Patch-Based Self-Supervised Learning for Road Damage Classification: A Case Study on the RDD2022 dataset of Indian test site

Poonam Jayhind Pardeshi [a, *], Shailja [a], Anushka Chaudhary [b], Arya [c], Manohar Yadav [a]

[a] Geographic Information System (GIS) Cell, Motilal Nehru National Institute of Technology Allahabad, Prayagraj- 211004, India–
(poonam.2024rgi01, shailja.2021rgi03, ssmyadav) @mnnit.ac.in
[b] School of Computer Science & Engineering, Lovely Professional University, Jalandhar- 144411, India–
anushkarchaudhary@gmail.com
[c] Department of Computing Technologies, SRM Institute of Science and Technology, Chennai- 603203, India–
choudhuryarya186@gmail.com

Keywords: Road, Image, Self-supervised learning, MobileNetV2, Classification, Road infrastructure monitoring.

Abstract

Timely detection of road surface damage is recognized as essential for maintaining safe and efficient transportation infrastructure. In developing countries such as India, damage types including cracks, potholes, and surface wear are worsened due to climatic conditions, heavy traffic, and inconsistent maintenance. Manual inspection is resource-intensive and non-scalable, emphasizing the need for automated, learning-based approaches. This study proposes a lightweight, patch-based self-supervised learning (SSL) framework using MobileNetV2 for classifying five road damage types in the India-specific subset of the Road Damage Detection 2022 (RDD2022) dataset. Although RDD2022 supports deep learning for road damage detection, patch-wise modeling remains largely unexplored. The methodology comprises four stages: Image patching, SSL-based pretraining with augmentation, supervised fine-tuning on labeled patches, and evaluation. SSL facilitates representation learning from unlabeled data, crucial in domains with limited annotations. Combined with patch-based sampling, localized damage features are captured, improving performance under intra-class imbalance. MobileNetV2 is selected for its fast convergence and edge-device compatibility, making it suitable for deployment in low-resource settings. The proposed model achieves 78% overall accuracy and a weighted F1-score of 78% on the test set. Training accuracy improves steadily over 25 epochs, reaching over 91%, while validation accuracy stabilizes at approximately 78%. Compared to standard CNN architectures, competitive performance is achieved without large pretrained models or high-end computational resources. The approach supports real-time inference, geospatial integration, and potential applications in infrastructure monitoring and urban planning. Validation against state-of-the-art models confirms the framework's effectiveness and relevance for scalable, region-specific road damage classification.

1. Introduction

The safety, efficiency, and sustainability of transportation systems are enhanced through the timely detection and classification of road surface damage. India has the second-largest road network in the world, while 90% of the total passenger traffic uses road network to commute (India Brand Equity Foundation [IBEF], 2025; Chorada et al., 2023). The Ministry of Road Transportation and Highways Department has also stated that one of the main reasons for accidents is road deterioration (Ministry of Road Transport and Highways [MoRTH], 2023). In developing countries, road damages such as cracks, potholes, rutting, bumps, crosswalk blur, and white line blur are caused by climatic variability, high vehicle loads, and insufficient maintenance practices.

A benchmark dataset, Road Damage Detection 2022 (RDD2022), was released to support the development of deep learning-based methods for the automatic detection and classification of road damage. It is comprised of 47,420 road images collected from six countries: India, Japan, the Czech Republic, Norway, the United States, and China (Arya et al., 2022). Despite recent advancements in deep learning, the performance of conventional supervised models remains limited due to the high costs and labour-intensive nature of large-scale annotation.

To address this challenge, this study adopts self-supervised learning (SSL) as an effective approach for representation learning without extensive manual labels. A patch-based SSL framework employing MobileNetV2 is proposed for classifying five road damage types within the India-specific subset of RDD2022. Bounding box annotations are used to extract fixed-size patches, each assigned to one of five categories: longitudinal crack, transverse crack, alligator crack, pothole, or other surface corruption. The framework is designed to learn discriminative features for accurate multi-class classification, thereby providing a more detailed understanding of road surface conditions than is achieved through defect classification. Although severity, size, or density quantification is not performed in this work, the spatially localized classification outputs offer a foundation for future severity assessment and maintenance prioritization.

compared with state-of-the-art models, the proposed edge-compatible CNN model outperforms traditional training in data-scarce setting. The integration of this framework is expected to support real-time road condition monitoring and to facilitate sustainable urban planning.

2. Related Work

The automation of road damage detection is being increasingly studied due to its potential to enhance transportation safety and reduce infrastructure maintenance costs for smart cities. Various deep learning techniques are being proposed ranging from classification of patch-based damages to object detection on full road images and pixel-level segmentation methods.

*Corresponding Author: poonam.2024rgi01@mnnit.ac.in (Poonam Jayhind Pardeshi)

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-5-W2-2025-459-2025

The suitability of transfer learning for detecting road damage in diverse geographical areas has been investigated in several research. Arya et al.,2021, studied the utilization of the SSD MobileNet model pre-trained on the MSCOCO (Microsoft Common Objects in Context) dataset to train the required road damage detection models using transfer-learning. Compiled results for different models are shown which were applied on test data of India. The results demonstrate that models trained using Indian data either used purely or in mixed form of InJa models are observed to perform significantly better than those trained on Japan data. However, even models trained with Indian data are found to perform poorly on the Transverse crack category. This limitation is attributed to the presence of limited samples of Transverse crack category. The performance of models trained purely on Indian data is noted to be slightly better than that of the InJa models, despite the former being trained on smaller datasets. The experiments illustrate that an efficient road damage detection model for a given country can be developed by combining local data with data from another country (Arya et al., 2021). This study's findings on cross-country dataset mixing inform our approach by highlighting the importance of localized training data in improving model performance, especially for rare defect categories.

Alfarrarjeh et al., 2018, proposed a study on automating the detection of different types of road damages using smartphone images crowdsourced by city crews or the public. It uses YOLO for training a model to detect road damages as distinguishable objects in the analyzed images. The solution was able to achieve an F1 score up to 0.62 by augmenting more synthesized images to the low-cardinality classes of the training set and using a high value of the non-maximum suppression at prediction (Alfarrarjeh et al., 2018). Their work on YOLO-based object detection for road damage provides a baseline detection framework that our method builds upon with enhanced feature extraction and data balancing strategies.

