IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 43, No. 3, June, 2026, pp. 1125-1138
Journal homepage: http://iieta.org/journals/ts

Self-Supervised Learning for Road Distress Image Recognition

Yigang You, Wenkai Jiang*

Business School, Guangxi University of Finance and Economics, Nanning 530007, China

Corresponding Author Email: jiangwenkai@gxufe.edu.cn

https://doi.org/10.18280/ts.430305

Received: 12 December 2025
Revised: 10 April 2026
Accepted: 19 April 2026
Available online: 30 June 2026

Keywords:
self-supervised learning, road distress recognition, contrastive learning, masked autoencoding, multi-scale feature fusion, cross-domain adaptation

ABSTRACT

Road distress intelligent recognition constitutes a core enabling technology for smart transportation infrastructure maintenance. Existing deep learning-based detection methods rely heavily on large-scale manually annotated data, which incurs high labeling costs and strong data dependence. Mainstream self-supervised learning algorithms follow generic visual learning paradigms without incorporating the visual distribution characteristics of road distresses, failing to accommodate the intricate inspection environments of actual roadway networks. To address these issues, an end-to-end fully self-supervised recognition framework tailored to road distress images was constructed, enabling high-accuracy distress detection and cross-domain adaptive recognition driven by unlabeled data. First, a distress saliency representation system was built by integrating multi-directional gradient features and adaptive local texture features, and a distress-aware differentiated masking strategy was designed to guide the pretraining process toward focusing on core distress regions, thereby incorporating domain-specific prior knowledge of road scenes. Second, a dual-branch self-supervised pretraining architecture coupling contrastive learning and masked reconstruction was established, with a dynamic loss balancing mechanism to fuse global discriminative features and local textural details, compensating for the representational deficiencies of a single self-supervised paradigm. Meanwhile, an attention-guided multi-scale feature adaptive fusion module was constructed to hierarchically capture shallow structural details, intermediate geometric structures, and high-level semantic features, enabling adaptive aggregation of distress features across a wide range of scales and matching the inherent large-scale variation of pavement distresses. Finally, a lightweight visual prompt adaptation mechanism was introduced, with maximum mean discrepancy constraint for feature alignment on unlabeled target domains, effectively suppressing domain shift interference without updating backbone parameters. The proposed approach effectively breaks through the technical bottlenecks of scarce annotation resources and insufficient cross-domain adaptability in road distress detection, providing reliable technical support for large-scale intelligent roadway network inspection.

1. INTRODUCTION

Road infrastructure serves as the primary carrier of transportation systems, and its in-service health condition directly determines roadway network safety and traffic maintenance efficiency [1, 2]. Under the coupled effects of prolonged traffic loading, natural weathering, and environmental erosion, pavement surfaces are highly susceptible to typical distresses such as cracks, potholes, and alligator cracking. Without timely and accurate detection and maintenance, accelerated performance deterioration of the pavement may be induced, traffic safety hazards may be triggered, and roadway network operation and maintenance costs may be significantly increased. Traditional road distress detection has relied heavily on manual visual inspection, an operational mode constrained by subjective human experience, limited field of view, and low inspection efficiency, rendering it incapable of accommodating large-scale and routine roadway network health monitoring demands. With the rapid advancement of visual perception and deep learning technologies, image-based automated distress detection methods have gradually supplanted conventional manual inspection and have become the mainstream technical approach for intelligent transportation maintenance [3, 4]. At present, the dominant distress recognition models are built upon the supervised learning paradigm, where model performance is entirely dependent on large-scale, high-precision pixel-level annotated datasets. However, road distress images are characterized by complex scenes and diverse distress morphologies; manual pixel-wise annotation is labor-intensive and time-consuming, and inconsistencies in annotation criteria across different human labelers frequently lead to annotation bias. The prohibitively high data annotation costs and data quality deficiencies severely impede the large-scale deployment of intelligent detection models in real-world roadway scenarios [5, 6]. Self-supervised learning techniques are capable of generating supervisory signals from the inherent properties of the data themselves, thereby eliminating the reliance on manual annotations and enabling the acquisition of generalizable and robust visual

representations from massive unlabeled road images, offering a novel pathway to overcome the data bottleneck in road distress detection [7, 8]. Although self-supervised learning has achieved groundbreaking progress in generic visual recognition tasks, its application to the domain-specific scenario of road distress detection remains immature, with insufficient adaptation to the visual characteristics of pavement distresses and the complexity of road scenes, leaving numerous critical technical challenges yet to be resolved [9, 10].

Current self-supervised learning methods for road distress recognition still follow the training paradigms established for generic vision tasks, without domain-specific optimization tailored to road scenes and distress characteristics, resulting in notable deficiencies in overall recognition performance and scene adaptability [11, 12]. The proxy task designs of existing algorithms are generally devoid of prior knowledge specific to the road distress domain, with masking strategies and data augmentation policies both adopting generic fixed configurations that treat all image regions indiscriminately. Road distresses possess distinctive visual characteristics—cracks exhibit continuous linear gradient features, whereas potholes manifest as localized textural discontinuities. Generic training strategies fail to steer the model toward focusing on the effective core regions of distresses, leading to insufficient representational specificity acquired during pretraining [13, 14]. Meanwhile, each single self-supervised learning paradigm suffers from inherent representational deficiencies: contrastive learning emphasizes the extraction of global semantic discriminative features from images, yet struggles to capture fine-grained local textural details of minute distresses such as cracks; masked autoencoders are dedicated to pixel-level local texture reconstruction, but tend to overfit low-level visual features while losing distress-specific semantic information. The complementary advantages of these two mainstream paradigms cannot be synergistically integrated, resulting in incomplete representational information [15, 16]. Furthermore, road distresses exhibit extreme scale variability—fine cracks occupy exceedingly small pixel extents, whereas potholes and large-area alligator cracking span considerably broader coverage. Existing self-supervised models predominantly adopt single-scale feature extraction architectures, which are incapable of simultaneously capturing and modeling multi-scale distress features with sufficient precision, thus failing to balance detection accuracy across both fine-grained and large-area distresses [17, 18]. Finally, significant domain shift issues exist among road images acquired under varying geographic regions, illumination conditions, and pavement material types [19, 20]; the general representations learned by current self-supervised algorithms lack sufficient robustness [21, 22], resulting in severe performance degradation during cross-scene deployment, which cannot meet the demanding requirements of complex and diverse real-world roadway network inspection [23].

To address the aforementioned industry pain points and technical deficiencies, an end-to-end fully self-supervised recognition framework for road distress images is constructed, forming a multi-level, full-pipeline technical optimization system. A differentiated masking generation strategy is designed, grounded in the visual priors of road distresses, so that focused learning on distress regions is enabled during the pretraining process. A dual-branch collaborative pretraining architecture, combining contrastive and reconstruction objectives, is established so that global discriminative representations and local detailed representations are fused. An attention-guided multi-scale feature fusion module is built so that the challenge of modeling multi-scale distresses is addressed. A lightweight visual prompt adaptation mechanism is introduced so that efficient cross-domain scene transfer is achieved. These four core modules are progressively refined and synergistically coupled, forming a complete technical closed loop, by which the performance limitations of existing algorithms are effectively broken through, the dependence of the model on annotated data is substantially reduced, and distress recognition accuracy and generalization capability in complex scenarios are significantly enhanced.

The overall structure of this work is organized as follows: In Section 2, the core modules and technical principles of the proposed fully self-supervised distress recognition framework are systematically elaborated. In Section 3, the effectiveness and superiority of the algorithm are validated through multiple comparative experiments and ablation studies. In Section 4, the research findings are summarized, the limitations of the current technology are reviewed, and directions for future research are outlined.

2. METHODOLOGY

To address the critical issues in existing self-supervised learning methods for road distress recognition—namely, the lack of domain-specific priors, singular representational dimensions, insufficient multi-scale feature modeling, and inadequate cross-domain generalization capability—an end-to-end fully self-supervised distress image recognition framework tailored to complex road scenes is constructed. Through a multi-stage coupled functional module system, high-accuracy distress representation learning and adaptive scene transfer are achieved driven by unlabeled data. This framework abandons the generic training paradigms established for general vision tasks. First, a differentiated masking generation strategy is designed, grounded in the inherent texture and gradient visual features of road distresses, so that focused learning on distress regions is enabled during the pretraining process, and domain-specific prior knowledge of road scenes is incorporated into the self-supervised representation learning pipeline. On this basis, a dual-branch pretraining architecture that couples contrastive learning and masked reconstruction is established, by which global semantic discriminative features and local detailed features are fused, thereby enhancing the representational capacity of the model. To accommodate the wide-ranging scale distribution characteristics of pavement distresses, an attention-guided multi-scale feature adaptive fusion mechanism is introduced into the framework, hierarchically integrating detail-level, structural, and semantic features from different network stages so that accurate modeling and aggregation of multi-scale distress features are achieved. Finally, a lightweight visual prompt adaptation mechanism is employed to perform cross-domain feature alignment, effectively suppressing performance degradation caused by scene domain shift while preserving parameter efficiency. Each core module operates in a synergistic manner, following the progressive logic of feature screening, representation optimization, multi-scale enhancement, and cross-domain adaptation, thereby forming a comprehensive self-supervised learning technical system. The design principles, network architectures, and optimization

methodologies of each module are elaborated in detail in the following subsections.

