# CROSS-DOMAIN ROAD DAMAGE CLASSIFICATION USING REGULARIZED SELF-SUPERVISED REPRESENTATION LEARNING

*D.V. Agrawal*[*], *V. Gupta*[†], *C.R. Krishna*[*]

**Abstract:** Road surface abrasions significantly contribute to vehicle collisions and mechanical failures worldwide. Traditional machine learning-based methods for road damage detection typically rely heavily on extensive manual annotations, making them costly, labour-intensive, and inefficient. To address this challenge, this paper proposes a label-efficient image processing framework based on self-supervised representation learning for road damage classification. Our approach integrates contrastive learning with a regularized redundancy reduction method, enabling the extraction of rich, discriminative features directly from unlabelled data. Contrastive learning separates positive and negative samples to learn robust feature representations, while a cross-correlation loss maximizes information content by minimizing redundancy. Regularization through variance and covariance loss terms ensures feature diversity and prevents informational collapse in the learned representations. Extensive evaluations in both in-domain and cross-domain scenarios demonstrate that our proposed method achieves superior performance compared to supervised techniques, even when trained with substantially fewer labelled samples. Thus, this work provides an effective, economical, and scalable solution to the critical challenges faced in automated road maintenance. The downstream task considered in this study is multi-class classification of road damage categories rather than binary damaged-versus-undamaged road detection.

Key words: *road damage classification, image processing, self-supervised learning, contrastive learning, domain adaptation, variance and covariance regularization, road maintenance automation*

*Received: April 10, 2025*
*Revised and accepted: August 18, 2025*

**DOI:** 10.14311/NNW.2025.35.004

[*] Deepika Vikas Agrawal – Corresponding author; C. Rama Krishna; National Institute of Technical Teachers Training and Research, Chandigarh, India, E-mail: deepika.cse22@nitttrchd.ac.in, rkc@nitttrchd.ac.in

[†] Varun Gupta; Chandigarh College of Engineering and Technology, Chandigarh, India, E-mail: varungupta@ccet.ac.in

# 1. Introduction

The roadway system is a cornerstone of national transportation infrastructure, playing a critical role in economic growth and public safety. Effective road maintenance is essential for ensuring community safety, optimizing transportation efficiency, and minimizing economic impacts. Without proactive management, road surface anomalies such as potholes and cracks. Failure to promptly detect and repair these defects can lead to severe vehicle accidents, substantial property damage, increased fuel consumption, and considerable economic losses. Indeed, suboptimal road conditions globally result in significant financial and human costs. For instance, in 2022, various nations allocated billions of euros toward road infrastructure repairs driven by inadequate maintenance practices [13]. In the United States alone, motorists incurred approximately \$15 billion in vehicle repair costs due to poor road conditions [7, 24]. According to the World Health Organization (WHO), road traffic accidents claim approximately 1.19 million lives annually and cause injuries in between 20 to 50 million individuals, many resulting in permanent disabilities. Such incidents remain the leading cause of death for individuals aged 5–29, with annual fatalities expected to rise to 2.4 million by 2030, making road traffic the fifth leading cause of death globally [23, 30, 31].

Despite advancements in road infrastructure technology, traditional road inspection methods predominantly rely on manual inspection processes. These methods are inherently slow, labour-intensive, subjective, and prone to human error, significantly delaying defect detection and repair. Consequently, they contribute to increased maintenance costs and compromise public safety. Recent advancements in machine learning offer promising opportunities for automating road damage detection, potentially enhancing operational efficiency, reducing costs, and improving road safety. Nevertheless, contemporary supervised machine learning methods require substantial amounts of annotated data, the collection of which remains expensive and laborious, thereby limiting their widespread implementation.

To bridge this critical gap, we propose a novel self-supervised learning framework specifically designed for efficient road damage classification. Self-supervised learning techniques generate pseudo-labels directly from unlabelled data, enabling effective representation learning without requiring extensive manual annotations [27]. Typically, this process involves two phases: a pretext task, where models learn meaningful representations from pseudo-labeled data, and a downstream task, where these representations are fine-tuned for specific classification objectives. Notable approaches in self-supervised learning include contrastive learning methods such as SimCLR [9], MoCo [18], and redundancy reduction techniques like Barlow Twins [32] and VICReg [5]. While contrastive learning focuses on differentiating similar from dissimilar data pairs to create robust feature embeddings, redundancy reduction emphasizes the maximization of information by minimizing redundant features within the learned representations.

In this work, we focus on label-efficient classification of road damage categories using benchmark datasets containing annotated damaged-road samples. The employed datasets do not include a separate no-damage or clear-road class. Therefore, the present study does not address binary damaged-versus-undamaged road detection, but rather classification among damage categories. This scope is particularly relevant whether self-supervised pretraining can improve downstream category-level discrimination and cross-domain transfer under limited-label settings.

## 1.1 Contributions

This study introduces an innovative regularized redundancy reduction contrastive learning framework that synergistically combines the strengths of contrastive learning and redundancy reduction strategies. Specifically, our approach uses contrastive learning to differentiate between positive and negative samples, complemented by a cross-correlation loss to minimize redundancy and enhance information content. Regularization techniques, including variance and covariance loss terms, further enrich the learned representations by preventing norm collapse and ensuring informational diversity. The primary contributions of this research include:

1. Proposing a novel self-supervised regularized redundancy reduction contrastive learning framework, significantly reducing the dependence on labelled data and enabling cost-effective and scalable solutions for automated road damage classification.

2. Utilizing contrastive learning at the core of the methodology to produce robust, generalizable feature representations from unlabelled datasets, substantially lowering annotation requirements.

3. Introducing a cross-correlation loss function to optimize the correlation structure within embeddings, encouraging high intra-dimensional correlations while minimizing inter-dimensional redundancy.

4. Implementing variance and covariance regularization strategies to prevent norm and informational collapse, thereby ensuring the richness and diversity of learned feature representations.

5. Demonstrating superior performance over existing supervised and state-of-the-art approaches, even when trained on significantly reduced labelled datasets (20% or 50%).

6. Establishing strong domain adaptability by successfully evaluating the proposed method in cross-domain contexts, highlighting its capacity for robust knowledge transfer across distinct datasets.

The remainder of the paper is structured as follows: Section 2 reviews existing literature on road damage classification methodologies. Section 3 presents a detailed description of the proposed framework. Section 4 introduces the datasets used in this research. Section 5 discusses the experimental setup and the comprehensive analysis of the obtained results. Finally, Section 6 concludes the paper with a summary of the findings and directions for future research.

## 2. Related Work

Recent advancements in machine learning and deep learning have significantly improved automated road damage detection. Broadly, these methods can be categorized into three main approaches: traditional machine learning (ML), custom-designed convolutional neural networks (CNN), and pre-trained network-based methods. Although these methods show promising results, they have limitations, such as high dependence on labelled datasets, insufficient generalizability, and suboptimal performance across domains. Our research addresses these critical issues.

### 2.1 Traditional Machine Learning-Based Approaches

Traditional ML techniques primarily utilize manually crafted features and classifiers for road anomaly detection. Zhang et al. [15] employed a support vector machine (SVM) [19] to classify road cracks into multiple categories using a self-curated dataset comprising 250 images. Egaji et al. [14] evaluated various ML algorithms-including naïve Bayes, $k$-nearest neighbour (KNN), logistic regression, SVM, and random forest tree for pothole detection, relying on their collected dataset. Similarly, Carlos et al. [8] developed a pothole and speed bump detection method employing SVM and random forest classifiers on smartphone-based collected data. Furthermore, Bhatlawande et al. [6] proposed a comprehensive framework to classify various road irregularities, assessing several classifiers (SVM, KNN, naïve Bayes, logistic regression, and random forest), ultimately identifying random forest as the most reliable classifier for pothole detection.

