FSRDD: An Efficient Few-Shot Detector for Rare City Road Damage Detection

Binyi Su, Hua Zhang, Zhaohui Wu, and Zhong Zhou

***Abstract*— Road damage detection (RDD) is indispensable for safe autonomous driving. Existing RDD models focus on designing feature representations following expert knowledge. However, collecting and labeling all types of samples is time-consuming and leads to insufficient training data. To alleviate the adverse effect of few training samples, a novel few-shot road damage detector (FSRDD) is proposed in this paper to detect rare road damages. The proposed FSRDD includes three stages. First, fully annotated abundant base classes are leveraged to train a base detector, where ghost attention (GA) and proposal feature metric (PFM) modules are developed to eliminate the redundant information and measure the proposal features, respectively. Second, the recognition branch of the detector is fine-tuned using a few samples of all classes. Finally, the test set is inferred with the help of an offline scale-aware prototypical calibration block (SPCB). Extensive experiments show that our FSRDD achieves 10-shot rare road damage detection with 33.4% and 12.9% mAP50 on RDD and CNRDD datasets, respectively, significantly outperforming state-of-the-art methods.**

***Index Terms*— Road damage, deep learning, few-shot detection, fine-tuning.**

I. Introduction

POOR city road conditions would have a significant negative impact on traffic safety, driving efficiency, and vehicle quality. Therefore, road condition estimation has become an important research direction in the field of intelligent transportation. However, it is a challenging task owing to various types of disturbance factors, such as different road conditions and difficulty in collecting data on rare road damages.

Current common methods of road condition estimation depend primarily on the visual inspection of experienced workers, which is inefficient and time-consuming [1]. Meanwhile, many researchers [1], [2], [3], [4] have developed vision-based approaches for road surface inspection using image filters or hand-crafted features. In detail, Subirats *et al.* [1] presented a traditional wavelet-based method to detect pavement cracks. Medina *et al.* [2] adopted the Gabor filter and AdaBoost classifier to detect road crack damage. Moreover, Kapela *et al.* [3] applied the histograms of oriented gradients (HOG) as a feature representation technology to recognize three types of cracks in asphalt surfaced pavements. Hu *et al.* [4] developed a novel local binary pattern method to locate the road cracks under complex background disturbances. However, conventional road damage detection methods have several limitations: first, the feature representation of road damage in these methods mainly relies on manually designed descriptors, which require expert knowledge and complex parameter adjustment. Second, each method solves a specific issue in a real scenario and has poor generalization ability. Third, previous studies [1], [2], [3], [4] have only considered a limited number of road damage types.

Deep learning (DL)-based methods have been used in road damage detection to solve these problems. Maeda *et al.* [5] releases a road damage detection (RDD) dataset in VOC format and employs an improved single shot multi-box detector (SSD) to inspect eight road damage types. To augment the training dataset, Maeda *et al.* [6] applies a generative adversarial network to generate pseudo-road damages. Furthermore,

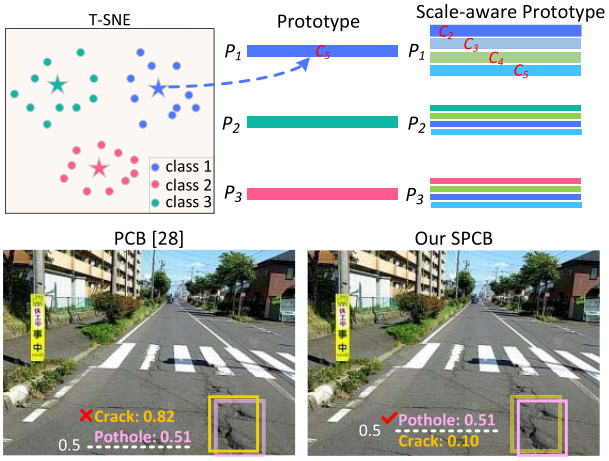

Fig. 1. The prototype in previous methods [20], [26], [27], [28] is single-scale. We propose a scale-aware prototype to calibrate the classification of few-shot object detection (FSOD). C_2, C_3, C_4, and C_5 are multi-scale layers of the feature-extraction network. P_1, P_2, and P_3 represent prototypes.

Manuscript received 15 December 2021; revised 6 May 2022 and 26 July 2022; accepted 13 September 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB2100603, in part by the National Natural Science Foundation of China under Grant 61872024 and Grant 62072454, in part by the Beijing Natural Science Foundation under Grant 4202084, and in part by the Basal Research Fund of Central Public Research Institute of China under Grant 20212701. The Associate Editor for this article was Z. Duric. *(Corresponding author: Zhong Zhou.)*

Binyi Su and Zhong Zhou are with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China (e-mail: subinyi@buaa.edu.cn; zz@buaa.edu.cn).