Alqethami et al. (2022) developed an efficient automated system for detecting road damages. The method was trained with four CNN models on a reliable Saudi dataset, and a transfer learning approach was applied with pre-trained models to address the challenges of data volume and training time. The results of classifying six types of damage showed that RoadNet obtained the highest accuracy, followed by AlexNet, VGG-16, and ResNet-34 (Alqethami et al., 2022). Their evaluation of multiple CNN architectures and use of transfer learning supports our choice of pre-trained models to address data scarcity and accelerate training.

Bouhsissin et al. (2025) introduced the DD-CNN-23Layers model, a deep learning-based approach for detecting and classifying various types of road damages. An indigenous dataset enabled the model to achieve better performance than several pre-trained YOLO models (v7 to v10) in both accuracy and speed. The model demonstrates efficiency and suitability for real-time, in-vehicle applications (Bouhsissin et al., 2025). The demonstrated efficiency of DD-CNN-23Layers for real-time detection aligns with our goal of developing a solution suitable for in-vehicle or mobile deployment.

Hasan et al. (2022) proposed a method that uses four backbone transfer learning models and three different detection heads. The model was trained with three optimizers and batch sizes of 4, 8, 16, and 32 in Faster R-CNN with ResNet-101. Training with different optimizers and batch sizes showed that the best F1-score was achieved with a batch size of 4 and the SGD optimizer with momentum (Hasan et al., 2022). Their experiments on optimizer and batch size variations inform our own training parameter selection for balancing accuracy and computational cost.

Maeda et al. (2018) developed a new large-scale dataset for road damage detection and classification. An SSD (Single Shot Detector) with MobileNet and Inception V2 achieved recalls and precisions greater than 71% and 77%, respectively, with an inference time of 1.5 seconds using a smartphone, which is beneficial in regions where experts and financial resources are lacking (Maeda et al., 2018). The release of a large-scale road damage dataset in their work forms part of the dataset foundation for our experiments and benchmarking.

Asif et al. (2024) investigated the role of dataset content in the effectiveness of self-supervised pretext tasks, with a focus on rotation prediction using MobileNetV2. Two distinct datasets, a cat image dataset and the satellite-based xBD building damage dataset were used to evaluate how object consistency, scene complexity, and domain characteristics influence feature learning. The study found that consistent, upright objects (cats) resulted in higher pretext and downstream classification accuracy, whereas diverse and less structured scenes in xBD increased misclassification and overfitting. The authors emphasized that aligning pretext task choice with dataset characteristics is crucial for maximizing the transferability of learned representations and improving generalization in downstream task (Asif et al., 2024). Their insights on dataset-specific suitability of self-supervised tasks directly influence our integration of rotation prediction as a pretext task to improve downstream classification.

Chen et al. 2025, proposed a Cross-Scale Overlapping Patch-Based Attention Network (COP-Net) for road crack detection and segmentation, aiming to address challenges of scale variation and complex backgrounds in pavement imagery. COP-Net comprises two modules: first is Scale Channel Attention (SCA) and Patch-based Cross-Scale Attention (PCA) module. The SCA module perform channel important across various scales. This results in comprehensive and globally informed channel attention feature map for crack detection. Where, PCA module focus on cross-scale considerations on feature maps, enabling the model to concurrently capture information from both large and small cracks. Experimental results on benchmark road crack datasets demonstrated that COP-Net outperformed existing CNN and transformer-based methods in terms of precision and recall, particularly for thin and discontinuous cracks. The study highlights the effectiveness of combining patch-based representation with attention-driven feature fusion for robust infrastructure defect detection (Chen et al., 2025). The COP-Net architecture's success in handling scale variation and fine defects validates our choice of incorporating patch-based techniques for improved defect classification.

3. Dataset and Methodology

Road damage poses significant challenges for both conventional and autonomous vehicles. For traditional vehicles, encountering these road hazards can lead to mechanical damage, increased maintenance costs, and even accidents due to sudden manoeuvres to avoid them (Bouhsissin et al., 2025). To address the challenges of an accurate road damage classification and to demonstrate feasibility of lightweight model for low-resource deployment MobileNetV2 with SSL is used in this study. Before training, the original dataset RDD2022 downloaded from Kaggle and it is extracted for Indian test site images for case study (Ali Abdelmenam & Alaa Gaber, 2025). The data from India includes images captured from local roads, state highways and national highways, covering the metropolitan (Delhi, Gurugram) as well as non-metropolitan regions (mainly from Haryana). All these images have been collected from plain areas using smartphone-mounted vehicles have been used to capture road images from plain regions. Road images were captured using a smartphone running a publicly available image-capturing application developed by Sekimoto Lab, The University of Tokyo.

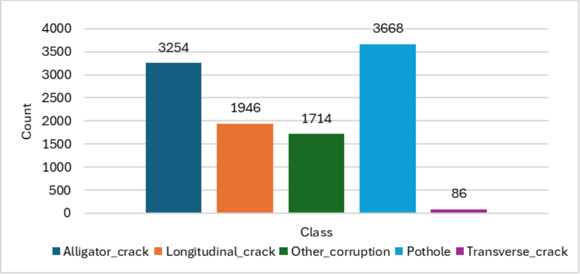

Figure 1. Statistics for the class-wise number of damage images included in the underlying dataset.

The smartphone was installed on the dashboard/windshield of the vehicle, and the updated version of the application was used to capture JPEG images of road damages at 720×960 pixels, which were then resized to 720×720 pixels. The image collection process is configured to prevent overlap or leakage when the vehicle is operated at an average speed of about 40 km/h (25 mph). On motorways and certain first-class roads outside urban areas, the vehicle speed is adjusted in accordance with local traffic regulations. Additionally, the dataset is primarily composed of images of flexible (asphalt) pavements (Arya et al., 2022).