### 2.2 Custom CNN-Based Methods

Custom-designed CNN architecture has demonstrated substantial effectiveness in road damage detection tasks. Patra et al. [28] introduced a CNN-based pothole detection system trained on a dataset of 3,424 real-world images, successfully deploying their model on an Android application scenario. Chu et al. [11] presented a CNN-based model trained on a dataset collected from the roads of Lahore, effectively identifying cracks. Zhang et al. [33] further advanced CNN approaches by customizing a ConvNext-based model consisting of 55 layers, achieving high accuracy in classifying road cracks using a dataset comprising 1,600 grayscale images.

### 2.3 Pretrained Networks-Based Methods

Leveraging pre-trained deep learning architectures has become popular due to their robust performance and reduced training effort. Aparna et al. [3] proposed a ResNet-based pothole detection framework utilizing thermal imagery, demonstrating effectiveness under challenging environmental conditions. Li et al. [26] employed an EfficientNet architecture to detect potholes on the MegaDepth and ConTrack datasets, showcasing notable efficiency and accuracy. Additionally, Cinar et al. [12] utilized a DenseNet121 [21] based model for the pothole detection task; however, their testing was performed on a very small dataset (67 images). Li et al. [25] used a self-collected dataset covering six types of road distresses, including potholes, cracks, and depressions. They compared the performance of ResNet-50 [17], ResNet-34, VGG-19 and VGG-16 [29] models on the dataset, where the ResNet-50 achieved the best results for both binary and multi-class classification tasks. Jana et al. [22] introduced a pothole detection approach based on merging data from self-acquisition and 'road-traversing knowledge' datasets to improve detection accuracy. They utilize seven pre-trained networks named ResNet-50 [17], ResNet-34, VGG-19 and VGG-16 [29], InceptionV3, DenseNet121, Xception [10], SEResNet-50 (squeeze and excitation ResNet) [20], EfficientNet-B3 [1] out of which DenseNet121, Xception, SEResNet-50 performed well. Gupta et al. [16] achieved excellent performance on multiple damage classes by applying a ResNet-based algorithm to the road damage 2020 (RDD2020) [4]. Ahmad et al. [2] classified pavement classes with MobileNet v2, ResNet50, and ResNet18; datasets were collected from the internet and taken with images of Pakistani roads.

### 2.4 Research Gaps

Despite promising developments, existing road damage detection methods encounter significant limitations, namely heavy reliance on extensive labelled datasets, inadequate performance, and limited generalization across different datasets. To overcome these constraints, this study introduces a label-efficient, self-supervised representation learning framework for road damage classification. The proposed methodology employs a novel regularized redundancy reduction contrastive learning framework, effectively generating high-quality, semantically rich representations from unlabeled images. This strategy significantly reduces the reliance on extensive labelled data, leading to robust performance even with limited annotations. Moreover, comprehensive cross-domain evaluations were conducted on two benchmark datasets – "Road crack classification" and "Road defects images" to validate the enhanced generalization capabilities and superior performance of the proposed model. Consequently, this approach not only substantially reduces data annotation costs but also provides a practical, scalable, and efficient solution for automated road damage assessment in real-world applications.

## 3. Proposed Approach

We propose a self-supervised representation learning approach for learning image representations for road damage classification, integrating regularized redundancy reduction with contrastive learning. This approach minimizes redundant information in the learned representations while maximizing their informational content. By leveraging contrastive learning, which distinguishes between similar and dissimilar data points, our approach effectively learns consistent image representations from unlabelled data. Additionally, our approach calculates the cross-correlation matrix between image embeddings of augmented views processed by twin networks and adjusts it to approximate the identity matrix. This process ensures that features learned from augmented views are similar while reducing redundancy among embedding dimensions. To further refine the learning process, we incorporate two regularization terms into the loss function: variance and covariance. The variance term maintains the variance along each dimension of the embedded vectors above a certain threshold, preventing norm collapse. The covariance term, on the other

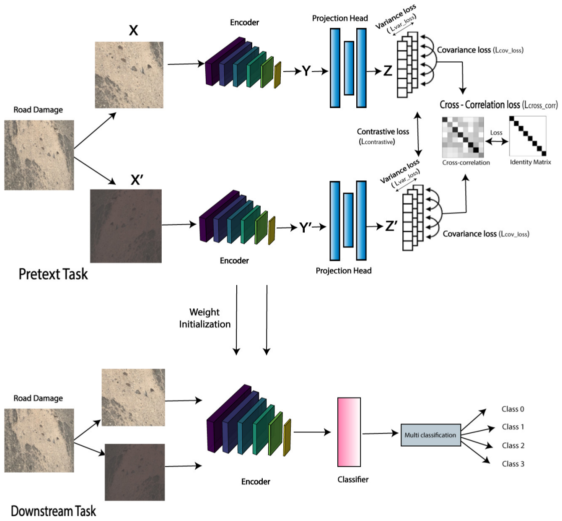


**Fig. 1** *Overview of the proposed self-supervised road damage classification framework combining contrastive learning, redundancy reduction, and variance-covariance regularization.*

hand, serves to decorrelate different embedding dimensions, preventing informational collapse—meaning that each feature dimension will encode its own separate useful information. In essence, contrastive learning enables models to acquire robust image representations irrespective of input variations, while redundancy reduction and regularization terms ensure these features are both informative and non-redundant. In short, we propose this approach where contrastive learning is combined with redundancy reduction and stable regularization. Fig. 1 represents the proposed approach for road damage classification.

The proposed approach consists of a two-phase training framework, which is illustrated in Fig. 1. The first phase, referred to as the pretext task, is dedicated to learning robust representations using a self-supervised representation learning approach. It allows the model to learn representations of the data without the need for labelled annotations. In the second phase, called the downstream task, learned representations are utilized to classify road-wise damage efficiently. Both phases are described in detail in the following sections.

## 3.1 Pretext Task

In the first instance of an unsupervised "pretext" phase, the model is trained solely on unlabelled data and processes the data to reproduce meaningful image representations. To do this, we utilize many data augmentations (resizing, horizontal and vertical flips, jittering, greyscale, gaussian blurring and affine) that produce two different versions of the same image. In this stage, we use a pre-trained ResNet-50 backbone and a projection head (consisting of two fully connected layers) to project input images to a latent embedding space. Such representations significantly compress vital visual knowledge, setting up a proper abstract basis for the subsequent task. We then use these learned embeddings in the next phase of supervised fine-tuning, where we perform road damage classification. As a result, fewer annotated samples are needed to obtain high performance, lowering the effort and resource consumption of labelling.

Our loss function combines four components: contrastive loss [9], cross-correlation loss [32], variance loss [5], and covariance loss [5] to make sure the model learns both discriminative and regularized representations. By optimizing these components jointly, the model achieves a strong generalization ability for future tasks. We define the overall loss function as in Eq. (1):

$$\begin{aligned} L_{\text{total}}(Z, Z') = &\lambda_1 \cdot L_{\text{cont}}(Z, Z') + \lambda_2 \cdot L_{\text{cross}}(Z, Z') \\ &+ \lambda_3 \cdot L_{\text{var}}(Z, Z') + \lambda_4 \cdot L_{\text{cov}}(Z, Z'). \end{aligned} \tag{1}$$

We set $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = 25$ and $\lambda_4 = 1$, for the values of our experiments, as suggested by literature [5,9,32]. The first loss term, the contrastive loss, induces the model to minimize the distance between the matching (positive) samples and maximize the distance between the dissimilar (negative) samples from each other in the embedding space. It is defined as Eq. (2):

$$L_{\text{cont}}(Z, Z') = -\log \frac{\exp\left(\frac{sim(z_i, z_j)}{\tau}\right)}{\sum_{k=1}^{2n} 1_{k \neq i} \exp(sim(z_i z_k / \tau))}. \tag{2}$$