Hua Zhang is with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China (e-mail: zhanghua@iie.ac.cn).

Zhaohui Wu is with the China Academy of Transportation Sciences, Beijing 100029, China (e-mail: wzh0005@gmail.com).

Digital Object Identifier 10.1109/TITS.2022.3208188

Yang *et al.* [7] proposes a feature pyramid and hierarchical boosting network for road crack detection. Subsequently, Tang *et al.* [8] uses a patch label distillation strategy to automatically detect various pavement distresses. Although existing DL-based models have achieved impressive performance in detecting road damages, they require a large number of annotated samples to train an excellent convolutional neural network (CNN)-based model. However, there exist multiple types of road damages only having few samples, such as crosswalk blur and longitudinal, and lateral crack construction joint part. These damages rarely appear on the road surface and are typically quickly repaired. Thus, traditional data-driven models would fail to detect them with few training samples.

Interestingly, few-shot object detection (FSOD) based on meta learning [9] or transfer learning [10] has attracted significant attentions. FSOD aims to train the model on base classes with abundant examples and novel classes with a small amount of training data. For meta-learning [9], the model first learns meta features from base classes that are generalizable to detect different object classes. It calculates the centroid of each support class, known as class prototype, from the support data, and classifies a query by measuring its similarity to all prototypes. While for transfer learning [10], a learning model employs the balanced samples of all classes to fine-tune the weights of existing detectors trained on abundant samples to develop a few-shot detector. Compared to meta learning-based methods, transfer learning-based methods [10], [27] present the advantages of easy training and a good effect. Therefore, the transfer learning-based method is employed in this paper to detect the infrequent road damage.

The most popular and effective solution for transfer learning-based methods is to introduce an offline prototypical calibration block (PCB) [28], which does not need to be trained. The PCB establishes prototypes and calibrates the class of a query by measuring its similarity with these prototypes. However, existing prototypes [20], [26], [27], [28] are single-scale representations of a support class, resulting in a scarcity of the scale robustness problem in class calibration. To tackle this issue, a scale-aware prototype is proposed to calibrate the classification, which represents a support class with the centroids of different scale layers in a feature-extraction network, as shown in Fig. 1. A scale-aware prototypical calibration block (SPCB) is proposed based on the scale-aware prototype. The SPCB leverages multi-scale hierarchical calibration to effectively narrow such a gap between the multi-scale distributions of objects. In addition, the real-world road damage is hampered by complex background disturbances. To address this problem, an attention-based module is proposed to suppress negative features yielded by the noise background.

Specifically, our few-shot road damage detector (FSRDD) adopts a three-stage transfer-learning scheme to detect the rare categories of the road damage. The road damage detector, such as Faster RCNN [11], is first trained on data-abundant base classes, and then fine-tune only the last layers of the detector with a small balanced training set consisting of both base and novel classes. Finally, a test set is inferred with the assistance of an offline SPCB. To obtain high-quality classification results, the SPCB aggregates multiple scale-aware calibration results to rectify the classification score of the predicted box. Simultaneously, a novel Ghost Attention (GA) module is introduced into the region proposal network (RPN) of the Faster RCNN to reveal the information underlying intrinsic feature maps and eliminate redundant information. Furthermore, a proposal feature metric (PFM) module is designed, which leverages scaled cosine similarity to measure the proposal features and further boosts the few-shot road damage detection performance. The main contributions of this paper are summarized as follows.

1) A few-shot road damage detector (FSRDD) is proposed to detect rare road damages, which presents the advantage of high efficiency.
2) GA and PFM modules are proposed to eliminate redundant information and construct a discriminative representation space, respectively.
3) A SPCB is proposed by aggregating multi-scale hierarchical calibration results to revise the classification of the predicted box.
4) Extensive experiments on RDD dataset and CNRDD dataset demonstrate that our method significantly outperforms state-of-the-art methods.

This paper is organized as follows: Section II presents an overview of the related works. Section III describes the proposed method. Section IV shows the extensive experiments. Finally, Section V concludes the paper.

II. Related Work

A. Conventional Object Detection

Conventional DL-based object detectors can be roughly divided into two categories: one and two-stage detectors. One-stage object detectors such as YOLO [12] and SSD [13] do not depend on the proposal and directly predict the class and bounding box. Gan *et al.* [14] proposes a one-stage detector based on M2det to detect road damage. This method can extract shallow and deep features hidden in an image, improving the performance of small road damage detection. Mao *et al.* [15] introduces YOLOv3 with an improved aspect ratio sensitive loss to boost the performance of road damage detection. Recently, YOLO-MF [16] modified by an acceleration algorithm and a median flow algorithm is developed to count the number of road cracks.

