inductive biases. Specifically, standard global pooling (GAP) in Vision Transformers suffers a catastrophic, asymmetric collapse under unstructured impulse noise, defaulting to widespread

textural aggregation. In stark contrast, the proposed attention-driven [CLS] mechanism effectively mitigates this vulnerability. By selectively routing semantic context from surviving uncorrupted patches, the [CLS] token maintained a resilient and balanced performance (BPI = 61.71%), comparable to the structural stability of the supervised CNN baseline.

For practical deployment ITS, this study reveals a clear architectural trade-off. Specifically, the StoP-JEPA framework yields the most statistically stable diagnostic equilibrium under severe Gaussian noise, whereas the local convolutions of YOLO act as natural low-pass filters, maintaining superior structural topology under extreme optical blur. Ultimately, achieving reliable real-world infrastructure monitoring requires selecting architectural paradigms whose intrinsic mathematical priors align with the specific environmental perturbations of the deployment domain.

References

[1] Abdelwahed, S.H., Sharobim, B.K., Wasfey, B., Said, L.A., 2025. Advancements in real-time road damage detection: a comprehensive survey of methodologies and datasets. Journal of Real-Time Image Processing 22, 1–13.

[2] 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.

[3] 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`.

[4] Hendrycks, D., Dietterich, T., 2019. Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261 .

[5] Hendrycks, D., Mazeika, M., Kadavath, S., Song, D., 2019. Using self-supervised learning can improve model robustness and uncertainty. Advances in neural information processing systems 32.

[6] Hoang, N.D., Nguyen, Q.L., 2019. A novel method for asphalt pavement crack classification based on image processing and machine learning. Engineering with Computers 35, 487–498.

[7] Liao, S., 2024. Road damage detection algorithm based on optimised you only look once version 8, in: 2024 5th International Conference on Computer Engineering and Application (ICCEA), IEEE. pp. 1381–1384.

[8] Liu, J., Wang, Z., Ma, L., Fang, C., Bai, T., Zhang, X., Liu, J., Chen, Z., 2024. Benchmarking object detection robustness against real-world corruptions. International Journal of Computer Vision 132, 4398–4416.

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

[10] Monemi, M., C.M.R.M.B.M., Latva-aho, M., 2025. Tutorial on joint embedding predictive architectures (jepa): Foundations, applications, and future directions. .

[11] Rodriguez-Rodriguez, J.A., López-Rubio, E., Ángel-Ruiz, J.A., Molina-Cabello, M.A., 2024. The impact of noise and brightness on object detection methods. Sensors 24, 821.

[12] 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.

[13] Zakeri, H., Nejad, F.M., Fahimifar, A., 2017. Image based techniques for crack detection, classification and quantification in asphalt pavement: a review. Archives of Computational Methods in Engineering 24, 935–977.

[14] 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.

[15] Zhou, S., Canchila, C., Song, W., 2023. Deep learning-based crack segmentation for civil infrastructure: Data types, architectures, and benchmarked performance. Automation in Construction 146, 104678.

[16] Zhu, A., Wang, B., Xie, J., Ma, C., 2023. Mff-yolo: an accurate model for detecting tunnel defects based on multi-scale feature fusion. Sensors 23, 6490.