In this loss function, $sim(z_i, z_j)$ is the cosine similarity between the two embeddings $z_i$ and $z_j$. Furthermore, the loss computes the negative logarithm of the exponential ratio of the similarity between positive pairs to the similarity between all pairs $\tau$ is the temperature parameter, and $2n$ is the number of augmented samples in total (two for each original image) where the indicator function $1_{k \neq i}$ excludes $i$ from the denominator. The second loss term, i.e., cross-correlation loss [9], correlates the representations of multiple augmented views from the same image. It is formulated in Eq. (3):

$$L_{\text{cross}}(Z, Z') = \sum_i (1 - C_{ii})^2 + \lambda \cdot \sum_i \sum_{j \neq i} C_{ij}^2. \tag{3}$$

Here, $C$ denotes the cross-correlation matrix of the embeddings from the augmented views. The first component $(\sum_i (1 - C_{ii})^2)$ promotes the diagonal entries of $C$ to be equal to the identity matrix while $\sum_i \sum_{j \neq i} C_{ij}^2$ penalizes off-diagonal correlations. $\lambda$ is a hyperparameter that controls the trade-off between the two terms of the loss function. The value of $\lambda$ has been set to $\lambda = 5 \cdot 10^{-3}$, as mentioned in the literature [32].

The third loss term, i.e., variance regularization term, ensures that each embedding dimension has at least some variability, preventing features from collapsing along a single dimension. For a batch of embeddings, let $z^j$ represents all values of the $j$th dimension in the batch. The loss is defined as Eq. (4):

$$L_{var}(Z, Z') = \frac{1}{d} \sum_{j=1}^{d} \max\left(0, \gamma - std\left(\sqrt{var(z^j) + \epsilon}\right)\right), \tag{4}$$

where

$d$ is the dimensionality of the embedding vectors.

$z^j$ represents all values of the $j$th dimension across the batch $Z$.

$\gamma$ is the target standard deviation (which can be set to be 1).

$std\left(\sqrt{var(z^j)+\epsilon}\right)$ is the regularized standard deviation.

This loss term encourages the variance of embeddings along each dimension to be at least $\gamma$. It ensures that only dimensions where the standard deviation is less than $\gamma$ contribute to the loss. If the standard deviation is already above $\gamma$, the term becomes zero, and no penalty is applied [5].

The covariance loss encourages statistical independence among different embedding dimensions to prevent informational collapse. We compute the covariance matrix $C(Z)$ of the embedding $Z$ as in Eq. (5):

$$C(Z) = \frac{1}{n-1} \sum_{i=1}^{n} (z_i - \overline{z}\ )\ (z_i -\ \overline{z}\ )^{\mathrm{T}}, \tag{5}$$

where

$$\overline{z} = \frac{1}{n} \sum_{i=1}^{n} z_i.$$

The covariance loss focuses on the off-diagonal elements of $C(Z)$ as in Eq. (6):

$$L_{\mathrm{cov}}(Z, Z') = \frac{1}{d} \sum_{i \neq j} [C(Z)]_{i,j}^2, \tag{6}$$

where $n$ is the number of samples in the batch, and $[C(Z)]_{i,j}$ is the $i, j$th entry of covariance matrix. By minimizing these off-diagonal elements, the different features in a representation remain uncorrelated, thereby preventing informational collapse.

During the pretext stage, all available images are used in an unlabeled manner for representation learning, irrespective of their downstream category labels.

## 3.2 Downstream Task

Section 3.2 focuses on taking a network, i.e., ResNet50, that has already been trained in a self-supervised manner using a pretext task and adapting it to perform a supervised downstream classification task—in this case, classifying images of road damage. After the model has learned generic and useful features during the pretext phase, it is adapted for the specific classification problem by replacing the projection head with a new prediction head consisting of two fully connected layers. This simple architectural adjustment, coupled with fine-tuning, allows the model to leverage the previously learned representations and rapidly achieve high performance on the downstream classification task, even with a smaller labelled dataset.

During the pretext task, two augmented views of the same input image are passed through two identical encoder branches with shared weights. These twin branches are used only to compute the self-supervised objective. After pretraining, a single encoder initialized with the learned weights is retained for the downstream task. The projection head during pretraining is discarded and replaced with a task-specific classifier composed of two fully connected layers.

The pretext stage is fully self-supervised and does not use class labels. Class labels are used only in the downstream stage. In the fine-tuning setting, the pretrained encoder and classifier are jointly optimized using labelled training images. In the feature extraction setting, the pretrained encoder is frozen and only the classifier is trained using labelled training images.

The pseudocode for the proposed approach is shown as Algorithm 1.

**Algorithm 1** Pseudocode (PyTorch style).

```
# Training loop
for epoch in range(no_epoch): do
  loss = 0.0
  for data in train_dataloader: do
    # Extract two different augmentations and pass both augmentations
    to get projections
    x_i, x_j = extract_views_from(data)
    z_i = model(x_i)
    z_j = model(x_j)
    # Contrastive Loss
    sim_matrix = torch.matmul(z_i, z_j.T)
    labels = torch.arange(z_i.size(0)).cuda()
    cont_loss = F.cross_entropy(sim_matrix / self.temperature, labels)
    lmbda = 5e-3
    bs = z_i.size(0)
    emb = z_i.size(1)
    z_iNorm = (z_i - z_i.mean(0)) / z_i.std(0)
    z_jNorm = (z_j - z_j.mean(0)) / z_j.std(0)
    # Cross Correlation Loss
    crossCorMat = (z_iNorm.T  z_jNorm) / bs
    # Extract on-diagonal and off-diagonal elements from the
    cross-correlation matrix
    on_diag = torch.diagonal(crossCorMat).add_(-1).pow_(2).sum()
    n, m = crossCorMat.shape
    off_diagonal = crossCorMat.flatten()[:-1].view(n - 1, n + 1)[:, 1:].flatten()
    off_diag = off_diagonal.pow_(2).sum()
    cross_corr_loss = on_diag + lmbda * off_diag
    # Variance Loss
    x = z_i - z_i.mean(dim=0)
    y = z_j - z_j.mean(dim=0)
    var_x = torch.sqrt(x.var(dim=0) + 0.0001)
    var_y = torch.sqrt(y.var(dim=0) + 0.0001)
    var_loss = torch.mean(F.relu(1 - var_x)) / 2
                + torch.mean(F.relu(1 - var_y)) / 2
    # Covariance loss
    cov_x = (x.T  x) / (bs - 1) # Covariance matrix for z_i
    cov_y = (y.T  y) / (bs - 1) # Covariance matrix for z_j
```

```
        cov_loss = get_off_diagonal(cov_x).pow_(2).sum().div(emb)
                  + get_off_diagonal(cov_y).pow_(2).sum(). div(emb)
        # Combined loss and optimizer
        lmbda1=1, lmbda2=1, lmbda3=25, lmbda4=1
        combined_loss = lmbda1 * cont_loss + lmbda2 * cross_corr_loss
                  + lmbda3 * cov_loss + lmbda4 * var_loss
        optimizer.zero_grad()
        combined_loss.backward()
        optimizer.step()
    end for
end for
```

## 4. Dataset Description

In this subsection, we briefly describe the road damage datasets used in this work. Their main characteristics are summarized in Tab. I.

| Dataset | Total images | Train images | Test images | No. of classes |
|---|---|---|---|---|
| Road crack classification | 1600 | 1280 | 320 | 4 |
| Road defects images | 400 | 320 | 80 | 4 |

**Tab. I** *Dataset details.*

Details for each of these datasets are presented below:

1) Road crack classification dataset[1] [33]: Open-source dataset for training and testing road crack classifiers. The dataset, which was cropped from the DSPS (Data Science for Pavement Symposium) dataset published by the Federal Highway Administration (FHWA), contains four scenarios of road cracks: block, longitudinal, alligator and transverse. Moreover, as it offers a wide range of categorization, this dataset is often used to benchmark deep learning and computer vision models for various road crack analysis tasks.