The dataset follows 54-15-15 split proportion with 6793 training images, 1934 in test and 1941 images available for validation after data cleaning. Class-wise 3254 images are available in Alligator crack, 1946 in Longitudinal crack, 1714 in Other corruption and 86 available in Transverse crack class.

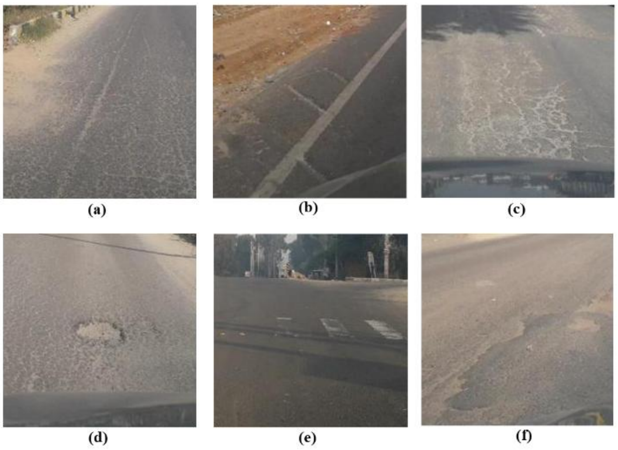

Figure 2. Sample images road damage categories considered for Indian test site. a. Longitudinal crack, b. Transverse crack, c. Alligator crack, d. Pothole, e. Other corruption, f. Other corruption.

These are the five damage classes for classification while class imbalance is also analyzed (Arya et al., 2021). The dataset with this number of images achieved good performance with data diversity and preserved representation. The original road images in the RDD2022 dataset contains multiple damages within single image. To focus learning on the localized damage characteristics, the patches of size 128×128 is extracted which is centered around the annotated bounding boxes. The patch extraction reduces the background variability within the image and also increases the number of training samples, as each damage annotation yields a unique patch. By learning each patch damage as independent training sample, model generalizes over variation in shape, texture and damage orientation. This strategy also aligns well with lightweight CNN architectures, which benefit from small and fixed input resolutions.

Further, the proposed approach adopts a two-stage learning framework, in which self-supervised learning is combined with supervised fine-tuning for patch-level road damage classification. The overall methodology is visualized in the pipeline shown in Figure 3.

3.1 Self-Supervised Pretraining

Self-supervised pretraining stage is based on the RotNet (Rotation) framework, where the encoder is pretrained to predict geometric transformations on unlabeled data to identify the different rotation angles e.g. 0^0, 90^0, 180^0, and 270^0 applied to input patches (Gidaris et al., 2018). The core intuition behind using this rotation-based pretraining is that it helps the model to learn both spatial and visual patterns that are informative for distinguishing rotated views of the road scenes. This involves capturing features such as surface texture, relative positioning of cracks, road markings and other types of road damages. While, rotation-based self-supervised learning improves the quality of learned features, it should be used with care during fine-tuning. It adds random rotation in fine-tuning which may cause the model to ignore orientation details that could help classification. Here, encoder is trained to classify the correct rotation label, helps to extract underlying structural and contextual features from road damages. The encoder employed for this task is based on MobileNetV2, a lightweight convolutional neural network architecture in which depth wise separable convolutions and inverted residual bottleneck blocks are utilized to balance accuracy and efficiency (Howard et al., 2019). Each damage patch image is processed through a sequence of convolutions and residual blocks to produce a compact feature embedding, which is then passed through global average pooling and a fully connected layer which maps the features to four rotation classes, the overall processing of MobileNetV2 encoder is shown in Figure 4. The categorical cross-entropy loss as the objective function during training is defined over the softmax probabilities of the predicted rotation classes. The set of encoder weights is obtained through this process that is optimized for feature discrimination under rotational transformations, the weights are then transferred to the supervised classification task.

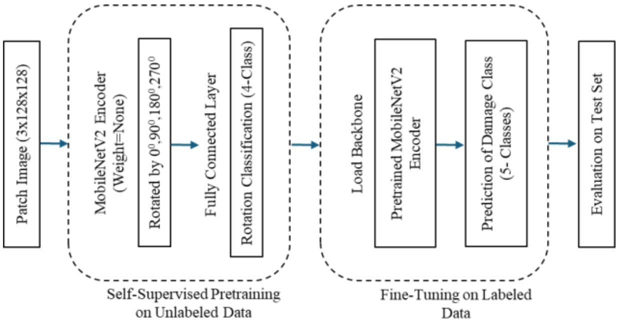
Patch Image (3x128x128)
MobileNetV2 Encoder (Weight=None)
Rotated by 0^0,90^0,180^0,270^0
Fully Connected Layer
Rotation Classification (4-Class)
Load Backbone
Pretrained MobileNetV2 Encoder
Prediction of Damage Class (5- Classes)
Evaluation on Test Set
Self-Supervised Pretraining on Unlabeled Data
Fine-Tuning on Labeled Data

Figure 3. Proposed two-stage pretraining with fine-tuning pipeline for classification.

3.2 Supervised Fine-tuning

After SSL pretraining, the learned weights of pretrained MobileNetV2 encoder is retained, its final rotation prediction head is replaced with a supervised classification layer for five road damage classes. Model is trained on 25 epochs using Adam optimizer with learning rate of 1×10^{-4}. Weighted cross-entropy is used to address the class imbalance issue; Weights are based on the inverse frequency of classes. The lower sample count for the transverse crack introduces a class imbalance that can bias the learning process towards majority classes. But weighted cross-entropy was applied to compensate for this imbalance by assigning higher loss penalties to minority classes. The limited representation may constrain the model's ability to generalize to unseen transverse crack patterns, as its recall being lower than its relative to precision. Feature maps are pooled and fed into the classification layer and validation performance is monitored. The full MobileNetV2 architecture, including its encoder and classification head is shown in Figure 4. This two-stage approach enables efficient transfer learning while reducing reliance on large annotated datasets and maintaining generalization in sparse data conditions.