2.1 Distress-aware differentiated masking generation strategy

The pretraining performance of masked autoencoders is highly correlated with the masking sampling strategy. The global random masking mechanism adopted by general-purpose algorithms indiscriminately masks and learns from all image patches without leveraging semantic differences in image content to optimize training emphasis. The feature distribution of road distress images exhibits significant scene-specific characteristics, with distress regions densely concentrating edge mutations and texture anomalies, which constitute the core effective information supporting distress recognition tasks. A uniform masking sampling approach attenuates the training weight of distress regions, rendering the model inefficient in capturing distress-specific features and substantially degrading the learning quality of self-supervised pretraining. To address this issue, a distress-aware differentiated masking generation strategy is constructed, grounded in the visual distribution patterns of road distresses, by which adaptive masking sampling is achieved through quantitative assessment of regional distress saliency, guiding the model to focus on high-value distress regions during the pretraining phase and accomplishing domain-specific representation learning tailored to road scenes.

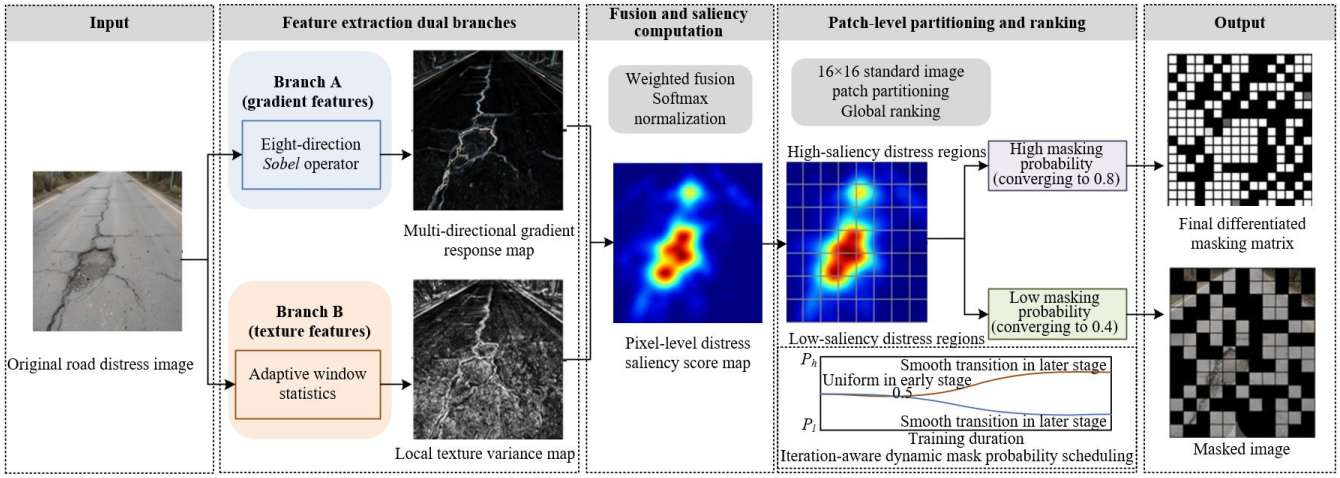

Figure 1. Flowchart of the distress-aware differentiated masking generation strategy

To precisely characterize the intra-image differences in distress saliency, a refined saliency detection system is built by integrating multi-directional gradient features and adaptive texture features, enabling accurate localization of potential distress regions. The flowchart of the distress-aware differentiated masking generation strategy is presented in Figure 1. Gradient feature extraction is performed using an eight-direction Sobel operator, which extends the conventional four-direction operator by incorporating diagonal gradient computations, thereby effectively enhancing the feature response to irregular distresses such as diagonal cracks. The global gradient response map is computed as follows:

$$G=\sum_{d\in D}|\nabla_d X|, \quad D=\{0^\circ,22.5^\circ,45^\circ,67.5^\circ,90^\circ,112.5^\circ,135^\circ,157.5^\circ\} \tag{1}$$

where, X denotes the input road distress image, and $\nabla_d X$ represents the gradient response of the image at the corresponding angle. Meanwhile, an adaptive window strategy is introduced for local texture variance computation, in which the statistical window size is dynamically adjusted according to the local gradient density. Large windows are employed in texture-uniform pavement regions, whereas small windows are adopted in texture-complex regions suspected of containing distresses, thereby improving the accuracy of texture feature detection while maintaining computational efficiency. The local texture variance is formulated as:

$$V(i,j)=\frac{1}{k^2}\sum_{p=i-\lfloor k/2\rfloor}^{i+\lfloor k/2\rfloor}\sum_{q=j-\lfloor k/2\rfloor}^{j+\lfloor k/2\rfloor}\left(X(p,q)-\mu_{i,j}\right)^2 \tag{2}$$

where, k denotes the adaptive window size and $\mu_{i,j}$ represents the mean pixel value within the local window. After normalization, the gradient feature map and the texture variance map are fused through a weighted summation, and the pixel-level distress saliency score map is obtained via Softmax mapping, expressed as:

$$S=softmax(\alpha\cdot\hat{G}+\beta\cdot\hat{V}) \tag{3}$$

where, $\hat{G}$ and $\hat{V}$ denote the normalized gradient response map and texture variance map, respectively. The fusion weights α and β are set to 0.6 and 0.4, respectively. This parameter combination is determined through grid search on the RDD2022 validation set samples, and is shown to optimally accommodate the feature distribution characteristics of road distresses.

Based on the pixel-level saliency scores, patch-level differentiated masking construction is performed, enabling precise alignment between the masking rules and the distress distribution. The input image is uniformly partitioned into standard 16×16 patches, and all patches are globally ranked according to their average saliency scores, by which high-saliency distress regions and low-saliency background regions are distinguished. Differentiated masking probabilities are assigned to these two categories so that adaptive focused learning is achieved during the pretraining task. The generation rule for the final masking matrix is defined as:

$$M_i=\begin{cases}1, \text{ if } i\in\text{high-saliency region and } rand(\,)<r_{high}\\ 1, \text{ if } i\in\text{low-saliency region and } rand(\,)<r_{low}\\ 0, \text{ otherwise}\end{cases} \tag{4}$$

where, M_i denotes the masking state of the i-th image patch, with a value of 1 indicating that the patch participates in the masked reconstruction task and a value of 0 indicating that the patch features are directly preserved; *rand*( ) represents a random number uniformly distributed in the interval; and r_{high} and r_{low} correspond to the masking probabilities for high- and low-saliency regions, respectively. Through this mechanism, a higher reconstruction learning weight is imposed on the core distress regions, enabling thorough exploitation of distress texture and structural features, while redundant learning from ineffective background regions is attenuated, thereby enhancing the effectiveness and specificity of the pretraining representations.

To balance model training stability and targeted feature learning, an iteration-aware dynamic masking probability scheduling mechanism is designed, enabling a smooth transition of the training strategy. The masking probabilities for high- and low-saliency regions are adaptively updated with respect to the training epoch, with the scheduling functions formulated as:

$$r_{high}(t)=0.5+0.3\cdot\sigma\left(\frac{t\text{-}50}{20}\right),\ r_{low}(t)=0.5\text{-}0.1\cdot\sigma\left(\frac{t\text{-}50}{20}\right) \tag{5}$$

where, t denotes the current training iteration epoch, and σ( ) represents the Sigmoid activation function. In the early training stage, the masking probabilities for the two types of regions are kept consistent, both maintained at a level of 0.5, so that the model preferentially learns the fundamental visual features of road scenes and establishes a holistic understanding of pavement images. As training iterations proceed, the masking probabilities are smoothly adjusted within the specified intervals, ultimately converging to fixed values—the high-saliency region masking probability converges to 0.8, and the low-saliency region masking probability converges to 0.4. This progressive scheduling strategy effectively mitigates the representational learning bias induced by differentiated masking at the early training stage, prevents the model from converging to suboptimal local minima, gradually reinforces the model's capacity for learning distress features, and ensures the overall effectiveness of self-supervised pretraining.

2.2 Contrastive-reconstruction dual-branch collaborative self-supervised pretraining

Masked self-supervised learning and contrastive self-supervised learning possess distinct advantages in feature learning; a single learning paradigm is insufficient to simultaneously satisfy both the global semantic discriminability and the local textural granularity required for road distress recognition. Building upon the differentiated masking sampling mechanism established in the preceding section, a contrastive and reconstruction dual-branch collaborative self-supervised pretraining framework is constructed, in which deep integration of the two self-supervised tasks is achieved through a shared feature encoding network. The architecture of the contrastive-reconstruction dual-branch collaborative self-supervised pretraining framework is illustrated in Figure 2. The overall framework consists of a shared visual Transformer encoder, a contrastive learning task head, and a masked reconstruction task head. The encoder adopts the ViT-B/16 base architecture, comprising twelve Transformer encoding layers, with each layer equipped with twelve attention heads, a hidden feature dimension of 768, and a feed-forward network dimension of 3072. The encoder parameters are first initialized on the generic image dataset ImageNet-21K to acquire basic visual representational capacity, and subsequently domain-adaptive pretraining is performed on unlabeled road distress images, providing a stable and generalizable feature extraction foundation for the dual-branch collaborative learning.