2) Road defects images dataset[2]: The dataset, obtained utilizing mobile cameras, is intended for creating computer vision models that can be employed for detecting defects in infrastructure like roads, bridges, and buildings automatically. It consists of 400 images 100 each for cracks, potholes, patches and surface defects—taken in different light and environmental scenarios. The core of this streamlines laborious manual inspections and improves accuracy, efficiency, and cost in the maintenance and safety of infrastructure.

It should be noted that both benchmark datasets used in this study contain labelled categories of road damage and do not include a separate "no-damage" or

[1] https://github.com/tjboise/RCCD/tree/main/crackdata

[2] https://www.kaggle.com/datasets/patelmihir/road-defects-nonaugmented?select=Surface_Defects

"clear-road" class. Specifically, the Road crack classification dataset comprises four crack categories, while the Road defects images dataset contains four defect categories: cracks, potholes, patches, and surface defects. Therefore, the downstream task in this work is formulated as multi-class road damage classification among damaged-road categories rather than binary damaged-versus-undamaged road detection. Consequently, the proposed framework is not evaluated for distinguishing healthy road surface from damaged ones.

## 5. Experimentation Details

This section discusses the design and implementation of experiments to verify the proposed method for road damage classification. Two independent databases are the Road crack classification dataset and the Road defects images dataset. We implement it in Python using the open-source PyTorch library and run it on a workstation with NVIDIA GeForce RTX 4090 GPU (24 GB memory). The training process is split into two stages.

The model is trained in two stages; in the first phase (pretext), it is trained on unlabeled data using the proposed approach. In the second (downstream) phase, we evaluate the learned representations using labelled data. We explore two primary methods of evaluation:

1. Fine-tuning: We train both the backbone (encoder) and the prediction head (two fully connected layers) on labelled data.
2. Feature Extraction: The backbone is left frozen, and only the prediction head (two fully connected layers) is trained on the labelled data. In doing so, we emphasize the advantages of learning flexible, reusable representations in a self-supervised manner rather than optimizing the full network from scratch.

We measure performance under varying annotation scenarios using 100%, 50% and 20% of the available labelled samples. When applying the entire dataset (100%), we train using 128 batch size, while for the 50% and 20%, the batch size is reduced to 8 as the number of remaining images in the dataset are less.

We adopt three primary experimental settings to assess our approach thoroughly:

**Experiment 1** Evaluation of the proposed approach on the Road crack classification dataset (within-dataset evaluation): The pretext training and downstream task are performed on the Road crack classification dataset. The model is first trained in a self-supervised way, and the downstream task (classification) is performed on the same dataset using both evaluation methods (fine-tuning and feature extraction).

**Experiment 2** Evaluation of the proposed approach on the Road defect images dataset (within-dataset evaluation): The pretext and downstream training phases are both conducted using the Road defect images dataset. Initially, the model undergoes self-supervised training, followed by evaluation on the same dataset using two distinct approaches: fine-tuning and feature extraction.

**Experiment 3** Generalization across datasets: In this experiment, the model is pre-trained (self-supervised) on one dataset and fine-tuned on another dataset (downstream) and vice-versa. This configuration tests the model's capacity to transfer knowledge learned from road damage data of various types and distributions.

## 5.1 Hyperparameter Optimization

Throughout these experiments, extensive tuning of hyperparameters was carried out, involving testing multiple values. After thorough experimentation, the best-performing configurations were finalized for the experiment. The hyperparameter and augmentation settings for the experiments are given in Tab. II.

| Parameters | Pretext task | Downstream task |
|---|---|---|
| Batch size | 128 | 128 (100% data),<br>8 (20% and 50% data) |
| Optimization algorithm | Adaptive moment<br>estimation (ADAM) | Stochastic gradient<br>descent (SGD) |
| Initial learning rate | 0.000001 | 0.001 |
| Regularization | 0.001 | 0.0005 |
| Horizontal flip probability | 0.5 | 0.5 |
| Grayscale transformation | 0.2 | 0.2 |
| Blur kernel size | [21,21] | [21,21] |
| Gaussian blur intensity | 0.5 | 0.5 |

**Tab. II** *Hyperparameter & its values.*

As shown in Tab. II, a few hyperparameters are kept the same for both the pretext and downstream phases, whereas others are tuned to fulfil the specific demands of each task.

A five-fold cross-validation is used for each experiment. The dataset is randomly divided into five subsets, with each subset being taken as the test fold in turn and the other four used for training and validation. These performance metrics – accuracy, precision, recall, and F1-score are logged for every loop of iterations. They are then averaged and constitute a thorough analysis of the generalizability and robustness of the model across diverse segments of the data. Such a systematic framework allows for consistent performance analysis of the approach for different scenarios in terms of training settings, data availability and combinations of datasets, providing important information about the practical applicability and flexibility of the proposed method.

In each fold, one subset is treated as the held-out test set, while the remaining subsets are used for training and validation. The reported accuracy, precision, recall, and F1-score values in Sections 5.2–5.4 correspond to test-fold performance averaged across the five folds. Since both benchmark datasets contain only damage-category labels, all experiments reported in Section 5.2–5.4 evaluate multi-class classification performance among road damage categories.

## 5.2 Experiment 1

### Pretext and Downstream on Road Crack Classification Dataset

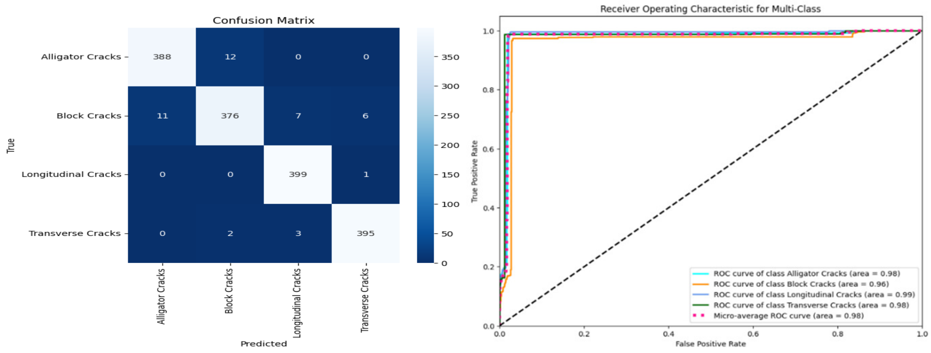
This section delves into the details of Experiment 1, which involves performing both the pretext and target tasks on the Road crack classification dataset. During pretext, the proposed approach learned representation using the proposed loss function while the downstream task focuses on the classification of road damage.

**Experiment 1.1 – Fine-tuning:** Tab. III shows the fine-tuning results for different percentages on the Road crack classification dataset. Fine-tuning involves training the entire network during the downstream task.

| Exp. | Data used (%) | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) |
|---|---|---|---|---|---|
| 1.1.1 | 100 | 97.38 | 97.36 | 97.37 | 97.36 |
| 1.1.2 | 50 | 96.18 | 96.28 | 96.17 | 96.27 |
| 1.1.3 | 20 | 90.00 | 90.27 | 90.10 | 89.69 |

**Tab. III** *Exp. 1.1 – Fine-tuning results on the Road crack classification dataset.*

When the model is trained on the complete dataset, the accuracy of the model is excellent, 97.38%. Interestingly, using only 50% of the labelled data, the accuracy still can reach 96.18%. Remarkably, it achieves a high accuracy of 90% with just 20% of the labelled data remaining, showcasing a significant reduction in the reliance on large-scale annotation. This behaviour holds for the other metrics of evaluation (precision, recall, F1-score), confirming the strength and label efficiency of the proposed framework. Fig. 2a presents the confusion matrix on 100% data, detailing the true positives, false positives, false negatives, and true negatives for each crack type.