The second category is RCNN [11] and its series [17]. These methods first employ a RPN to extract many proposals of potential objects. These proposals are then further refined and classified by a subsequent fully connected network, which outputs the final detection results: category and position. Malini *et al.* [18] proposes a modified VGG architecture in Faster RCNN to detect the pavement damages. Xu *et al.* [19] adopts a two-stage Mask R-CNN to inspect road cracks, achieving an excellent detection performance. However, the above conventional approaches require a large number of annotated samples for training, which are expensive to obtain. Thus, exploring FSOD methods is extremely necessary.

B. Few-Shot Object Detection

1) Meta-Learning: Meta-learning expects to learn the task-level meta knowledge, which allows a detector to rapidly detect novel classes using a few training samples. Kang *et al.* [20] designs a novel meta-YOLO detector that extracts general meta-features and reweights these features with query embedding. Meta RCNN [21] transforms Faster RCNN into a meta-learner, which uses a soft-attention mechanism to re-weight region of interest (RoI) features and detect few-shot objects. Based on Meta RCNN, FSIW [22] aggregates more complex features learned from abundant base classes, achieving better detection performance in novel categories. To fully exploit the features of novel objects, Hu *et al.* [23] designs a dense relation distillation with context-aware aggregation (DCNet) to tackle the few-shot detection problem. Built on meta-YOLO [20], CME [24] first converts the few-shot detection problem to a few-shot classification problem, and achieves better novel class detection performance.

2) Metric-Learning: Metric learning is also commonly referred to as similarity learning. The purpose of metric learning is to measure the similarity between two input images or features. It can be generalized to novel categories with few training samples. Commonly used metric approaches include cosine similarity, dot-product similarity, and euclidean distance. Dong *et al.* [25] introduces a simple metric learning-based method for few-shot road damage classification. This method employs a novel metric loss to minimize the distance between the same class and maximize the distance between different classes. In this paper, a Proposal Feature Metric module is designed based on the scaled cosine similarity, which achieves impressive performance in measuring the representation space of proposal features.

3) Transfer-Learning: There are several transfer learning-based detectors for FSOD. Wang *et al.* [10] proposes a two-stage fine-tuning approach (TFA) for novel instance detection. This method fine-tunes the last layers of the existing detector while freezing other parameters to enable the detector to adapt to novel categories. Subsequently, several works (FSCE [27] and DeFRCN [28]) established on TFA are proposed. FSCE [27] introduces the supervised contrastive learning to acquire robust feature representations for novel objects. DeFRCN [28] proposes an offline single-scale PCB to calibrate the misclassified box of novel classes. However, this existing method [28] focuses on single-scale calibration, which neglects useful information from different scales. Thus, an SPCB is proposed to rectify the classification results, which fills the representation gap between the multi-scale distributions of objects.

C. Ghost Operation and Attention Mechanism

The ghost module [29] is a variant of convolution operation, which can produce more informative features with less computational cost. The attention mechanism [30] is used to imitate the human visual processing mechanism, which can suppress irrelevant information and focus on objects. In recent years, attention mechanism has been widely used in damage detection. Su *et al.* [32] proposes a complementary attention block to suppress the noise background in photovoltaic anomaly detection [33], [34], [35], [36]. Dong *et al.* [25] introduces a channel-spatial attention module to extract robust features in road damage detection. Inspired by the aforementioned studies, we design a novel Ghost Attention (GA) module to eliminate the redundant information in road images.

III. Method

In this section, the preliminary knowledge of FSOD is first briefly introduced. The base detector with two additional modules: GA and PFM is then described. Finally, our three-stage fine-tuning approach (FSRDD)are presented.

A. Few-Shot Detection Setting

There are a number of samples for base classes C_b and a few samples for novel classes C_n. We expect to train an efficient detector based on the unbalanced dataset $D = \{(x^*,\ y^*),\ x^* \in X^*,\ y^* \in Y^*\}$, where x^* denotes the input image and $y^* = \{(cls_i,\ box_i),\ i = 1, \ldots, \widetilde{M}\}$ denotes the classes $cls \in C_b \cup C_n$ and the ground-truth bounding-box coordinates box of $\widetilde{M}$ object instances in the image x^*. However, because of the few annotated samples for novel class C_n, the detection model cannot be trained efficiently. Therefore, we firstly use abundant samples of the base classes to train a detector and transfer the general knowledge learned from the base classes to the novel categories.