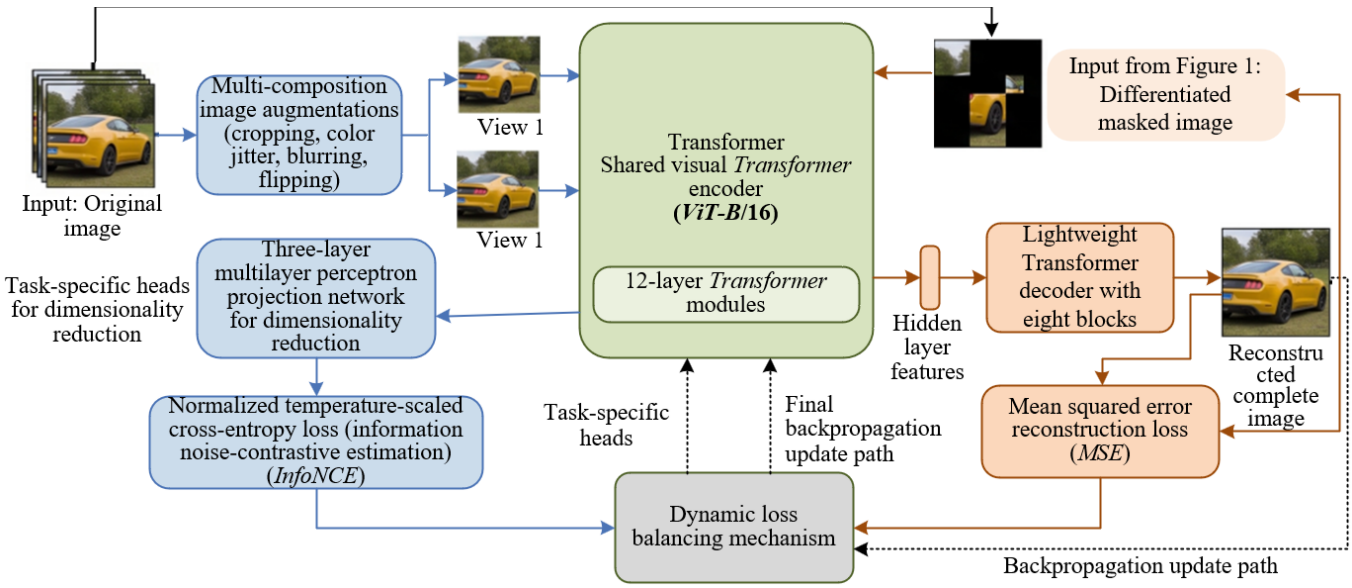

Figure 2. Architecture of the contrastive-reconstruction dual-branch collaborative self-supervised pretraining framework

The contrastive learning branch is centered on the instance discrimination task, by which the model is constrained to learn global semantic features through multi-view image transformations, thereby enhancing the discriminative capability across different road scenes and distress types. A multi-composition image augmentation strategy is designed to construct paired training samples, where two semantically consistent but visually distinct views of the same image are generated through joint transformations including scale-adaptive random cropping, multi-level color perturbation,

Gaussian blurring, and horizontal flipping, effectively expanding the feature space of road scenes. The augmented views are fed into the shared encoder for feature extraction, and subsequently passed through a three-layer stacked multi-layer perceptron projection network for dimensionality reduction and semantic mapping, with the layer dimensions sequentially set to 768, 2048, 2048, and 128, and batch normalization and activation functions configured at each layer for feature regularization. The branch is optimized using the normalized temperature-scaled cross-entropy loss, by which the representational distribution is optimized through sample feature similarity measurement. The loss function is formulated as:

$$L_{cont}=-\log\frac{exp(sim(z_i,z_j)/\tau)}{\sum_{k=1}^{2B}1_{[k\neq i]}\,exp(sim(z_i,z_k)/\tau)} \tag{6}$$

where, z_i and z_j denote the dual-view features generated from the same original image, *sim* represents the cosine similarity function, τ denotes the temperature scaling parameter, set to 0.07 in this work for modulating the smoothness of the feature distribution, B denotes the training batch size, and the indicator function is employed to exclude the similarity computation of the sample with itself, thereby ensuring the validity of the instance discrimination task.

The masked reconstruction branch is dedicated to fitting local pixel and texture details, compensating for the deficiency of contrastive learning in capturing fine-grained distress features. A lightweight Transformer decoder is designed for this branch, comprising eight decoding layers, each equipped with eight attention heads and a hidden dimension of 512, by which reconstruction accuracy is ensured while computational overhead is reduced. The differentiated masking matrix is applied to the original input image, with the pixel information of unmasked regions preserved to construct corrupted images. The corrupted images are fed into the shared encoder for hidden feature extraction, and the complete pavement image information is subsequently reconstructed through the decoder. During training, the pixel reconstruction error is computed exclusively for the masked regions, concentrating feature learning on high-value distress areas. The adopted mean squared error reconstruction loss is defined as:

$$L_{rec}=\frac{1}{|M|}\sum_{i:M_i=1}\|X_i-\widehat{X}_i\|_2^2 \tag{7}$$

where, M denotes the image masking matrix, $|M|$ represents the total number of masked patches, and X_i and $\widehat{X}_i$ correspond to the pixel features of the original image and the reconstructed image, respectively. By constraining the pixel reconstruction fidelity in the masked regions, the model is effectively driven to learn the local texture and structural features of cracks, potholes, and other distress types with high precision.

To balance the learning weights of the dual-branch tasks and achieve synergistic optimization of global semantic and local detail features, an iteration-adaptive dynamic loss weighting mechanism is introduced for joint training. The two task branches share the encoder parameters, while only the independent projection head and decoder parameters are retained to adapt to the characteristics of their respective tasks, thereby avoiding feature redundancy while ensuring differentiated learning capability across the two tasks. The overall loss function is formulated as a dynamically weighted combination of the dual-branch losses, expressed as:

$$L_{total}=L_{cont}+\lambda(t)\cdot L_{rec} \tag{8}$$

where, $\lambda(t)$ denotes the iteration-dependent balancing coefficient, which is dynamically updated according to a linear decay rule. The update formula is given by:

$$\lambda(t)=\lambda_{max}-(\lambda_{max}-\lambda_{min})\cdot t/T \tag{9}$$

where, the upper weight limit λ_{max} is set to 2.0, the lower weight limit λ_{min} is set to 0.5, t denotes the current training iteration epoch, and T is the total number of pretraining epochs, fixed at 200. Through this mechanism, the reconstruction task is prioritized in the early training stage, enabling rapid learning of basic pavement textures and distress structural information. As training progresses, the contribution weight of the contrastive learning task is gradually increased, thereby reinforcing the model's global semantic discriminative capability. Ultimately, complementary fusion and synergistic optimization of the two types of self-supervised representations are achieved.

2.3 Multi-scale feature adaptive fusion module

Road distresses exhibit extremely non-uniform scale distribution characteristics. Fine cracks contain only a small number of pixel-level texture features, whereas potholes and contiguous alligator cracking cover extensive pavement regions, with scale spans exceeding two orders of magnitude. A single deep feature output can only emphasize either global semantic modeling or local detail capture, failing to accommodate the representational requirements of distresses at different scales, which results in the loss of small-scale distress features and insufficient semantic representation of large-scale distresses. To overcome the limitations of single-scale feature modeling, a multi-scale feature adaptive fusion module is designed, by which hierarchical capture of multi-level feature information and dynamic aggregation are achieved through the synergistic mechanism of scale-adaptive embedding and attention-guided fusion, constructing a unified representation system adapted to full-scale road distresses. The structure of the multi-scale feature adaptive fusion module is illustrated in Figure 3.

Leveraging the hierarchical feature representation characteristics of the visual Transformer encoder, a scale-adaptive embedding submodule is constructed to perform the selection and regularization of multi-scale distress features. The Vision Transformer encoder comprises twelve Transformer encoding layers, with features from different network layers exhibiting distinct representational emphases. Shallow-layer features preserve fine-grained high-frequency spatial information and are highly sensitive to small-scale distresses such as fine cracks. Mid-layer features integrate spatial structure and basic semantic information, adapting to medium-scale pavement damage. Deep-layer features focus on global contextual semantic modeling, effectively characterizing the overall distribution patterns of large-area distresses. The feature outputs from the third, seventh, and eleventh layers are selected as the sources of multi-scale representation, corresponding to detail, structure, and semantic core features, respectively. Through independent linear transformations, the three types of features are uniformly projected into a 768-dimensional hidden space, eliminating dimensional discrepancies across hierarchical features and providing regularized and unified feature inputs

for subsequent cross-scale fusion.

To overcome the limitations of fixed-weight concatenation in fusion, a lightweight attention-guided fusion mechanism is adopted, by which the fusion weights of each scale feature are adaptively assigned according to the distress distribution state of the input image. First, the regularized multi-scale features are transformed through scale-specific projection matrices to achieve feature space alignment, with the computation formulated as:

$$\widetilde{F}_s = W_s F_s \quad s \in \{1,2,3\} \tag{10}$$

where, F_s denotes the original multi-level features extracted by the network, W_s represents the learnable projection matrix specific to each scale, and $\widetilde{F}_s$ denotes the standardized features after dimensional and spatial alignment. Subsequently, the semantic contribution of each hierarchical feature is computed through a global learnable query vector and scale-specific key matrices, and the adaptive fusion weights are obtained after normalization:

$$a_s = \frac{exp(q^\top K_s \widetilde{F}_s)}{\sum_{s'=1}^{3} exp(q^\top K_{s'} \widetilde{F}_{s'})} \tag{11}$$

where, q denotes the global learnable query parameter, K_s represents the scale-specific key projection matrix, and the weight coefficients satisfy the normalization constraint. Combined with the attention weights and the value-projected features, adaptive fusion of multi-scale information is accomplished. The final fused feature is computed as:

$$F_{fused} = \sum_{s=1}^{3} a_s \cdot V_s \widetilde{F}_s \tag{12}$$

where, V_s denotes the scale-specific value projection matrix. This mechanism dynamically adjusts the contribution ratio of features at each level—enhancing the weight of shallow detail features for fine-crack images while emphasizing deep semantic features for large-area distress images—thereby achieving scale-adaptive intelligent fusion.