**(a)** **(b)**

**Fig. 2** *(a) Confusion matrix on Road crack classification dataset; (b) ROC curve on Road crack classification dataset.*

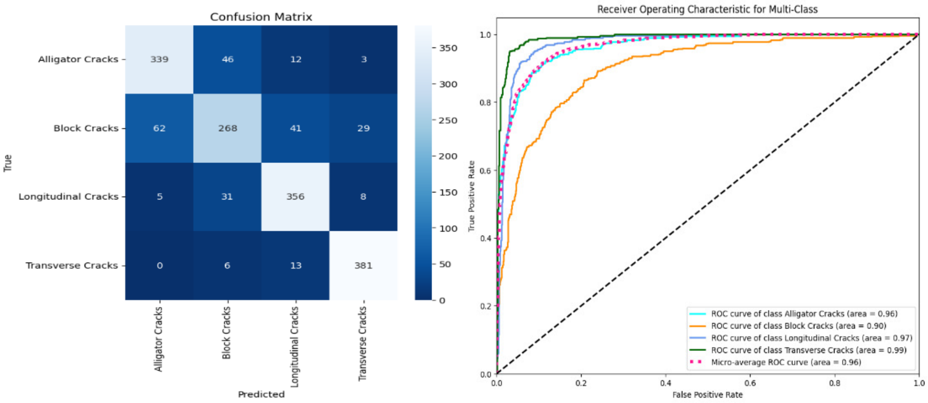
The confusion matrix in Fig. 2a shows how well the model distinguishes among alligator, block, longitudinal, and transverse cracks. The corresponding ROC curves are shown in Fig. 2b, which also demonstrates that the model showed reliable discriminative performance.

**Experiment 1.2 – Feature Extraction:** In this case, the encoder portion of the network is frozen, and only fully connected layers are trained for the downstream (classification) task. The results are summarized in Tab. IV.

| Exp. | Data used (%) | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) |
|---|---|---|---|---|---|
| 1.2.1 | 100 | 84.00 | 83.67 | 84.01 | 83.73 |
| 1.2.2 | 50 | 81.38 | 80.66 | 81.37 | 80.67 |
| 1.2.3 | 20 | 78.12 | 77.63 | 78.12 | 75.83 |

**Tab. IV** *Exp. 1.2 – Feature extraction results on the Road crack classification dataset.*

Tab. IV shows that by using the full dataset; the proposed approach achieves 84% accuracy and similar values for precision, recall and F1-score without ever training the encoder for a classification task. For 50% of the data, it achieved 81% accuracy and similar values for precision, recall, and F1-score. This approach achieve a very promising performance of over 75% across all metrics, even with just 20% of the data being labelled, which reinforces the idea that the representations learned are highly generalizable and informative. Figs. 3a & 3b offer more details, where Fig. 3a shows the confusion matrix on complete data for several crack types & Fig. 3b shows the corresponding ROC curve.



**(a)** **(b)**

**Fig. 3** *(a) Confusion matrix on Road crack classification dataset; (b) ROC curve on Road crack classification dataset.*

## 5.3 Experiment 2

### Pretext and Downstream on Road Defects Images Dataset

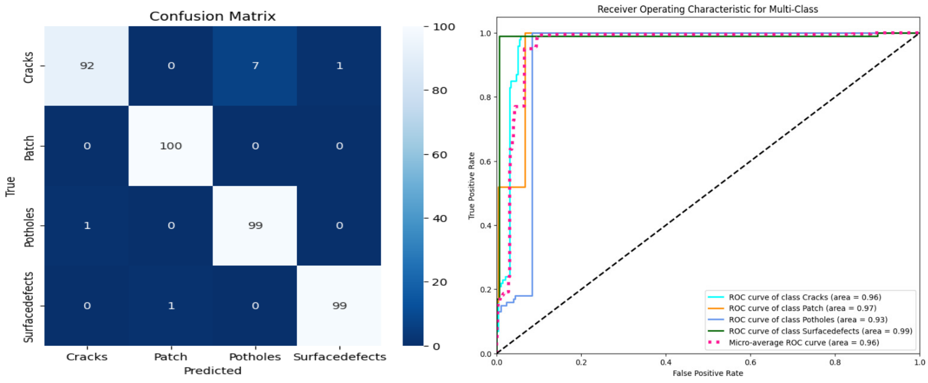
In this section, we illustrate the performance of the proposed method on the Road defects images dataset. In the downstream phase, similar to the past experiment, we compare two training schemes: fine-tuning, where the trained network is subsequently trained end to end, and feature extraction, where only fully connected layers are trained and the encoder is frozen. We further evaluate label efficiency for this approach using subsets of the dataset, including 100%, 50%, and 20% of the available annotations.

**Experiment 2.1 – Fine-tuning:** Tab. V presents the fine-tuning results obtained using the proposed approach for the road defect images dataset.

| Exp. | Data used (%) | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) |
|---|---|---|---|---|---|
| 2.1.1 | 100 | 98.00 | 98.03 | 98.01 | 97.99 |
| 2.1.2 | 50 | 97.50 | 97.61 | 97.50 | 97.62 |
| 2.1.3 | 20 | 95.60 | 95.15 | 95.00 | 94.99 |

**Tab. V** *Exp. 2.1 – Fine-tuning results on Road defects images dataset.*

The model obtained an accuracy of 98% and similar values of precision, recall and F1-score on complete data. It achieves an accuracy of 97.5% even on half of the data and an accuracy of 95.6% with only one-fifth of the labelled data, which is quite impressive. Fig. 4a showcases the confusion matrix summarizing the classification outcomes, highlighting the true positives, false positives, false negatives, and true negatives for each road damage category, namely cracks, potholes, patches, and surface defects, as predicted by the model.



(a) (b)

**Fig. 4** *(a) Confusion matrix on Road defects images dataset (b) ROC curve on Road defects images dataset.*

Additionally, Fig. 4b illustrates the ROC curves, which indicate commendable performance.

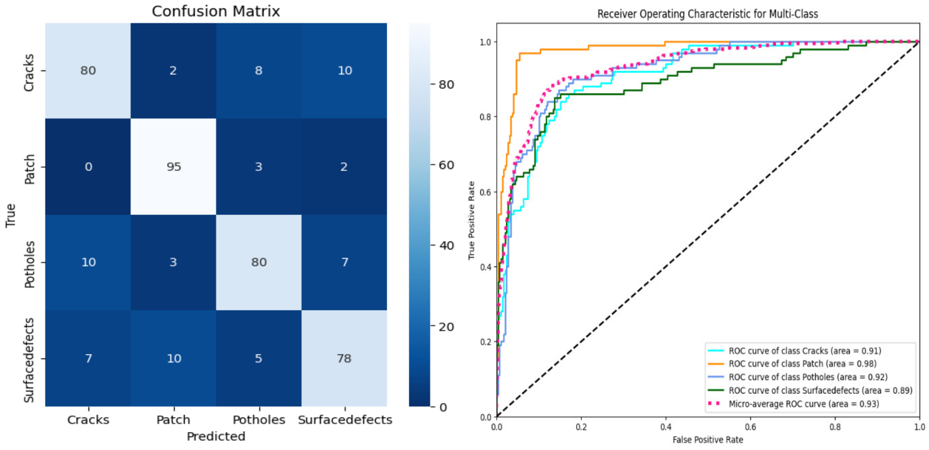
**Experiment 2.2 – Feature Extraction:** The feature extraction results using the proposed approach on the Road defects images dataset in terms of accuracy, precision, recall, and F1-score are shown in Tab. VI.

| Exp. | Data used (%) | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) |
|---|---|---|---|---|---|
| 2.2.1 | 100 | 83.25 | 83.14 | 83.25 | 83.12 |
| 2.2.2 | 50 | 70.00 | 70.50 | 70.00 | 69.68 |

**Tab. VI** *Exp 2.2 – Feature extraction results on Road defects images dataset.*

For the feature extraction scenario, the model accuracy is at 83% using the whole dataset, as is the precision, recall and F1-score. With half of the data, the model gives us satisfactory results without the need to retrain the encoder. However, when using a 20% split, the remaining dataset is too small, leaving only 16 images for validation. Thus, this experiment is not conducted as it has generated random and meaningless results. The above results suggest that the learned representations are sufficiently rich to yield satisfactory classification performance, albeit with limited resources. Fig. 5a (confusion matrix) and Fig. 5b (ROC curve) show the classification quality per defect category, further proving the effectiveness of the learned representations on full data.