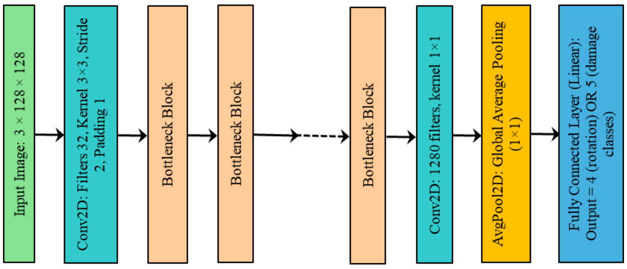
Input Image: 3 × 128 × 128
Conv2D: Filters 32, Kernel 3×3, Stride 2, Padding 1
Bottleneck Block
Bottleneck Block
Bottleneck Block
Conv2D: 1280 filters, kernel 1×1
AvgPool2D: Global Average Pooling (1×1)
Fully Connected Layer (Linear): Output = 4 (rotation) OR 5 (damage classes)

Figure 4. MobileNetV2 architecture with task-specific classification head.

4. Results

The performance of the proposed framework is thoroughly evaluated on the Indian subset. The quantitative evaluation focuses on accuracy, precision, recall, F1-score and qualitative prediction behaviour across damage classes. Results are presented through the classification report, confusion matrix, learning curves, and class-wise prediction visualizations.

4.1 Quantitative Evaluation

The classification accuracy of 78% is achieved with proposed MobileNetV2 model, on the extracted test set having 1934 image patches pretrained using a self-supervised RotNet task and subsequently fine-tuned using supervised learning. Where, the Pothole class records the highest F1-score of 0.82, with recall of 0.79 and precision of 0.82, indicating that it is consistently identified with high reliability.

Class	Precision	Recall	F1-score	Support Images
Longitudinal crack	0.70	0.75	0.73	342
Transverse crack	0.87	0.72	0.79	18
Alligator crack	0.79	0.80	0.80	609
Other corruption	0.71	0.75	0.73	304
Pothole	0.85	0.79	0.82	661
Accuracy			0.78	1934
Macro average	0.78	0.76	0.77	1934
Weighted average	0.79	0.78	0.78	1934

Table 1. MobileNetV2 classification report showing class-wise precision, recall, and F1-score on the RDD2022 India test set.

Alligator cracks also perform well, having an F1-score of 0.80, indicating a good balance with recall 0.80 and precision 0.79. Longitudinal crack and other corruption classes has shown moderate results, both with F1-scores of 0.73. These classes have showed greater variability due to their intra-class variation or structural similarity to other damage classes. However, transverse cracks which are underrepresented in the dataset with only 18 images, achieved an F1-score of 0.79 with a relatively high precision of 0.87 and a lower recall of 0.72. These results suggest that while the model confidently predicts transverse cracks, it fails to recall a proportion of true classes. The weighted average F1-score is computed as 0.78, and the macro averaged F1-score is 0.77, showing a balanced performance across both minor and major classes, despite the dataset's inherent imbalance.

4.2 Class-Wise Performance and Error Distribution

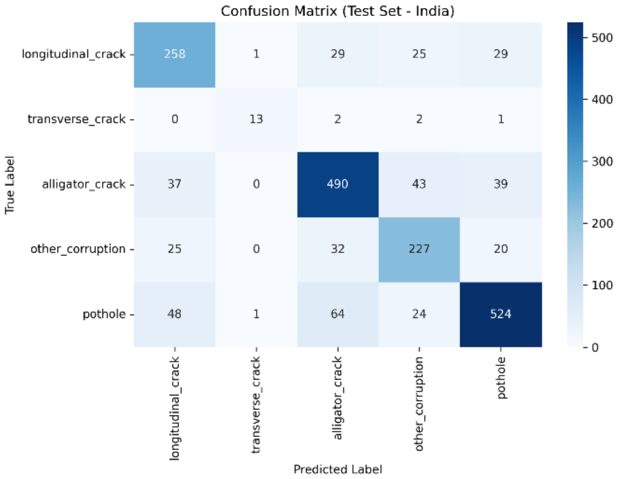

Figure 5. Confusion matrix showing class-wise prediction performance of MobileNetV2 with SSL.

The detailed view of class-wise prediction is represented using confusion matrix, shown in Figure 5. The matrix is showing strong diagonal distribution, indicating accurate prediction across most of the classes.

However, few off-diagonal elements reveal some misclassifications. Specifically, confusion is observed between alligator cracks and longitudinal cracks, and between potholes and other corruption classes. The overlapping surface patterns and structural textures of cracks that visually resemble each other are contributed for these misclassifications. For example, the zigzag and interconnected nature of alligator cracks may visually appear as fragmented longitudinal cracks features, and dark patches of road staining or shadowing may resemble as potholes. Such confusion patterns highlight difficulty in differentiating structurally similar defects, particularly under variations in occlusion, variations in lighting and resolution inherent to real-world road imagery.

4.3 Training and Convergence Analysis

The training and validation accuracy and loss curve plots provide insight into the model's convergence behavior over 25 epochs of supervised finetuning. The model successfully learns task-specific features, as shown by the steady and consistent rise in training accuracy, which reaches about 91%. Validation accuracy, though slightly lower, converges towards 78% closely aligning with final evaluation metrics. There is no major underfitting, as both training and validation accuracy improve over time. Minor overfitting is observed but the gap is narrow, indicating strong generalization capability. Figure 6 and Figure 7, are showing plots for accuracy and loss curves for supervised classification.

The training loss drops rapidly from 1.37 to 0.12, showing effective convergence after SSL pretraining. The validation loss peaks around 0.8 to 0.9, with minor oscillations. To mitigate this, several regularization techniques were explored, including dropout, early stopping, and different data augmentation strategies. Initial experiments with dropout layers using dropout values ranging 0.3 to 0.5 are conducted by incorporating them into the classification head and intermediate blocks.