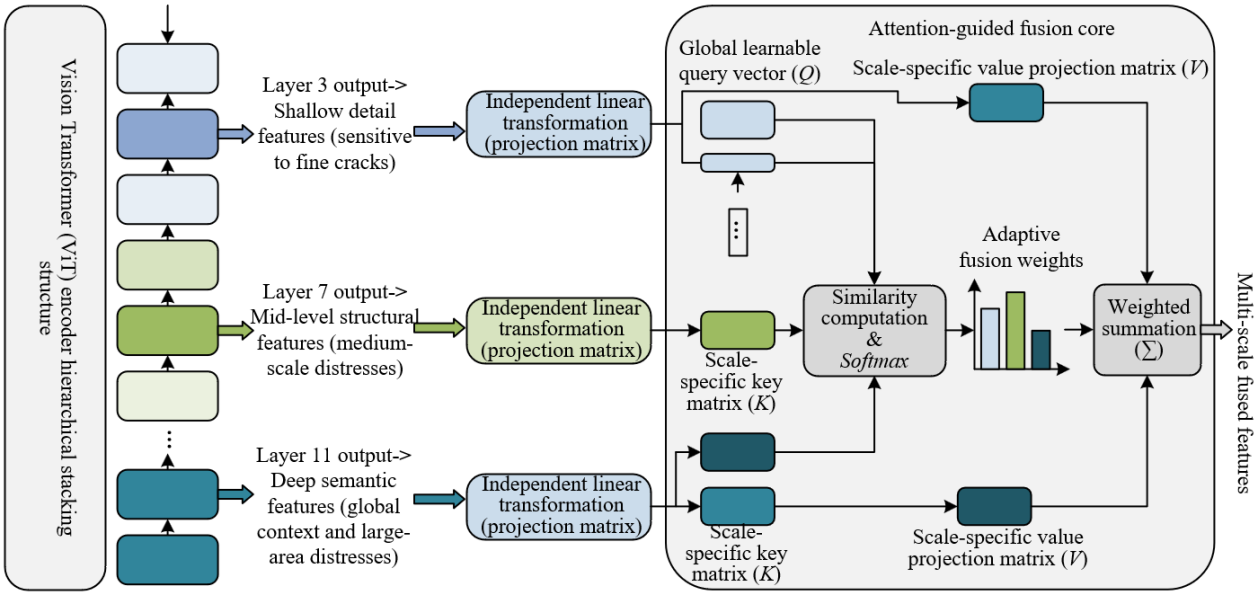

Figure 3. Structure of the multi-scale feature adaptive fusion module

To ensure semantic synergy among multi-scale features while preserving their differentiated representations, a cross-scale consistency constraint is introduced, by which semantic drift and representational redundancy across multi-level features are suppressed. This constraint guides the alignment of hierarchical features in the semantic space by measuring their spatial distances. The loss function is defined as:

$$L_{scale} = \sum_{s \neq s'} \| sg(F_s) - F_{s'} \|_2^2 \tag{13}$$

where, $sg(\,)$ denotes the stop-gradient operation, which fixes the parameter update path of the reference features and effectively prevents the mutual assimilation of features at different scales during iterative optimization, thereby preserving the unique scale-specific advantages of each level. This loss is incorporated into optimization exclusively during the self-supervised pretraining phase, with a weight coefficient set to 0.1, reinforcing feature synergy without compromising the diversity of multi-scale features, and further enhancing the completeness and robustness of the final fused representations.

2.4 Cross-domain self-supervised visual prompt adaptation

Through the multi-module collaborative self-supervised pretraining described in the preceding sections, the model is capable of acquiring refined, multi-scale distress representations from source-domain road data. However, significant domain distribution discrepancies exist among road images in real-world engineering scenarios. Variations in pavement materials, environmental illumination, and camera viewpoints across different geographic regions induce feature space shifts, leading to representational misalignment and substantial performance degradation when the pretrained model is directly transferred to unknown target domains. Conventional full fine-tuning approaches require updating all network parameters, which is not only computationally expensive but also prone to corrupting the generalizable distress features acquired during source-domain pretraining and triggering overfitting. To address these issues, a lightweight cross-domain visual prompt adaptation mechanism is introduced, in which all parameters of the

pretrained backbone network are completely frozen, and only learnable visual prompt vectors are employed to modulate the feature extraction logic of the model. Through this mechanism, target-domain scene adaptation is achieved with minimal parameter overhead, effectively preserving the high-quality source-domain representations while compensating for the performance deficits induced by domain shift. A schematic diagram of the cross-domain self-supervised visual prompt adaptation mechanism is presented in Figure 4. To generate domain-specific prompt information tailored to the distress feature distribution of the target domain, a lightweight two-layer perceptron is constructed as the prompt generation network, which dynamically outputs prompt vectors conditioned on target-domain image features and distress saliency priors. The network input integrates mid-layer features from the encoder and target-domain saliency score features. The activation features from the sixth layer of the Vision Transformer network are selected as the semantic input, as this layer captures both shallow spatial details and deep semantic information, accurately reflecting the underlying visual distribution of target-domain images. After dimensional concatenation of the input features and saliency features, prompt vector decoding is accomplished through hierarchical transformations and nonlinear activations. The overall mapping process is expressed as:

$$P=p_{\omega}\left(f_{\theta}(X_{target}),S_{target}\right) \tag{14}$$

where, p_{ω} denotes the prompt generation network, X_{target} represents the target-domain input image, S_{target} denotes the corresponding distress saliency score map, and P is the final output visual prompt vector. The dimensional transformation structure of the network is sequentially configured as the input dimension, a hidden layer dimension of 256, and the prompt vector dimension, with a Gaussian error linear unit nonlinear activation function incorporated in the intermediate layer to enhance feature mapping capability. The number of output prompt tokens is fixed at ten, ensuring both sufficient prompt informativeness and lightweight characteristics.

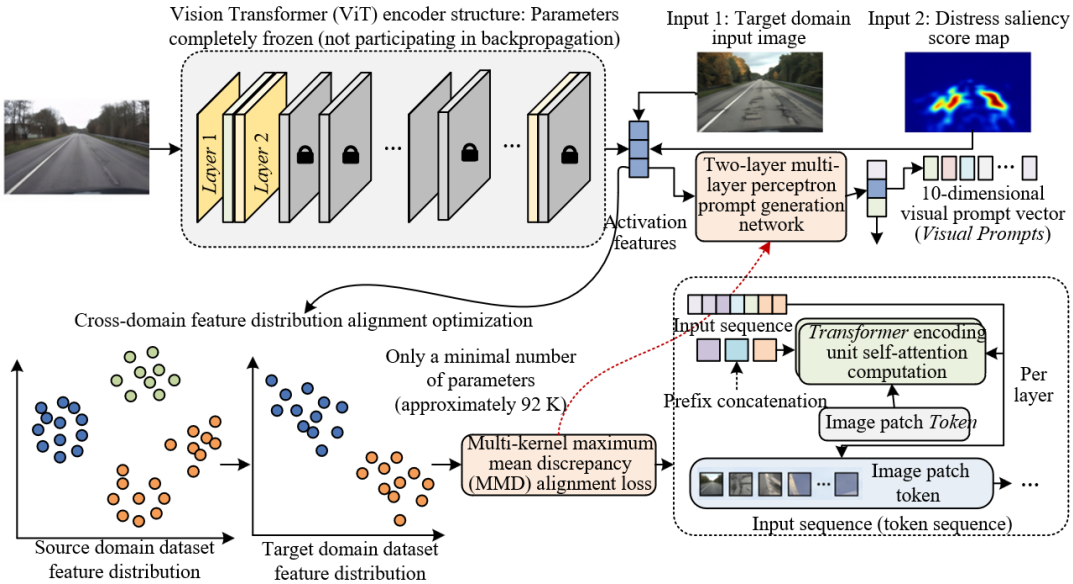

Figure 4. Schematic diagram of the cross-domain self-supervised visual prompt adaptation mechanism

To fully leverage the domain-regulating capability of visual prompts, a prefix concatenation strategy is adopted to embed the prompt vectors into each encoding layer of the Transformer network, enabling global-level feature rectification. At each layer, independent prompt tokens are prepended to the input sequence before the image patch tokens, and these tokens participate jointly with the patch features in the self-attention computation. Through this mechanism, the target-domain distress prior information carried by the prompt vectors is able to guide the feature modeling process at each network layer, progressively correcting the representational deviations induced by domain shifts. Independent parameters are learned for the prompt tokens across all layers, with the total newly added parameters amounting to only 92,160—less than 0.1% of the total parameter count of the ViT-B/16 backbone network. This embedding approach requires no modification to the backbone network architecture or pretrained weights, thereby maximally preserving the distress representational capacity acquired during self-supervised pretraining, while achieving precise adaptation to target-domain feature patterns through lightweight parameter learning.