Overall, Experiment 2 shows that the proposed method achieves high performance and significant robustness in less training data in both full fine-tuning and feature extraction modes. This stability in performance across different availability



(a) (b)

**Fig. 5** *(a) Confusion matrix on Road defects images dataset; (b) ROC curve on Road defects images dataset.*

scenarios emphasizes the approach's suitability for practical settings, where access to annotated data might be restricted.

## 5.4 Experiment 3

### Cross-Domain Evaluation

Experiment 3 investigates the cross-domain knowledge transfer capabilities of the proposed approach, specifically examining whether features learned during pretext tasks can generalize effectively to different downstream tasks.

**Experiment 3.1 – Fine-tuning in Cross-domain Settings:** In Experiment 3.1, two cross-domain fine-tuning scenarios were explored as described below.

- **Exp 3.1.1:** Pre-trained on Road defects images, fine-tuned on Road crack classification.
- **Exp 3.1.2:** Pre-trained on Road defects classification, fine-tuned on Road defects images.

The results of these experiments, summarized in Tab. VII, demonstrate significant transferability of learned representations.

| Exp. | Pretext dataset | Downstream dataset | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) |
|---|---|---|---|---|---|---|
| 3.1.1 | Road defects images | Road crack classification | 97.12 | 97.15 | 97.13 | 97.10 |
| 3.1.2 | Road crack classification | Road defects images | 81.50 | 82.09 | 81.50 | 80.80 |

**Tab. VII** *Exp. 3.1 – Fine-tuning results in cross-domain settings.*

Notably, the model achieved over 97% accuracy when transferring from Road defects images to Road crack classification. Conversely, transferring from Road crack classification using the Road defects images dataset in the pretext and fine-tuning the Road crack classification to Road defects yielded a solid accuracy of 81.5%.

**Experiment 3.2 – Feature Extraction in Cross-Domain Settings:** Experiment 3.2 evaluated feature extraction without retraining the encoder to test the robustness of learned representations across domains. Tab. VIII summarizes these findings.

Under the feature extraction settings, it can reach an accuracy of 68.31% in transferring from Road defects images to Road crack classification and 58.75% in the opposite direction (from Road crack classification to Road defects images), respectively. Although their values are lower than those obtained from fine-tuning, they are still much higher than the random accuracy of 25% that we would expect for a four-class classification problem.

| Exp. | Pretext dataset | Downstream dataset | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) |
|---|---|---|---|---|---|---|
| 3.2.1 | Road defects images | Road crack classification | 68.31 | 67.31 | 68.31 | 67.52 |
| 3.2.2 | Road crack classification | Road defects images | 58.75 | 57.90 | 58.75 | 55.86 |

**Tab. VIII** *Exp. 3.2 – Feature extraction results in cross-domain settings.*

## 5.5 Ablation Study

In this work, contrastive learning was performed using NT-Xent loss and redundancy reduction learning was performed using Barlow Twins loss. To further enhance the learned image representations, we incorporated regularization through a variance loss term to prevent norm collapse and a covariance loss term to minimize redundancy across dimensions and avert informational collapse. We evaluated the individual performance of contrastive and redundancy reduction methods and their combination, as well as the proposed approach on the Road crack classification dataset and Road defects images dataset using key performance metrics. Tab. IX shows the results of these comparisons across both datasets:

The results indicate that standalone contrastive learning and redundancy reduction exhibit good performance individually. However, combining contrastive and redundancy reduction significantly enhances results, demonstrating their complementary nature. Adding variance or covariance to this combination yields minor gain, and neither outperforms the combined baseline. The proposed approach links all of the components—contrastive, redundancy reduction, variance, and covariance

| Approach | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) |
|---|---|---|---|---|
| Road crack classification dataset | | | | |
| Contrastive | 97.06 | 97.05 | 97.06 | 97.04 |
| Redundancy reduction | 96.25 | 96.28 | 96.26 | 96.20 |
| Cont + Red_Reduc | 97.19 | 97.19 | 97.19 | 97.19 |
| Cont + Red_Reduc + Variance | 96.69 | 96.68 | 96.68 | 96.67 |
| Cont + Red_Reduc + Covariance | 96.75 | 96.78 | 96.76 | 96.75 |
| **Proposed approach** | **97.38** | **97.36** | **97.37** | **97.36** |
| Road defects images dataset | | | | |
| Contrastive | 95.00 | 95.29 | 95.01 | 94.85 |
| Redundancy reduction | 96.75 | 96.74 | 96.78 | 96.77 |
| Cont + Red_Reduc | 97.00 | 97.06 | 97.01 | 96.97 |
| Cont + Red_Reduc + Variance | 97.50 | 97.53 | 97.51 | 97.48 |
| Cont + Red_Reduc + Covariance | 97.25 | 97.30 | 97.26 | 97.24 |
| **Proposed approach** | **98.00** | **98.03** | **98.01** | **97.99** |

**Tab. IX** *Comparison across different approaches on both datasets.*

– inside one loss. As a combination of each, it outweighs in strength and achieves the overall best performance on all metrics, potentially leveraging the strengths of all approaches on both datasets.

### 5.6 Qualitative Analysis

To provide some qualitative insight alongside the quantitative results, we evaluated the proposed approach using class activation maps (CAM), which show which areas of an image are most salient to the model when making its predictions. CAMs provide visibility into an internal black-box mechanism of the model and can be useful for checking the quality of the learned features. Figs. 6 and 7 presents a few samples of class activation maps with their corresponding original images from both datasets.

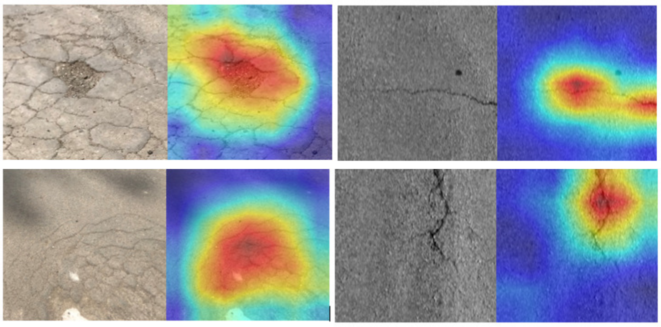


**Fig. 6** *Sample CAMs on Road crack classification dataset.*

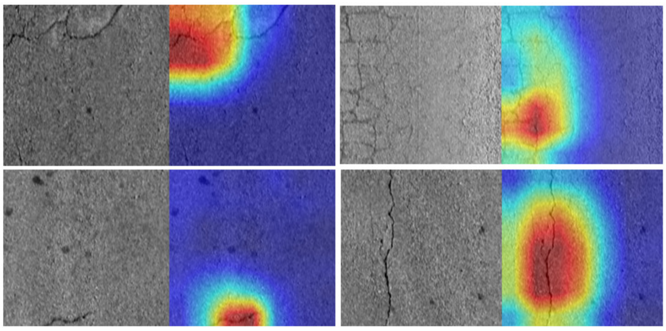


**Fig. 7** *Sample CAMs on Road defects images dataset.*

The class activation maps illustrate that the model successfully locates areas corresponding to road damage, as shown in Figs. 6 and 7. Such visualizations validate that the model learns meaningful features and focuses its attention on sufficient areas required for classification. When applied to various images, the highlighted regions were found to align well with expectations, reinforcing confidence in the model's performance.