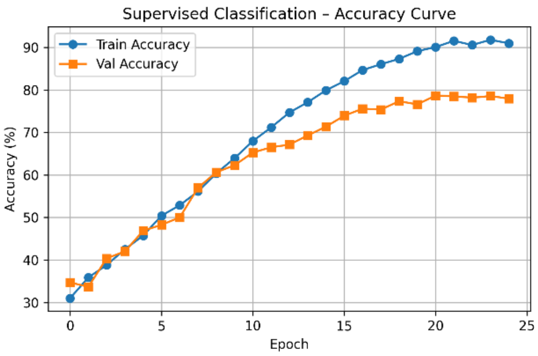

Figure 6. Training accuracy vs Validation accuracy plot for MobileNetV2.

These setting were found to reduce overfitting slightly, but they also lead to a noticeable drop in classification accuracy, particularly on fine-grained damage classes. Similarly, early stopping based on validation loss peak value is applied and observed that overfitting is prevented to some extent. However, training is often halted prematurely before the model reaches its optimal representation capacity. Extensive data augmentation experiments are also carried out to improve generalization. Augmentation such as horizontal and vertical flipping, color jittering, random cropping, brightness and contrast adjustments (Mumuni & Mumuni, 2022)are tested, these transformations are surprisingly found to degrade performance. A road damage patterns e.g. cracks, potholes, other surface corruptions have strong spatial and directional characteristics and many augmentations like flips are found to distort these patterns, causing the model to learn less semantically meaningful features. This issue in patch-based SSL arises due to geometric consistency is required to preserve the performance of pretext task representations.

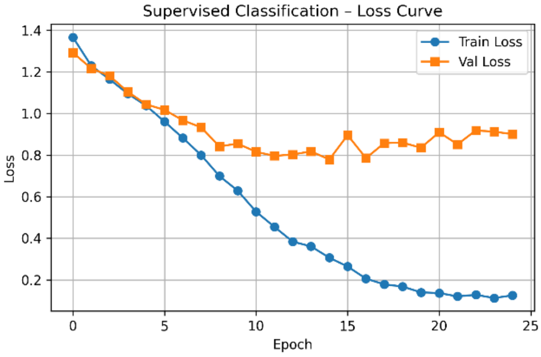

Figure 7. Training loss vs Validation loss plot for MobileNetV2.

Through experimentations, it is discovered that simple rotation-based augmentation offers a balanced trade-off. The diversity of training data is enhanced, while the geometric semantics of road damages are preserved (T. Chen et al., 2020). This strategy is observed to help overfitting in the later epochs without significantly impacting accuracy, as reflected in the stabilized validation curves. Therefore, rotation augmentation is adopted as the primary regularization method, while more aggressive augmentations are excluded to their negative impact on model performance. These observations highlight the sensitivity of road damage classification task to geometric distortions and support the decision to apply minimal, domain aware regularization strategy.

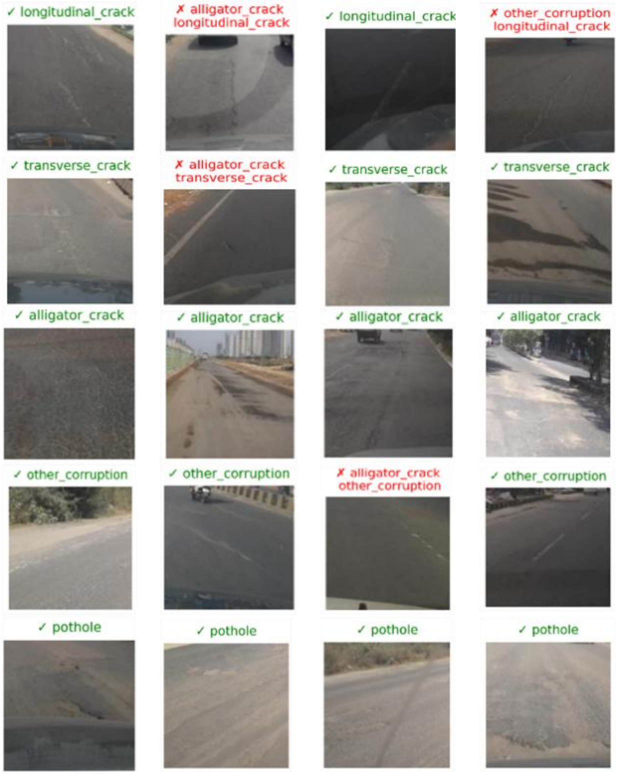

Figure 8. Classification prediction results across all damage classes, including both correct and incorrect predictions.

4.4 Visual Assessment of Model Predictions

Further, showing performance of MobileNetV2 with SSL on the Indian test site data Figure 8, represents a set of qualitative predictions for randomly selected samples from each damage class showing both correct and incorrect predictions. The model consistently predicts classes like transverse crack, pothole and alligator crack with high confidence, even under challenging illumination and occlusion effects. Most misclassification occurs between visually similar classes like transverse crack and alligator crack, longitudinal cracks and other corruption class. Other corruption class includes defects such as bump on road, rutting due to tires, crosswalk blur, white line blur, manholes etc (Arya et al., 2024). which introduces confusion due to its broad definition, leading to some overlap with structural crack patterns. This misclassification can be ambiguous even for human annotators. Incorrect predictions are relatively few in frequency compared to correct classifications, showing strong class separation achieved through self-supervised learning which is key challenge in real-world damage inspection scenarios.

4.5 Damage Class Prediction Analysis of Competitive SSL Models

To further support the quantitative evaluation, a detailed comparison of confusion matrices is performed for the top-performing baseline models trained with self-supervised learning, i.e., EfficientNet-B0, ResNet-18 and MobileNetV3-Large. This enables an understanding of model behavior at the class level and describing strength and limitation.

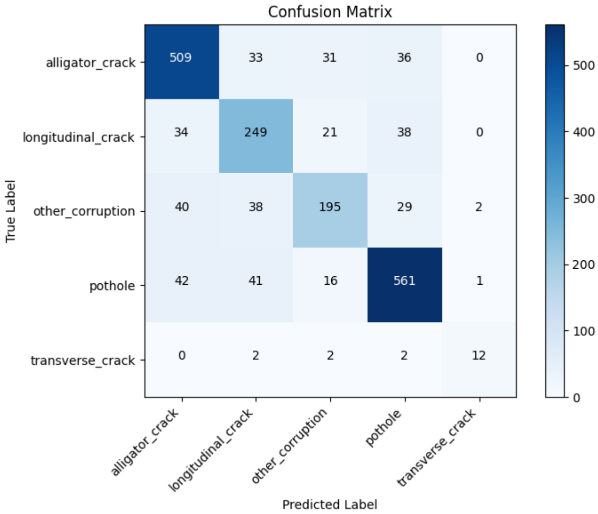

Figure 9. Confusion matrix for EfficientNet-B0 with SSL.