To constrain the learning direction of the prompt vectors and reduce the feature distribution discrepancy between the source and target domains, a domain-aware prompt alignment loss function is constructed, by which cross-domain feature alignment is achieved in an unsupervised manner based on the maximum mean discrepancy criterion. A multi-kernel Gaussian radial basis function combination is adopted to construct the metric function, integrating multiple kernel bandwidths to enhance the accuracy of distribution discrepancy measurement. The kernel bandwidths are set to 0.1, 1.0, and 10.0, respectively, with equal weights assigned to each kernel function. The cross-domain alignment loss and the maximum mean discrepancy computation are defined as:

$$L_{align}=MMD\left(E_{X_s\sim D_s}[f_{\theta}(X_s)],E_{X_t\sim D_t}[f_{\theta}(X_t \mid P)]\right) \tag{15}$$

$$MMD(P,Q)=\left\|E_{x\sim P}[\phi(x)]-E_{y\sim Q}[\phi(y)]\right\|_H^2 \tag{16}$$

where, D_s and D_t denote the source-domain dataset and the target-domain dataset, respectively, $\phi(\ )$ represents the high-dimensional feature mapping function, and H denotes the

reproducing kernel Hilbert space. This loss function backpropagates only to update the parameters of the prompt generation network, while the backbone network weights remain fixed throughout. Leveraging unlabeled target-domain data, feature distribution adaptive alignment is accomplished, thereby achieving efficient and low-cost cross-domain scene transfer while preserving the model's generalization capability.

2.5 Overall training procedure and optimizer configuration

A staged progressive training paradigm is adopted for the proposed model, by which optimization is performed sequentially according to the logic of representation learning, cross-domain adaptation, and downstream task fitting. The training objectives, parameter update strategies, and hyperparameter configurations of each stage are independent of one another, ensuring ordered convergence of the model's feature learning. The first stage consists of source-domain self-supervised pretraining, where the contrastive loss, masked reconstruction loss, and multi-scale consistency loss are jointly optimized on large-scale unlabeled road images, with all parameters of the shared encoder, dual-branch task heads, and multi-scale fusion module updated iteratively in synchronization. The AdamW optimizer is employed for parameter updates in this stage, with an initial learning rate of 1×10^{-4} and a weight decay coefficient of 0.05, which effectively constrains the magnitude of network parameters and suppresses overfitting. Training is conducted in a distributed manner across eight NVIDIA A100 graphics processing units, with a global batch size of 256 and a per-graphics processing unit batch size of 32. A cosine annealing learning rate schedule is adopted, with 10 warm-up iterations during which the learning rate is linearly increased from 1×10^{-6} to the initial learning rate, followed by a smooth decay as training proceeds. The total number of pretraining iterations is set to 200 epochs, ensuring that all modules adequately learn the generalizable representational features of road distresses.

The second stage consists of unsupervised cross-domain prompt adaptation training, with the core objective of achieving domain-adaptive transfer of the model to unseen target scenes. In this stage, all parameters of the pretrained backbone encoder are completely frozen, fully preserving the distress representational capacity acquired from the source domain, while only the lightweight prompt generation network is optimized. The domain-aware prompt alignment loss is employed to enforce feature distribution consistency between the target and source domains. The AdamW optimizer is retained for this stage, with the learning rate set to 1×10^{-3} and the weight decay coefficient adjusted to 0.01 to accommodate the lightweight network training characteristics, and the batch size is set to 128. The learning rate decay strategy is disabled, and a fixed learning rate is maintained throughout 50 training iterations. Through adaptive updates of the minimal set of learnable prompt parameters, the feature deviations induced by domain shift are precisely corrected, thereby efficiently enhancing the cross-scene generalization capability of the model without compromising the performance of the original pretrained model.

The third stage is dedicated to fine-tuning training and inference deployment for downstream tasks of road distress classification and segmentation, tailored to the requirements of real-world roadway network inspection. For the distress classification task, a linear classification layer is appended to the multi-scale fused features, and optimization is performed using the cross-entropy loss for category discrimination. For the pixel-level distress segmentation task, the fused features are upsampled to the original image resolution via transposed convolution, combined with one-dimensional convolution to achieve pixel-wise binary classification of distress and background. A layer-wise parameter update strategy is adopted during fine-tuning, where the parameters of the last four encoder layers are unfrozen to accommodate downstream task feature adaptation. The initial learning rate for the backbone network is set to 1×10^{-5}, while the initial learning rate for the classification and segmentation task heads is set to 1×10^{-3} . This differentiated learning rate configuration effectively balances the preservation of generalizable features and the fitting of task-specific features. The fine-tuning batch size is set to 64, with a total of 50 training epochs. After fine-tuning, the model can be directly deployed for road distress recognition and segmentation inference across various scenes.

3. EXPERIMENTS

3.1 Experimental setup

Multi-source public road distress datasets were employed for model training and performance evaluation, ensuring the objectivity and generalizability of the experimental results through cross-regional, multi-scenario, and multi-distress-type data samples. The primary experimental dataset is RDD2022, which comprised 47,420 road images collected from six countries—China, Japan, India, the Czech Republic, Norway, and the United States—comprehensively covering four typical pavement distress types: longitudinal cracks, transverse cracks, alligator cracking, and potholes. The dataset exhibits a wide range of image resolutions, faithfully reflecting the image quality fluctuations encountered in real-world road inspection processes. The dataset was uniformly partitioned into training, validation, and test sets at a ratio of 7:1:2. To further evaluate the cross-domain adaptation capability and complex-scene adaptability of the model, the datasets GRDDC and IRRDD were introduced for supplementary testing. The IRRDD encompassed five distress categories—cracks, potholes, rutting, pavement protrusions, and faded markings—effectively enriching the diversity and complexity of the test scenarios.

All experiments were developed and deployed based on the PyTorch 2.0 deep learning framework, with distributed training conducted on eight NVIDIA A100 40 GB high-performance graphics processing units. The model backbone adopted the ViT-B/16 architecture, with weights initialized from the ImageNet-21K pretrained parameters. All input images were uniformly resized to 224×224 resolution, with a patch size of 16. To mitigate experimental errors arising from random initialization and data sampling, all experimental groups were independently repeated three times for training and testing, and the final experimental results were reported as the mean values across the three runs, ensuring the reliability of the conclusions.

3.2 Experimental results and analysis

To systematically validate the module effectiveness, architectural rationality, data efficiency, and generalization superiority of the proposed fully self-supervised road distress

recognition framework, multiple sets of experiments were conducted in this section, including ablation studies, architectural comparison experiments, cross-domain testing experiments, labeling efficiency experiments, and state-of-the-art algorithm comparison experiments. Comprehensive performance validation and mechanistic analysis were performed from multiple dimensions, including module functionality, network architecture, data adaptability, and scene generalization.

3.2.1 Ablation study on the distress-aware differentiated masking strategy

To validate the optimization effect of the differentiated masking generation strategy on self-supervised pretraining representation learning, three sets of controlled experiments were established: a uniform random masking baseline, a static differentiated masking configuration, and a dynamically scheduled differentiated masking configuration. The remaining network structures and training parameters were fixed across all groups, and the classification and segmentation performance under different masking mechanisms were compared. The experimental results are presented in Table 1.

The experimental results in Table 1 clearly demonstrate the effectiveness of the differentiated masking generation strategy. The uniform random masking baseline model achieves only 76.8% in classification accuracy, 75.9% in weighted F1-score, and 58.2% in mean intersection over union, indicating that the generic random masking mechanism is insufficient to guide the model toward focusing on core distress regions. With the introduction of the static differentiated masking strategy, all three metrics are improved to 78.1%, 77.3%, and 60.5%, respectively, confirming that the distress-saliency-aware differentiated masking allocation effectively reinforces the learning weight of distress features during the pretraining phase. Upon incorporating the dynamic scheduling mechanism, the model performance is further enhanced to 79.3%, 78.6%, and 62.1%, representing improvements of 1.2 percentage points in classification accuracy and 1.6 percentage points in mean intersection over union over the static configuration. These results validate that the progressive masking probability scheduling strategy—by maintaining a uniform masking distribution in the early training stage to establish general visual cognition, and gradually intensifying differentiated focusing in the later stage to deepen distress feature learning—effectively mitigates the suboptimal local convergence issue induced by extreme differentiated masking, thereby ensuring stable convergence and improved representational quality throughout the pretraining process.

3.2.2 Ablation study on contrastive-reconstruction dual-branch collaborative pretraining

To validate the complementary advantages of the dual-branch collaborative self-supervised pretraining architecture, three pretraining schemes—contrastive learning only, masked reconstruction learning only, and dual-branch collaborative learning—were respectively constructed, and the representational learning efficacy of different learning paradigms was compared. The experimental results are presented in Table 2.