## 5.7 Comparative Analysis

In this section, we compare the proposed self-supervised learning-based approach with both fully supervised and state-of-the-art road damage classification methods. We start by benchmarking our approach with a standard supervised learning baseline over a range of data availability. We then compare our approach to previously reported solutions in the literature.

### Comparision with Supervised Learning-Based Approach on Road Defects Images Dataset

In the supervised learning experiment, we directly performed the downstream task and used the labelled data for training. We compare our approach with the supervised learning one on the Road defects images dataset in Tab. X.

| Approach | Data used (%) | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) |
|---|---|---|---|---|---|
| Supervised | 100 | 89.25 | 89.40 | 89.24 | 89.22 |
| | 50 | 73.50 | 73.82 | 73.48 | 73.02 |
| | 20 | 62.50 | 62.12 | 62.51 | 61.57 |
| SSL-based Proposed Approach | 100 | 98.00 | 98.03 | 98.01 | 97.99 |
| | 50 | 97.50 | 97.61 | 97.50 | 97.62 |
| | 20 | 95.60 | 95.15 | 95.00 | 94.99 |
| Improvement over supervised | 100 | 8.75 | 8.63 | 8.77 | 8.77 |
| | 50 | 24.00 | 23.79 | 24.02 | 24.60 |
| | 20 | 33.10 | 33.03 | 32.49 | 33.42 |

**Tab. X** *Comparison with supervised learning approach on the Road defects images dataset.*

The experimental results in Tab. X show that our SSL-based approach outperforms the supervised learning method for the Road defects images dataset. It is clear that in terms of improved accuracy, on the native data level across all usage levels, the SSL-based approach performs much better than another approach, where we gained an improvement of 8.75% in accuracy with the complete dataset, 24% with half of a dataset, 33.1% with just 20% of the dataset. Similarly, all the other metrics, including precision, recall, and F1-score, also demonstrated significant improvements. These results highlight the efficiency and robust nature of the SSL-based approach, particularly in scenarios with limited labelled data, making it a scalable and cost-effective solution for road damage classification task.

### Comparision with Supervised Learning and Prior-Art on Road Crack Classification Dataset

In our approach, we found only one prior work addressing road damage using the Road crack classification dataset. Tab. XI presents a comparison of the proposed SSL-based approach with the supervised learning and prior-art on the Road crack classification dataset. This comparison highlights the significant advancements achieved by our method in terms of accuracy, precision, recall, and F1-score.

| Approach | Data used (%) | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) |
|---|---|---|---|---|---|
| Supervised | 100 | 77.50 | 77.29 | 77.50 | 76.65 |
| | 50 | 65.75 | 66.01 | 65.76 | 65.33 |
| | 20 | 53.12 | 55.91 | 53.12 | 53.11 |
| Zhang et al. [33] (prior-art) | 100 | 85.70 | 85.80 | 82.90 | 83.40 |
| Proposed approach | 100 | 97.38 | 97.36 | 97.37 | 97.36 |
| | 50 | 96.38 | 95.28 | 95.17 | 95.27 |
| | 20 | 90.00 | 90.27 | 90.01 | 89.69 |
| The improvement over the supervised | 100 | 19.88 | 20.07 | 19.87 | 20.71 |
| | 50 | 30.63 | 29.27 | 29.41 | 29.94 |
| | 20 | 36.88 | 34.36 | 36.89 | 36.58 |
| The improvement over the prior-art | 100 | 11.68 | 11.56 | 14.47 | 13.96 |
| | 50 | 10.68 | 9.48 | 15.07 | 11.87 |
| | 20 | 4.30 | 4.47 | 7.11 | 6.29 |

**Tab. XI** *Comparison with prior-art on the Road crack classification dataset.*

The findings outlined in Tab. XI demonstrate the clear superiority of the novel SSL-based approach over traditional supervised learning mechanisms for classifying road cracks. With full utilization of the data, the SSL-based approach achieved a sizable improvement of 19.88% accuracy. Utilizing half of the data, we still saw that the SSL approach far outperforms the supervised approach by 30.63% accuracy. Perhaps most impressive was the SSL-based approach prowess, even when restricted to just one-fifth of the information, with an accuracy improvement of 36.88%. Similarly, all the other metrics, including precision, recall, and F1-score, also demonstrated significant improvements.

Moreover, it can be seen that the improvements obtained by the proposed SSL-based approach over the prior-art on the Road crack classification dataset are substantial. When using the entire set of data, the proposed approach delivers an 11.68% improvement in accuracy from the prior art. Likewise, precision, recall, and F1-score are significantly improved by 11.56%, 14.47%, and 13.96%, respectively. These results show that leveraging all data using the proposed approach is capable of achieving state-of-the-art performance. With half of the data, the proposed approach achieves a 10.68% higher accuracy than the prior-art, with precision, recall, and F1-score being 9.48%, 15.07%, and 11.87% higher than the prior-art. With only 20% of the data used, the proposed approach is 4.3% higher than the accuracy obtained by the prior art and consistently improves precision, recall, and F1-score.

These results demonstrate that the proposed SSL-based approach consistently outperforms supervised learning and prior-art methods, particularly under limited-label settings. The findings also indicate that the learned representations transfer effectively across datasets, highlighting the robustness of the proposed framework.

A limitation of the present study is that the employed benchmark datasets do not contain undamaged or clear-road images as a separate class. Therefore, the proposed framework should be interpreted as a multi-class road damage classification system rather than a binary damaged-versus-undamaged detector.

# 6. Conclusion

Automatic analysis of road damage is important for efficient infrastructure maintenance and public safety. In this work, we proposed a self-supervised learning-based approach that reduces dependence on large annotated datasets, which remain a major limitation of conventional supervised methods. The proposed regularized redundancy-reduction contrastive learning framework combines the strengths of contrastive learning and redundancy reduction to learn robust, informative, and non-redundant feature representations. The inclusion of variance and covariance regularization further improves representation stability and prevents collapse. Experimental results on two benchmark datasets, under both in-domain and cross-domain settings, demonstrate that the proposed approach consistently outperforms fully supervised counterparts, especially under limited-label conditions. Notably, the model trained with only 20% of the labelled data achieves superior performance compared with supervised and prior-art methods trained on the full dataset. Class activation maps further show that the learned representations are interpretable and focus on the most relevant damaged regions. A limitation of the present study is that the employed benchmark datasets do not include a separate no-damage or clear-road class. Therefore, the current work should be interpreted as multi-class road damage classification among damaged-road categories rather than binary damaged-versus-undamaged detection. Future work will extend the proposed framework to include undamaged road scenes, road damage segmentation, and real-time deployment in practical road inspection scenarios.

## Acknowledgement

The author declares that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors

# References

[1] ADNAN F., AWAN M.J., MAHMOUD A., NOBANEE H., YASIN A., ZAIN A.M., EfficientNetB3-Adaptive Augmented Deep Learning (AADL) for Multi-Class Plant Disease Classification. *IEEE Access.* 2023, 11, pp. 85426–85440, doi: 10.1109/ACCESS.2023.3303131.