The confusion matrix of EfficientNet-B0 with SSL as shown in Figure 9, shows strong performance for potholes and alligator cracks prediction, with F1-scores of 0.85 and 0.83 respectively. The most common misclassifications are between longitudinal cracks and other corruption class, and minor between longitudinal cracks and alligator cracks. The matrix displays a primarily diagonal distribution in spite of these overlaps, and the overall classification accuracy is 78.90% with a macro F1-score of 0.7579. Transverse cracks, which are underrepresented in the dataset, are predicted with a high precision with 0.80, through their recall drops to 0.67, showing a tendency of conservative classification.

In the case of ResNet-18 with SSL as shown in figure 10, the model achieved the highest overall accuracy of 80.30% along with a macro F1-score of 0.7807 and weighted F1-score of 0.8038. The confusion matrix indicates better separability between structurally similar classes, such as longitudinal cracks and other corruption, which are less frequently misclassified relative to the other models. However, the matrix also reveals increased confusion between longitudinal cracks and potholes, as well as between alligator cracks and other corruption class. The model handles transverse cracks with relatively high precision of 0.76 and recalls of 0.72, reflecting a more balanced classification even in the presence of class imbalance.

The confusion matrix of MobileNetV3-Large with SSL as shown in Figure 11, shown slightly lower performance in comparison, with an overall accuracy of 77.61% and a macro F1-score of 0.7408. The confusion matrix reveals considerable overlap between potholes and other corruption, as well as between longitudinal cracks and alligator cracks. These misclassifications are likely due to similar surface textures and directional crack patterns, which the model finds difficult to differentiate. Further, the transverse cracks with very low samples result in a relatively low recall of 0.61, indicating that a significantly some true samples are missed, despite with decent precision of 0.79. Across all three models, potholes consistently emerge as the most reliably detected class, due to their distinct shape and darker pixel distribution. While, longitudinal cracks and other corruption class facing persistent challenges, with moderate F1-scores. MobileNetV3-Large and Small are more recent architectures, our patch-based SSL framework of MobileNetV2 showed higher accuracy than MobileNetV3-Large and MobileNetV3-Small. This difference is likely due to the way these architectures represent and process small patches with important structural details. MobileNetV3 models, designed for extreme efficiency, use fewer intermediate feature channels and hard-swish activations, which can reduce the ability to preserve important local features necessary for crack detection. MobileNetV2's inverted residual structure and ReLU6 activations maintain more informative feature maps, enabling better transfer from rotation-based pretraining to the target classification problem, resulting in higher accuracy under the same input conditions. And patch-based SSL setting gave advantage to the models for achieving better accuracy.

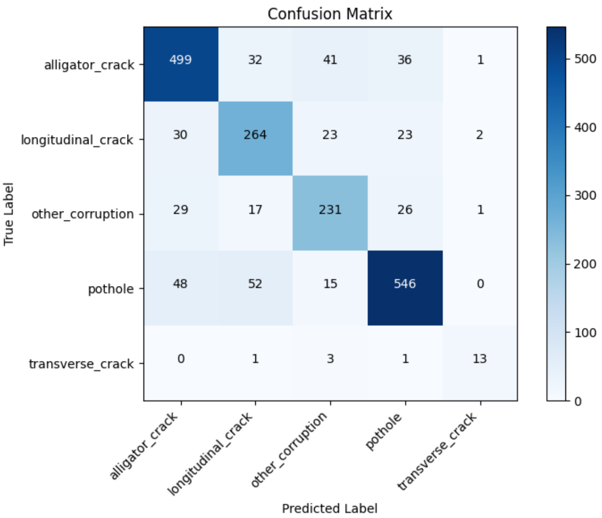

Figure 10. Confusion matrix for ResNet-18 with SSL.

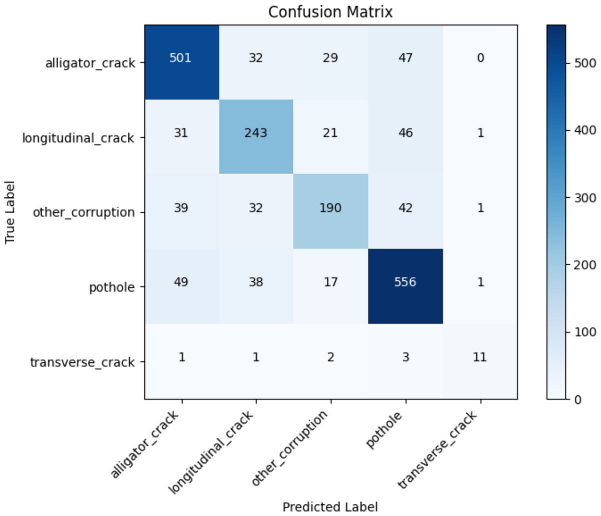

Figure 11. Confusion matrix for MobileNetV3-Large with SSL.

Although ResNet-18 with SSL slightly outperforms others, the MobileNetV2 with SSL still demonstrates the best balance of accuracy, efficiency, and generalization, especially under constrained computational settings.

5. Discussion

The efficacy of the proposed MobileNetV2 model integrated with patch-based self-supervised learning (SSL), a comprehensive comparison is conducted against several state-of-the-art CNN architectures across key performance metrics such as classification accuracy, macro and weighted F1-scores, model complexity in terms of parameters and FLOPs, memory used, inference time and throughput. The comparative evaluation is summarized in Table 2 and Table 3; It offers critical insights into the trade-offs between model performance and computational efficiency. Factors that are considered essential for real-world deployment of road damage classification systems, particularly in low-resource environments such as mobile or edge computing devices.