Table 1. Ablation study on the distress-aware differentiated masking strategy

Masking Strategy	Masking Rate Configuration	Accuracy (%)	Weighted F1 (%)	Mean Intersection Over Union (%)
Uniform random masking (baseline)	$r = 0.75$	76.8	75.9	58.2
High masking rate only	$r_{high} = 0.80$, $r_{low} = 0.50$ (without dynamic scheduling)	78.1	77.3	60.5
Differentiated masking (proposed)	$r_{high} = 0.80$, $r_{low} = 0.50$ (with dynamic scheduling)	79.3	78.6	62.1

Table 2. Dual-branch collaborative pretraining vs. single-branch pretraining

Pretraining Configuration	Contrastive Loss *Lcont*	Reconstruction Loss *Lrec*	Accuracy (%)	Weighted F1 (%)	Mean Intersection Over Union (%)
Contrastive learning only	√	×	77.5	76.7	59.8
Reconstruction only (masked autoencoder)	×	√	76.2	75.1	58.9
Dual-branch collaborative (proposed)	√	√	79.3	78.6	62.1

The experimental results in Table 2 reveal the inherent representational deficiencies of single self-supervised learning paradigms in road distress recognition tasks. The contrastive-learning-only branch model achieves a classification accuracy of 77.5%, a weighted F1-score of 76.7%, and a mean intersection over union of 59.8%, indicating that while contrastive learning offers certain advantages in global semantic discrimination, its capacity for modeling local textural details of distresses is notably insufficient. The masked-reconstruction-only branch model underperforms the contrastive branch in classification accuracy (76.2%), weighted F1-score (75.1%), and mean intersection over union (58.9%), suggesting that although the masked reconstruction paradigm excels at pixel-level texture fitting, its excessive focus on low-level visual features leads to the loss of high-level semantic discriminative information. The proposed dual-branch collaborative pretraining framework, through shared encoder parameter integration of the representational advantages of both self-supervised tasks, and by virtue of the dynamic loss balancing mechanism that enables progressive synergistic optimization—with reconstruction dominating in the early stage to learn foundational textures and contrastive learning dominating in the later stage to reinforce semantic discrimination—achieves 79.3%, 78.6%, and 62.1% on the three metrics, respectively. These values represent improvements of 1.8, 1.9, and 2.3 percentage points over the optimal single-branch scheme, conclusively demonstrating the effectiveness of dual-branch collaborative learning in compensating for the representational deficiencies of single paradigms.

3.2.3 Performance evaluation of the multi-scale feature adaptive fusion module

To verify the modeling capability of the attention-guided adaptive fusion mechanism for multi-scale distress features, three strategies—single-scale deep features, simple

concatenation of multi-scale features, and adaptive attention-based fusion—were compared, with segmentation accuracies for fine cracks, large-area distresses, and global scenes separately evaluated. The experimental results are presented in Table 3.

The experimental data in Table 3 fully demonstrate the significant impact of the scale variability of road distresses on feature modeling. The single-scale deep feature strategy (Layer 11) achieves a mean intersection over union of 68.7% on large-area distress segmentation, but attains only 54.3% on fine cracks and other small-scale distresses, with a gap of 14.4 percentage points between the two, reflecting the severe deficiency of single-depth features in preserving spatial details. The simple concatenation strategy improves fine crack segmentation accuracy to 61.8%, yet causes the accuracy on large-area distresses to decline to 65.2%, indicating that the high-frequency noise introduced by equal-weight fusion interferes with the stable expression of deep semantic features. The proposed attention-guided adaptive fusion mechanism, through dynamic weight allocation that enables on-demand configuration of feature resources, further elevates fine crack segmentation accuracy to 63.5%, restores large-area distress accuracy to 68.9%, and achieves an overall mean intersection over union of 65.8%. This represents an overall improvement of 2.7 percentage points over the simple concatenation strategy, conclusively demonstrating the core advantage of the adaptive fusion mechanism in simultaneously accommodating detection accuracy across different scales of distresses.

3.2.4 Evaluation of cross-domain generalization capability

To evaluate the generalization performance of the model in complex cross-scene roadway networks, a cross-domain experimental scheme was constructed based on the RDD2022 multi-country dataset, with training conducted on a single source domain and testing performed across multiple target domains. The cross-domain detection accuracy of various algorithms was compared, and the experimental results are presented in Table 4.

Table 3. Performance evaluation of the multi-scale adaptive fusion module

Feature Fusion Strategy	Fine Crack Mean Intersection Over Union (%)	Large-Area Distress Mean Intersection Over Union (%)	Overall Mean Intersection Over Union (%)
Single-scale (Layer 11, deep)	54.3	68.7	60.5
Simple concatenation of multi-scale features	61.8	65.2	63.1
Attention-guided adaptive fusion (proposed)	63.5	68.9	65.8

Table 4. Cross-domain generalization capability evaluation (RDD2022 six-country subsets: training on one country → testing on the remaining five, mAP@0.5%)

Method	Japan→Test	India→Test	Czech→Test	Norway→Test	USA→Test	Average
Supervised (source domain training, direct testing on target)	42.3	38.7	35.2	40.1	36.8	38.6
Self-supervised pretraining (without prompt adaptation)	54.6	49.8	46.3	52.4	47.9	50.2
Supervised + domain adaptation (visual prompt adaptation)	58.2	53.1	49.7	56	51.3	53.7
Proposed method (self-supervised pretraining + visual prompt adaptation)	64.8	60.2	55.7	62.5	58.1	60.3

The cross-domain experimental results in Table 4 systematically reveal the performance discrepancies of different methods in addressing domain shift issues. The conventional supervised model achieves an average accuracy of only 38.6% across the six-country cross-domain tests, with severe performance degradation relative to source-domain training—particularly dropping to 35.2% on the Czech→Test subset—fully reflecting the high dependence of supervised learning on the distribution of training data. The standard self-supervised pretraining model, while requiring no annotated data and exhibiting stronger generalizable representational capacity, achieves an average accuracy of 50.2% across the six countries, yet still lacks target-domain-specific feature alignment capability. The supervised approach combined with visual prompt adaptation raises the average accuracy to 53.7%, but the improvement remains limited due to the generalization bottleneck of supervised pretraining representations. The proposed method, leveraging distress-specific self-supervised pretraining to obtain highly generalizable foundational representations, and subsequently correcting target-domain feature distributions through the lightweight visual prompt mechanism, achieves detection accuracies of 64.8%, 60.2%, 55.7%, 62.5%, and 58.1% on the five target-domain subsets of Japan, India, the Czech Republic, Norway, and the United States, respectively, with a six-country average accuracy of 60.3%. This represents an absolute improvement of 6.6 percentage points over the optimal comparison scheme, conclusively validating the significant advantage of the "generalizable self-supervised representation + lightweight domain adaptation" paradigm in cross-domain road distress recognition.

3.2.5 Labeling efficiency experiment analysis

To verify the capability of the proposed self-supervised framework in reducing manual annotation costs, fine-tuning experiments with varying proportions of labeled data were conducted, and the learning efficiency of different models under scarce annotation scenarios was compared. The experimental results are presented in Figure 5.

The labeling efficiency experimental results in Figure 5 fully demonstrate the significant advantage of the proposed self-supervised framework in reducing annotation costs. Under the extremely scarce scenario with only 1% labeled data, the ResNet-50 and ViT-B/16 models trained from scratch achieve accuracies of only 32.1% and 29.8%, respectively, rendering them practically unusable. The generic ImageNet-pretrained model, benefiting from large-scale general visual prior knowledge, attains 58.4%, yet its performance remains

limited due to scene discrepancies. The proposed method achieves a classification accuracy of 67.5% under the 1% labeling condition, representing an improvement of 9.1 percentage points over the generic pretraining scheme. More critically, with only 10% labeled data, the proposed method reaches an accuracy of 76.4%, already surpassing the full-data performance of traditional models trained from scratch—76.4% for ResNet-50 and 75.8% for ViT-B/16—while approaching the full fine-tuning performance of the generic pretrained model (78.2%). This indicates that the proposed method can compress the manual annotation cost for road distress recognition tasks by approximately 90%. Under the full annotation condition, the proposed method further achieves 79.3%, validating the favorable compatibility between the road-scene-oriented self-supervised pretraining representations and the downstream supervised fine-tuning pipeline.

3.2.6 Comprehensive comparison with state-of-the-art self-supervised algorithms

To comprehensively validate the overall superiority of the proposed algorithm over existing road distress self-supervised methods, recent mainstream specialized algorithms were selected for horizontal comparison on both classification and segmentation tasks. The experimental results are presented in Table 5.

The comprehensive comparison results in Table 5 further establish the leading position of the proposed method in the field of self-supervised road distress recognition. The patch-based self-supervised learning method based on the Simple Framework for Contrastive Learning of Visual Representations (SimCLR) achieves a classification accuracy of 78.0%, yet attains a mean intersection over union of only 56.8%, reflecting the inherent limitation of the single contrastive learning paradigm in pixel-level segmentation tasks. The spatial contexts-informed self-supervised learning method (based on the masked autoencoder), despite incorporating spatial context modeling, achieves only 75.4% classification accuracy and 57.5% mean intersection over union, both inferior to the proposed method. The Crack-Segmenter method attains a mean intersection over union of 59.1% on the segmentation task, but its classification accuracy reaches only 76.8%. The Contrastive Learning for Adaptive Generalized Pavement Surface Distress Detection (CL-PSDD) method achieves 79.0% classification accuracy, approaching the 79.3% of the proposed method, yet it is only applicable to image-level classification tasks and lacks pixel-level segmentation capability. The PROBE method, while introducing a visual prompt mechanism for cross-domain adaptation and achieving a mean intersection over union of 63.2%, still falls 2.6 percentage points behind the proposed method. The proposed method achieves 79.3% classification accuracy, 78.6% weighted F1-score, and 65.8% mean intersection over union, outperforming all existing methods across all three core metrics, and stands as the only unified framework that simultaneously attains state-of-the-art performance on both classification and segmentation tasks, fully demonstrating the comprehensive advantages of dual-branch collaborative pretraining and multi-scale feature fusion.