[2] AHMAD C.F., SAYEGH A., CHEEMA A., QAYYUM W., EHTISHAM R., SAGHIR S., Ahmad A. Classification of different size of potholes based on surface area using convolutional neural network. *Discover Applied Sciences.* 2024, 6(9), p. 492, doi: 10.1007/s42452-024-06207-3.

[3] APARNA, BHATIA Y., RAI R., GUPTA V., AGGARWAL N., AKULA A. Convolutional neural networks based potholes detection using thermal imaging. *Journal of King Saud*

*University – Computer and Information Sciences*. 2022, 34(3), pp. 578–588, doi: 10.1016/j.jksuci.2019.02.004.

[4] ARYA D., MAEDA H., GHOSH S.K., TOSHNIWAL D., SEKIMOTO Y., RDD2020: An annotated image dataset for automatic road damage detection using deep learning. *Data Brief*. 2021, 36, p. 107133, doi: 10.1016/j.dib.2021.107133.

[5] BARDES A., PONCE J., LECUN Y. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. In: *The Tenth International Conference on Learning Representations*. ICLR, 2022.

[6] BHATLAWANDE S., DESHPANDE A., DESHPANDE S., SHILASKAR S. Proactive Detection of Pothole and Walkable Path for Safe Mobility of Visually Challenged. In: *2022 3rd International Conference for Emerging Technology (INCET), 2022*. 2022, pp. 1–5.

[7] BUČKO B., LIESKOVSKÁ E., ZÁBOVSKÁ K., ZÁBOVSKÝ M. Computer Vision Based Pothole Detection under Challenging Conditions. *Sensors*. 2022, 22(22), doi: 10.3390/s22228878.

[8] CARLOS M.R., GONZALEZ L.C., WAHLSTROM J., CORNEJO R., MARTINEZ F. Becoming Smarter at Characterizing Potholes and Speed Bumps from Smartphone Data-Introducing a Second-Generation Inference Problem. *IEEE Trans Mob Comput*. 2021, 20(2), pp. 366–376, doi: 10.1109/TMC.2019.2947443.

[9] CHEN T., KORNBLITH S., NOROUZI M., HINTON G. A Simple Framework for Contrastive Learning of Visual Representations. In: III, Hal Daumé, Singh, Aarti, ed. *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 2020, pp. 1597–1607.

[10] CHOLLET F. Xception: Deep Learning with Depthwise Separable Convolutions. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA. 2017, pp. 1800–1807.

[11] CHU H., SAEED M., RASHID J., MEHMOOD M., AHMAD I., IQBAL R., GHULAM A. Deep Learning Method to Detect the Road Cracks and Potholes for Smart Cities. *Computers, Materials and Continua*. 2023, 75(1), pp. 1863–1881, doi: 10.32604/cmc.2023.035287.

[12] CINAR N., KAYA M. An automated pothole detection via transfer learning. In: *2022 International Conference on Decision Aid Sciences and Applications, DASA 2022*, Chiangrai, Thailand. IEEE Inc., 2022, pp. 1355–1358.

[13] CUMBRERA F. de Q. Statista Home Page. In: *Global investment in road maintenance, 2022* [Online]. Statista [viewed 2024-10-07]. Available from: https://www.statista.com/statistics/1479073/global-investment-in-road-maintenance

[14] EGAJI O.A., EVANS G., GRIFFITHS M.G., Islas G. Real-time machine learning-based approach for pothole detection. *Expert Syst Appl*. 2021, 184, doi: 10.1016/j.eswa.2021.115562.

[15] GAO M., WANG X., ZHU S., GUAN P. Detection and Segmentation of Cement Concrete Pavement Pothole Based on Image Processing Technology. *Math Probl Eng*. 2020, 2020(1), p. 1360832, doi: 10.1155/2020/1360832.

[16] GUPTA Y., CHAUHAN F., SINGLA K. Analysis of Different Deep Learning Algorithms for Road Surface Damage Detection. In: *2023 International Conference on Disruptive Technologies (ICDT)*, Greater Noida, India. 2023, pp. 13–17.

[17] HE K., ZHANG X., REN S., SUN J. Deep Residual Learning for Image Recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA. IEEE, 2016, pp. 770–778.

[18] HE K., FAN H., WU Y., XIE S., GIRSHICK R. Momentum Contrast for Unsupervised Visual Representation Learning. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA. 2020, pp. 9726–9735.

[19] HEARST M. A., DUMAIS S. T., OSUNA E., PLATT J., SCHOLKOPF B. Support vector machines. *IEEE Intelligent Systems and their Applications*. 1998, 13(4), pp. 18–28, doi: 10.1109/5254.708428.

[20] HU J., SHEN L., ALBANIE S., SUN G., WU E. Squeeze-and-Excitation Networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2017, pp. 7132–7141, https://api.semanticscholar.org/CorpusID:140309863.

[21] HUANG G., LIU Z., VAN DER MAATEN L., WEINBERGER K.Q. Densely Connected Convolutional Networks. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA. IEEE, 2017, pp. 2261–2269.

[22] JANA S., MIDDYA A.I., ROY S., Participatory Sensing Based Urban Road Condition Classification using Transfer Learning. *Mobile Networks and Applications*. 2023, 29, pp. 42–58, doi: 10.1007/s11036-023-02118-6.

[23] JUSTO-SILVA R., FERREIRA A. Pavement maintenance considering traffic accident costs. *International Journal of Pavement Research and Technology*. 2019, 12(6), pp. 562–573,doi: 10.1007/s42947-019-0067-3.

[24] KLCO P., KONIAR D., HARGAS L., PASKALA M. Automated Detection Of Potholes Using YOLOv5 Neural Network. In: *Transportation Research Procedia*. Elsevier B.V., 2023, pp. 1150–1155.

[25] LI D., DUAN Z., HU X., ZHANG D., ZHANG Y. Automated classification and detection of multiple pavement distress images based on deep learning. *Journal of Traffic and Transportation Engineering (English Edition)*. 2023, 10(2), pp. 276–290, doi: 10.1016/j.jtte.2021.04.008.

[26] LI Y., LIU C., GAO Q., WU D., LI F., DU Y. ConTrack Distress Dataset: A Continuous Observation for Pavement Deterioration Spatio-Temporal Analysis. *IEEE Transactions on Intelligent Transportation Systems*. 2022, pp. 1–14, doi: 10.1109/TITS.2022.3201968.

[27] NOROOZI M. , VINJIMOOR A. , FAVARO P., PIRSIAVASH H. Boosting Self-Supervised Learning via Knowledge Transfer. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, IEEE Computer Society, 2018, pp. 9359–9367.

[28] PATRA S., MIDDYA A.I., ROY S. PotSpot: Participatory sensing based monitoring system for pothole detection using deep learning. *Multimed Tools Appl.* 2021, 80(16), pp. 25171–25195, doi: 10.1007/s11042-021-10874-4.

[29] QASSIM H., VERMA A., FEINZIMER D. Compressed residual-VGG16 CNN model for big data places image recognition. In *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, Las Vegas, NV, USA. IEEE, 2018, pp. 169–175.

[30] RATHEE M., BAČIĆ B., DOBORJEH M. Automated Road Defect and Anomaly Detection for Traffic Safety: A Systematic Review. *MDPI*. 2023, doi: 10.3390/s23125656.

[31] World Health Organization Home Page. In: *Road traffic injuries 2023*. [Online]. WHO [viewed 2024-10-07]. Available from: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

[32] ZBONTAR J., JING L., MISRA I., LECUN Y., DENY S. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In: *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*. PMLR, 2021, pp. 12310–12320.

[33] ZHANG T., WANG D., LU Y. Benchmark Study on a Novel Online Dataset for Standard Evaluation of Deep Learning-based Pavement Cracks Classification Models. *KSCE Journal of Civil Engineering*. 2024, 28(4), pp. 1267–1279, doi: 10.1007/s12205-024-1066-8.