The results show that MobileNetV2 with SSL, found to achieve highly competitive classification performance while maintaining remarkable computational efficiency. Specifically, an accuracy of 78%, macro F1-score of 0.770, and weighted F1-score of 0.780 are attained. ResNet-18 with SSL, whose accuracy is recorded at 80.30%, this marginal accuracy gain is obtained with nearly 5× more parameters, 6× more floating-point operations and over 5× the memory footprint required by ResNet-18. Furthermore, its inference time is observed to be faster by only 2ms, while a comparable throughput is offered 99.87 FPS for ResNet-18 and 80.47 FPS for MobileNetV2 both with SSL. The results shows that the MobileNet and ResNet backbones is more stable on an unseen dataset (Trinh et al., 2024). These findings are used to emphasize that the proposed pipeline delivers a far more optimal trade-off between accuracy and resource utilization.

When examined against EfficientNet-B0 with SSL, a widely adopted architecture known for its parameter efficiency, the proposed pipeline still demonstrated to be superior in several dimensions.

Model Variant	SSL Used	Accuracy	Macro F1	Weighted F1
MobileNetV2	**Yes**	**0.7800**	**0.7700**	**0.7800**
MobileNetV2	No	0.5610	0.4170	0.5520
MobileNetV3-Large	Yes	0.7760	0.7400	0.7740
MobileNetV3-Large	No	0.6980	0.6889	0.6992
MobileNetV3-Small	Yes	0.7642	0.7463	0.7642
MobileNetV3-Small	No	0.7001	0.6774	0.6986
EfficientNet-B0	Yes	0.7890	0.7579	0.7882
EfficientNet-B0	No	0.7540	0.6650	0.7550
ResNet-18	Yes	0.8030	0.7807	0.8038
ResNet-18	No	0.7760	0.7460	0.7750

Table 2. Comparison of model variants with and without self-supervised learning (SSL), reported using accuracy, macro F1 and weighted F1 scores.

Although EfficientNet-B0 slightly outperforms MobileNetV2 with SSL in terms of accuracy 78.90 vs 78.00%, this is achieved at the cost of nearly double the parameters i.e. 4.01M vs 2.23M, higher FLOPs 133.96M, and significantly slower inference speed 16.48ms vs 12.43ms, which results in a reduced throughput of 60.69 FPS. This further strengthens MobileNetV2 when integrated with SSL preserving both computational efficiency and competitive classification performance.

It is also noted that consistent performance improvements are introduced by SSL across all model variants. MobileNetV2 shown significant improvement of +21.9% accuracy with SSL, indicating that the self-supervised pretext tasks are able to provide highly transferable features suitable for road damage classification (Cao & Xiang, 2020; Zhao et al., 2024).Similar trends are observed for other models, although the relative gains vary for EfficientNet-B0 and ResNet-18. Suggesting that lighter models such as MobileNetV2 significantly benefited by SSL, possibly due to their limited representational capacity under purely supervised learning method.

Model Variant	SSL Used	Params	FLOPs	Inference Time
MobileNetV2	**Yes**	**2.23 M**	**104.18 M**	**12.43 ms**
MobileNetV2	No	2.23 M	104.18 M	12.55 ms
MobileNetV3-Large	Yes	4.21 M	76.97 M	11.91 ms
MobileNetV3-Large	No	4.21 M	76.97 M	11.67 ms
MobileNetV3-Small	Yes	1.52 M	20.27 M	6.94 ms
MobileNetV3-Small	No	1.52 M	20.27 M	7.42 ms
EfficientNet-B0	Yes	4.01 M	133.96 M	16.48 ms
EfficientNet-B0	No	4.01 M	133.96 M	18.81 ms
ResNet-18	Yes	11.18 M	595.86 M	10.01 ms
ResNet-18	No	11.18 M	595.86 M	12.31 ms

Table 3. Comparison of model variants with and without self-supervised learning (SSL), showing parameter count (Params), computational cost (FLOPs), and average inference time per image.

MobileNetV3-Small are identified as more lightweight than MobileNetV2, it achieves lower classification accuracy of 76.42%. The table 2 and table 3 both are showing comparison of different model variants with different performance metrics. This suggest that while model size is recognized as significant, MobileNetV2 has maintained a balance between feature depth and parameter. Although MobileNetV3-Large and ResNet-18 have shown high performance, their larger computational load makes them less suitable for low-end resources.

6. Conclusion

The study proposes, a lightweight and efficient road damage classification framework, leveraging MobileNetV2 architecture enhanced with patch-based self-supervised learning. The approach was evaluated on the Indian test site subset of the RDD2022 benchmark dataset with comprehensive experiments conducted across multiple deep convolutional architectures.

The results demonstrated that the proposed method not only delivers competitive classification accuracy of 78.00% but also significantly reduces computational overhead achieving up to 5× fewer parameters and FLOPs compared to ResNet-18, with significant performance trade-off. The SSL improved accuracy gains across all baseline models, highlighting SSL pretraining’s effectiveness, especially for low-capacity models trained with low labeled input. The quantitative and qualitative metrics of performance confirmed that the proposed model with SSL generalizes well across diverse road damage categories under occlusion, varying lighting, and surface texture conditions. The architecture is also sustaining inference speed above 80 FPS with a model size of under 9 MB, satisfying the constraints of mobile, drone-based and edge computing platforms.

Future work may extend this framework to multi-domain or cross-domain damage classification using a patch-based SSL approach applied to the complete RDD2022 dataset, which covers six countries, India, Japan, the Czech Republic, China, the United States, and Norway and may also incorporate severity assessment and pavement maintenance prioritization. This will enhance the model's ability to generalize across varying road types, environmental conditions and damage types. Such a study would allow us to investigate domain adaptation and domain generalization which are crucial for developing globally deployable road damage detection systems.

Acknowledgements

This work was supported by the Department of Science and Technology (DST), Ministry of Science and Technology, Government of India, under project grant number NGP/GS-09/Manohar/MNNIT/UP/2023.