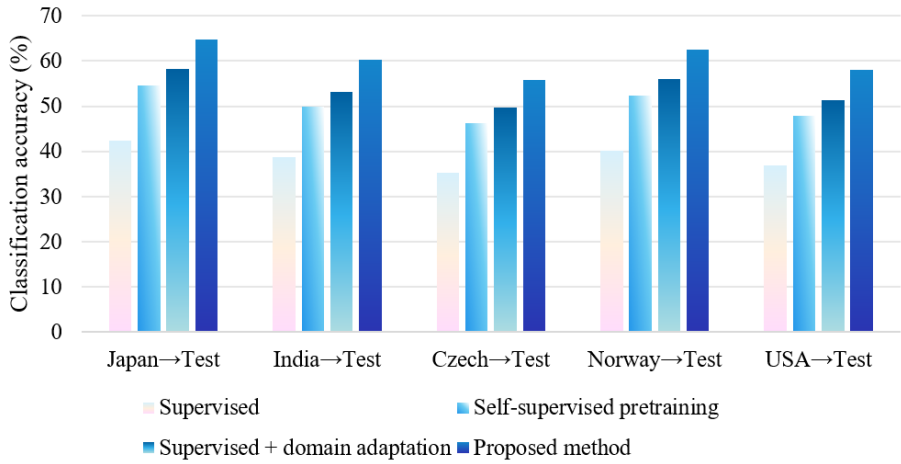

Figure 5. Labeling efficiency experiment (classification accuracy under varying proportions of labeled data)

Table 5. Comprehensive comparison with the latest self-supervised road distress detection methods

Method	Pretraining Paradigm	Backbone	Accuracy (%)	Weighted F1 (%)	Mean Intersection Over Union (%)
Patch-based self-supervised learning	Simple Framework for Contrastive Learning of Visual Representations (SimCLR)	MobileNetV2	78	78	56.8
Spatial contexts-informed self-supervised learning	Masked autoencoder	ViT-B/16	75.4	74.2	57.5
Crack-Segmenter	Diffusion	ResNet-50	76.8	76	59.1
Contrastive Learning for Adaptive Generalized Pavement Surface Distress Detection (CL-PSDD)	SimCLR (improved)	ResNet-50	79	—	—
PROBE	Visual prompt	ViT-B/16	—	—	63.2
Proposed method	Contrastive-reconstruction dual-branch	ViT-B/16	79.3	78.6	65.8

3.2.7 Stepwise ablation study of modules

To validate the independent effectiveness and synergistic gain characteristics of each core module, the base masked autoencoder-pretrained single-scale model was adopted as the baseline, and the proposed innovative modules were progressively stacked upon it to quantify the performance contribution and parameter overhead of each component. The experimental results are presented in Figure 6.

The stepwise ablation results in Figure 6 clearly demonstrate, through a progressive performance improvement chain, the independent effectiveness and synergistic gain characteristics of each core module in the proposed framework. The baseline model (masked autoencoder pretraining with single-scale features) achieves a classification accuracy of 74.1%, a weighted F1-score of 73.0%, and a mean intersection over union of 55.8%. After sequentially stacking each innovative module, performance exhibits a steady increasing trend: the distress-aware differentiated masking module brings performance improvements of 2.1, 2.1, and 3.1 percentage points on the three metrics, respectively, with zero parameter increment, proving the high efficiency of prior knowledge embedding at the data level. The contrastive branch collaborative pretraining further contributes gains of 1.9, 2.1, and 2.3 percentage points, refining the model's comprehensive representational capacity for both global semantics and local details. The multi-scale adaptive fusion module yields particularly notable gains on the segmentation task, with the mean intersection over union improving by 4.2 percentage points in a single step, fully validating the critical role of this module in addressing the fundamental challenge of the large-scale span of road distresses. The cross-domain visual prompt adaptation module finalizes the optimization of generalization performance with an extremely low parameter increment of only 0.1 M. The overall framework, at the cost of only 2.4 M additional parameters (accounting for 2.8% of the total), achieves a comprehensive performance leap of 5.2 percentage points in classification accuracy, 5.6 percentage points in weighted F1-score, and a substantial 10.0 percentage points in mean intersection over union, fully demonstrating the favorable balance between high-precision recognition and lightweight deployment, as well as the engineering practical value of the proposed method.

To intuitively verify the distress representation, cross-domain adaptation, and fine-grained segmentation capabilities of the proposed self-supervised learning framework in complex road scenes, real road images containing shoulder textures, lane markings, perspective variations, and continuous longitudinal cracks are selected for visualization experiments. As shown in Figure 7, the distress saliency responses form continuous and concentrated high-response regions along the crack main body and its fine branches, while effectively suppressing irrelevant activations induced by pavement granularity, roadside background, and regular markings, indicating that the multi-directional gradient and local texture priors accurately guide the model to focus on core distress regions. The reconstructed results after differentiated masking largely recover the crack orientation, edge morphology, and neighboring textures, demonstrating that the collaborative mechanism of contrastive learning and masked reconstruction is capable of jointly satisfying global semantic discrimination and local structure modeling. The multi-scale feature responses further reveal that shallow-layer features maintain high sensitivity to fine-grained edges, mid-layer features enhance the spatial connectivity structure of cracks, and deep-layer features form more stable semantic aggregation of distresses, with the fused features achieving a superior balance between detail integrity and background suppression. The comparison before and after cross-domain adaptation indicates that the visual prompt mechanism significantly reduces false responses caused by lane markings and complex textures, while improving the continuity and localization stability of crack regions. The final segmentation results are highly consistent with the ground-truth crack contours, maintaining good alignment with narrow branches, width variations, and irregular boundaries. These results collectively confirm that the proposed method is capable of learning discriminative, scale-robust, and scene-generalizable road distress representations under unlabeled pretraining and lightweight domain adaptation conditions, providing reliable technical support for high-precision automated inspection and low-cost model transfer in complex roadway network environments.

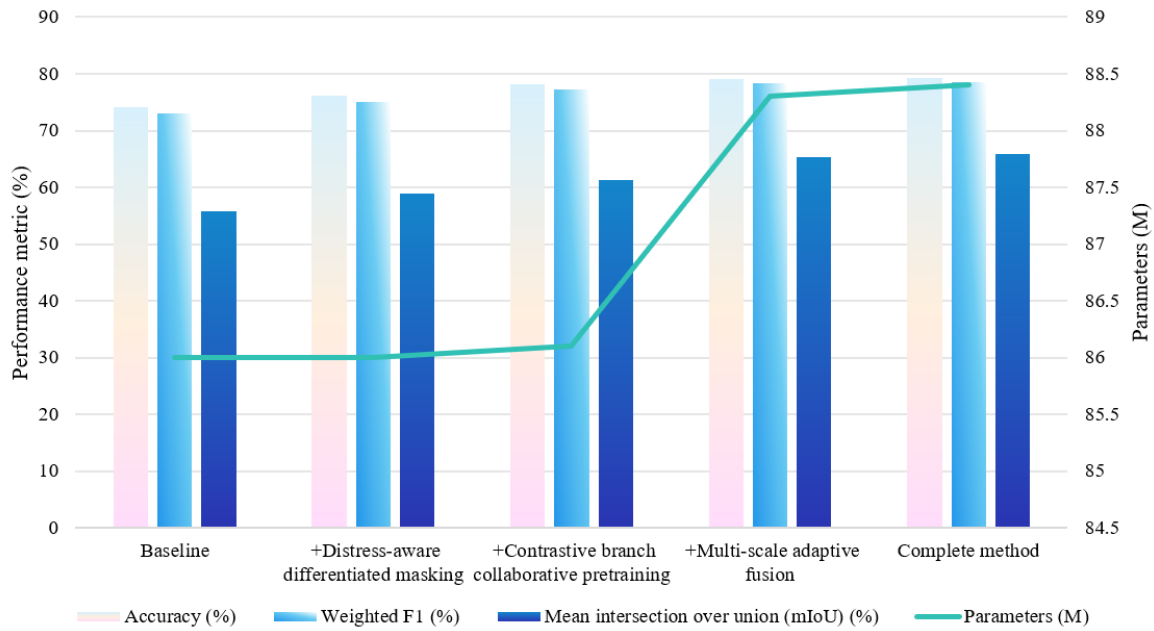

Figure 6. Ablation study of individual modules

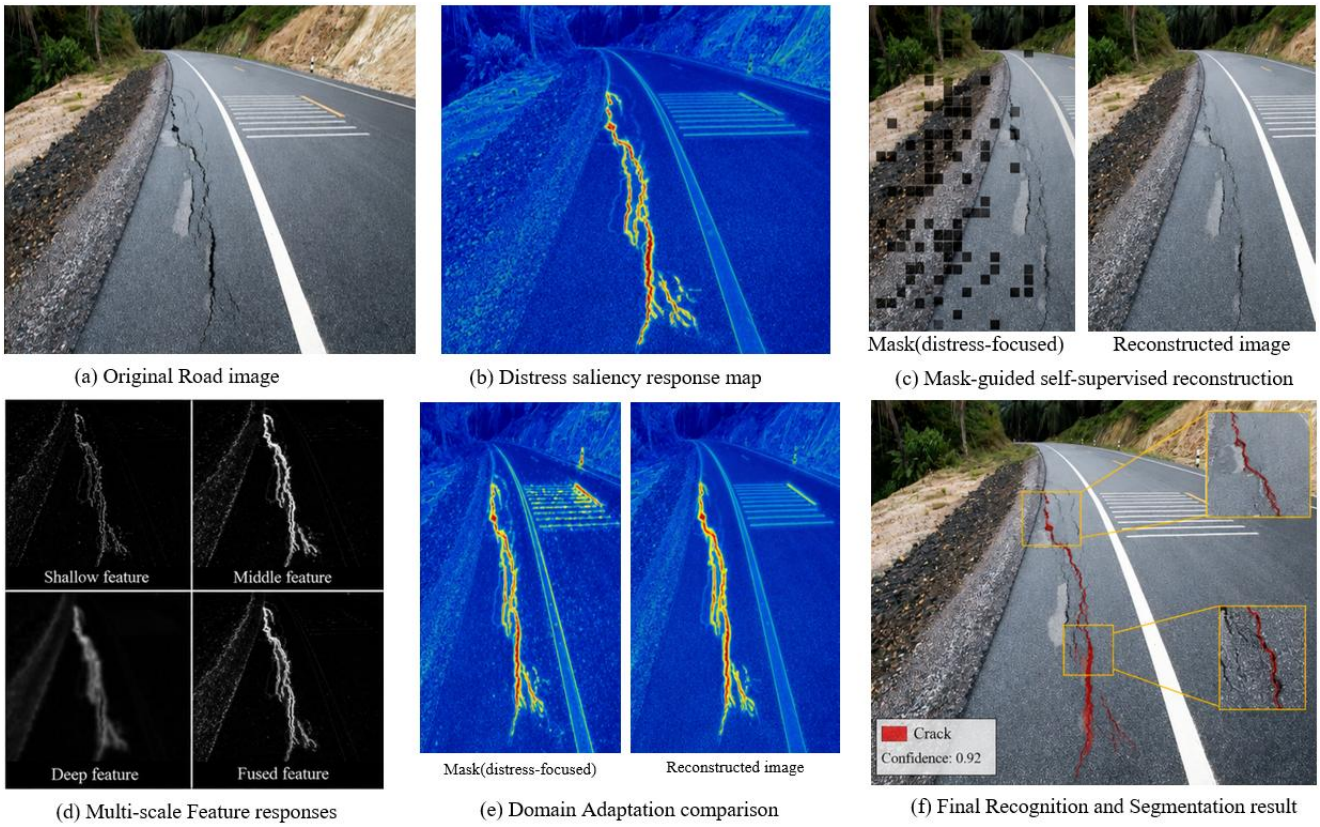

Figure 7. Implementation results of self-supervised distress recognition and image segmentation in complex cross-domain road scenes

4. CONCLUSION

Aiming at the intelligent operation and maintenance scenarios of smart transportation roadway networks, and targeting the critical issues of conventional road distress recognition algorithms—namely, heavy reliance on manually annotated data, poor domain adaptability, insufficient multi-scale modeling capability, and weak cross-domain generalization performance—an end-to-end fully self-supervised road distress image recognition framework was constructed in this work. Through the differentiated masking generation strategy, road-distress-specific visual priors were embedded into the framework, guiding the model to focus on high-value distress regions during the pretraining phase. By virtue of the dual-branch collaborative training mechanism combining contrastive learning and masked reconstruction, global discriminative representations and local detailed features were fused, compensating for the representational deficiencies of single self-supervised learning paradigms. In conjunction with the attention-guided multi-scale feature fusion scheme, the scale-differentiated distribution characteristics of pavement distresses were effectively accommodated, and efficient cross-domain feature alignment and scene transfer were achieved through the lightweight visual prompt adaptation mechanism. Systematic experimental results on multiple public road distress datasets demonstrated that the proposed method outperformed current mainstream supervised and self-supervised learning algorithms in both classification and segmentation tasks, and was capable of maintaining high-precision recognition performance under extremely low annotation conditions. With only 10% of manually annotated data, the detection performance of fully supervised models was matched, substantially reducing the dataset construction cost for intelligent distress detection. Ablation studies further validated the positive contributions and synergistic optimization properties of each core module, confirming the rationality and effectiveness of the overall framework design, and providing reliable technical support for large-scale, cross-scenario intelligent roadway network inspection tasks.

Future research may extend the application boundaries of the existing framework by advancing from static image recognition models to dynamic distress detection scenarios in road inspection video sequences, thereby enabling distress tracking and recognition in the temporal dimension. Meanwhile, multi-sensor fusion techniques may be incorporated, integrating light detection and ranging, infrared sensing, and other multi-source data to construct a multi-modal feature representation system, further enhancing the robustness of distress recognition under adverse operating conditions such as complex illumination and extreme weather, and promoting the deep iteration of intelligent road distress recognition technology from laboratory settings toward engineering deployment.

REFERENCES

[1] Yang, X., Zhang, J., Liu, W., Jing, J., Zheng, H., Xu, W. (2024). Automation in road distress detection, diagnosis and treatment. Journal of Road Engineering, 4(1): 1-26. https://doi.org/10.1016/j.jreng.2024.01.005

[2] El Hakea, A.H., Fakhr, M.W. (2022). Recent computer vision applications for pavement distress and condition

assessment. Automation in Construction, 146: 104664. https://doi.org/10.1016/j.autcon.2022.104664

[3] Arya, D., Maeda, H., Ghosh, S.K., et al. (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

[4] Guan, J., Yang, X., Ding, L., Cheng, X., Lee, V.C., Jin, C. (2021). Automated pixel-level pavement distress detection based on stereo vision and deep learning. Automation in Construction, 129: 103788. https://doi.org/10.1016/j.autcon.2021.103788

[5] Cano-Ortiz, S., Iglesias, L.L., del Árbol, P.M.R., Lastra-González, P., Castro-Fresno, D. (2024). An end-to-end computer vision system based on deep learning for pavement distress detection and quantification. Construction and Building Materials, 416: 135036. https://doi.org/10.1016/j.conbuildmat.2024.135036

[6] Arya, D., Maeda, H., Ghosh, S.K., Toshniwal, D., Sekimoto, Y. (2021). RDD2020: An annotated image dataset for automatic road damage detection using deep learning. Data in Brief, 36: 107133. https://doi.org/10.1016/j.dib.2021.107133

[7] Jing, L., Tian, Y. (2020). Self-supervised visual feature learning with deep neural networks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(11): 4037-4058. https://doi.org/10.1109/tpami.2020.2992393

[8] Ericsson, L., Gouk, H., Loy, C.C., Hospedales, T.M. (2022). Self-supervised representation learning: Introduction, advances, and challenges. IEEE Signal Processing Magazine, 39(3): 42-62. https://doi.org/10.1109/msp.2021.3134634

[9] Liu, X., Zhang, F., Hou, Z., et al. (2023). Self-supervised learning: Generative or contrastive. IEEE Transactions on Knowledge and Data Engineering, 35(1): 857-876. https://doi.org/10.1109/TKDE.2021.3090866

[10] Rani, V., Nabi, S.T., Kumar, M., Mittal, A., Kumar, K. (2023). Self-supervised learning: A succinct review. Archives of Computational Methods in Engineering, 30(4): 2761-2775. https://doi.org/10.1007/s11831-023-09884-2

[11] Roberts, R., Menant, F., Di Mino, G., Baltazart, V. (2022). Optimization and sensitivity analysis of existing deep learning models for pavement surface monitoring using low-quality images. Automation in Construction, 140: 104332. https://doi.org/10.1016/j.autcon.2022.104332

[12] Shim, S., Kim, J., Lee, S., Cho, G. (2022). Road damage detection using super-resolution and semi-supervised learning with generative adversarial network. Automation in Construction, 135: 104139. https://doi.org/10.1016/j.autcon.2022.104139

[13] Fan, Z., Li, C., Chen, Y., et al. (2020). Automatic crack detection on road pavements using encoder-decoder architecture. Materials, 13(13): 2960. https://doi.org/10.3390/ma13132960

[14] Fan, Z., Li, C., Chen, Y., et al. (2020). Ensemble of deep convolutional neural networks for automatic pavement crack detection and measurement. Coatings, 10(2): 152. https://doi.org/10.3390/coatings10020152

[15] Jaiswal, A., Babu, A.R., Zadeh, M.Z., Banerjee, D., Makedon, F. (2020). A survey on contrastive self-supervised learning. Technologies, 9(1): 2. https://doi.org/10.3390/technologies9010002

[16] Chen, Y., Mancini, M., Zhu, X., Akata, Z. (2022). Semi-Supervised and unsupervised deep visual learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(3): 1327-1347. https://doi.org/10.1109/tpami.2022.3201576

[17] Wang, S., Cai, B., Wang, W., et al. (2024). Automated detection of pavement distress based on enhanced YOLOv8 and synthetic data with textured background modeling. Transportation Geotechnics, 48: 101304. https://doi.org/10.1016/j.trgeo.2024.101304

[18] Luo, H., Li, J., Cai, L., Wu, M. (2023). STrans-YOLOX: Fusing swin transformer and YOLOX for automatic pavement crack detection. Applied Sciences, 13(3): 1999. https://doi.org/10.3390/app13031999

[19] Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Loy, C.C. (2023). Domain generalization: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(4): 4396-4415. https://doi.org/10.1109/TPAMI.2022.3195549

[20] Wilson, G., Cook, D.J. (2020). A survey of unsupervised deep domain adaptation. ACM Transactions on Intelligent Systems and Technology, 11(5): 1-46. https://doi.org/10.1145/3400066

[21] Kouw, W.M., Loog, M. (2019). A review of domain adaptation without target labels. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(3): 766-785. https://doi.org/10.1109/tpami.2019.2945942

[22] Fang, Y., Yap, P., Lin, W., Zhu, H., Liu, M. (2024). Source-free unsupervised domain adaptation: A survey. Neural Networks, 174: 106230-106230. https://doi.org/10.1016/j.neunet.2024.106230

[23] 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