References

Alfarrarjeh, A., Trivedi, D., Kim, S.H., Shahabi, C., 2018. A deep learning approach for road damage detection from smartphone images. In: *Proceedings of the 2018 IEEE International Conference on Big Data (Big Data)*, 10–13 Dec 2018, Seattle, WA, USA. IEEE, Piscataway, NJ, pp. 5201–5204. https://doi.org/10.1109/BigData.2018.8621899.

Abdelmenam, A., Gaber, A., 2025. *RDD2022 dataset*. Kaggle. https://www.kaggle.com/datasets/aliabdelmenam/rdd-2022/data (accessed 24 June 2025)

Alqethami, S., Alghamdi, S., Alsubait, T., Alhakami, H., 2022. RoadNet: efficient model to detect and classify road damages. *Applied Sciences (Switzerland)*, 12(22), 11529. https://doi.org/10.3390/app122211529.

Arya, D., Maeda, H., Ghosh, S.K., Toshniwal, D., Mraz, A., Kashiyama, T., Sekimoto, Y., 2021. Deep learning-based road damage detection and classification for multiple countries. *Automation in Construction*, 132, 103935. https://doi.org/10.1016/j.autcon.2021.103935.

Arya, D., Maeda, H., Ghosh, S.K., Toshniwal, D., Sekimoto, Y., 2024. RDD2022: A multi-national image dataset for automatic road damage detection. *Geoscience Data Journal*, 11(4), 846–862. https://doi.org/10.1002/gdj3.260.

Arya, D., Maeda, H., Sekimoto, Y., Omata, H., Ghosh, S.K., Toshniwal, D., Sharma, M., Pham, V.V., Zhong, J., Al-Hammadi, M., Shami, M.B., Nguyen, D., Cheng, H., Zhang, J., Klein-Paste, A., Mork, H., Lindseth, F., Seto, T., Mraz, A., Kashiyama, T., 2022. RDD2022 – The multi-national road damage dataset released through CRDDC'2022. *Figshare*. https://doi.org/10.6084/m9.figshare.21431547.v1.

Bouhsissin, S., Assemlali, H., Sael, N., 2025. Enhancing road safety: a convolutional neural network-based approach for road damage detection. *Machine Learning with Applications*, 20, 100668. https://doi.org/10.1016/j.mlwa.2025.100668.

Cao, L., Xiang, W., 2020. Application of convolutional neural network based on transfer learning for garbage classification. In: *Proceedings of the 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC)*, Chongqing, China, 12–14 June 2020. IEEE, Piscataway, NJ, pp. 1032–1036. https://doi.org/10.1109/ITOEC49072.2020.9141699.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations. *arXiv preprint*, arXiv:2002.05709. https://arxiv.org/abs/2002.05709.

Chorada, R., Kriplani, H., Acharya, B., 2023. CNN-based real-time pothole detection for avoidance road accident. In: *Proceedings of the 7th International Conference on Intelligent Computing and Control Systems (ICICCS 2023)*, 17–19 May 2023, Madurai, India. IEEE, Piscataway, NJ, pp. 700–707. https://doi.org/10.1109/ICICCS56967.2023.10142488.

Gidaris, S., Singh, P., Komodakis, N., 2018. Unsupervised representation learning by predicting image rotations. *International Conference on Learning Representations (ICLR)*. arXiv preprint arXiv:1803.07728. https://arxiv.org/abs/1803.07728.

Hasan, M.M., Sakib, S., Deb, K., 2022. Road damage detection and classification using deep neural network. In: *Proceedings of the 4th International Conference on Electrical, Computer and Telecommunication Engineering (ICECTE 2022)*, 29–31 Dec 2022, Rajshahi, Bangladesh. IEEE, Piscataway, NJ. https://doi.org/10.1109/ICECTE57896.2022.10114508.

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q.V., Adam, H., 2019. Searching for MobileNetV3. *arXiv preprint*, arXiv:1905.02244. https://arxiv.org/abs/1905.02244.

India Brand Equity Foundation (IBEF), 2025. *Road infrastructure in India.* https://www.ibef.org/industry/roads-india (accessed 14 June 2025).

Maeda, H., Sekimoto, Y., Seto, T., Kashiyama, T., Omata, H., 2018. Road damage detection and classification using deep neural networks with smartphone images. *Computer-Aided Civil and Infrastructure Engineering*, 33(12), 1127–1141. https://doi.org/10.1111/mice.12387.

Ministry of Road Transport and Highways (MoRTH), 2022. *Basic road statistics of India 2022.* Government of India. https://morth.nic.in/sites/default/files/Basic_Road_Statistics_of_India_2022.pdf (accessed 14 June 2025).

Mumuni, A., Mumuni, F., 2022. Data augmentation: a comprehensive survey of modern approaches. *Array*, 16, 100258. https://doi.org/10.1016/j.array.2022.100258.

Trinh, L., Anwar, A., Mercelis, S., 2024. Multiple data sources and domain generalization learning method for road surface defect classification. *arXiv preprint*, arXiv:2407.10197. https://arxiv.org/abs/2407.10197.

Zhao, Z., Alzubaidi, L., Zhang, J., Duan, Y., Gu, Y., 2024. A comparison review of transfer learning and self-supervised learning: definitions, applications, advantages and limitations. *Expert Systems with Applications*, 242, 122807. https://doi.org/10.1016/j.eswa.2023.122807.

Asif, A., Zakwan, M., Mahmood, M.H., 2024. Exploring the impact of dataset content on self-supervised pretext rotation learning: A comparative analysis using MobileNet. In: *Proceedings of the 21st International Bhurban Conference on Applied Sciences and Technology (IBCAST 2024)*, 107–112. https://doi.org/10.1109/IBCAST61650.2024.10877184.

Chen, P.H., Hsieh, J.W., Hsieh, Y.K., Chang, C.W., Huang, D.Y., 2025. Cross-scale overlapping patch-based attention network for road crack detection. *IEEE Transactions on Intelligent Transportation Systems*, 26(6), 7587–7599. https://doi.org/10.1109/TITS.2025.3558279.