B. Base Detector

As shown in Fig. 2, Faster RCNN with a feature pyramid network (FPN) [17] is used as the base detector, which can be divided into four parts: feature extraction (backbone and FPN), region proposal (RPN), RoI feature extraction (RoI pooling and RoI feature extractor), and prediction (box regressor and classifier). Specifically, a pre-trained ResNet101 [37] is adopted as the backbone to extract deep features of road images and FPN is applied to aggregate the multi-level features of different scales. These multi-level features are then transmitted to Ghost Attention RPN (GARPN), which employs two sub-networks to predict the proposals: objectness and box position. These proposals are aligned to the multi-level features to obtain the proposal features by RoI pooling. After being refined by the RoI feature extractor that consists of two fully connected layers, these proposal features are further measured by PFM module for the final class and box prediction.

1) Ghost Attention Module: The GA module is composed of a ghost module [29] and a soft-attention module [30] with a residual connection. These two modules are connected in series, and hard-swish and hard-sigmoid activation functions [31] are incorporated to boost the GA module to fit the data distribution of road images. As shown in Fig. 2, feature Y' is supposed as "ghost" of the intrinsic feature X' with linear transformation, where the linear transformation represents the convolution operation without bias item ($conv^{no\_bias}$). The input $X \in \mathbb{R}^{c \times w \times h}$ are the multi-level features from FPN, where c, w, and h denote the channel number, width, and

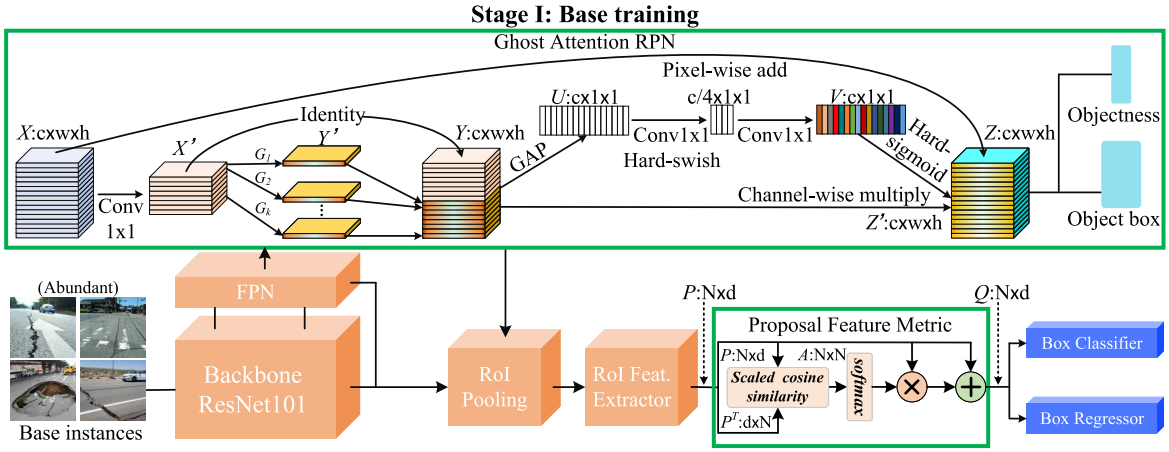

Fig. 2. The first stage: base training of our FSRDD. Base classes with abundant samples are used to train the detector and general knowledge is learned from the base classes. X, X', Y', Y, and Z represent the feature maps. G represents the linear operation used to generate the ghost feature map. N and d denote the number and dimension of proposal features, respectively.

height of feature maps. The intrinsic feature $X' \in \mathbb{R}^{\frac{c}{2}\times w\times h}$ is produced through the convolution operation with a small filter size of 1×1. Specifically,

$$X' = hard\_swish(BN(conv_{1\times1}^{no\_bias}(X))), \tag{1}$$

where BN denotes batch normalization. $hard\_swish$ is the activation function, which is defined as:

$$hard\_swish(x) = \begin{cases} 0, & if x \leqslant -3 \\ x, & if x \geqslant 3 \\ \frac{x(x+3)}{6}, & otherwise \end{cases} \tag{2}$$

To obtain the desired ghost feature $Y' \in \mathbb{R}^{\frac{c}{2}\times w\times h}$, a linear transformation is applied to process each intrinsic feature in X':

$$\begin{aligned} y_i' &= G_i(x_i') = hard\_swish(BN(conv_{3\times3}^{no\_bias}(x_i'))), \\ i &= 1, \ldots, k,\ k = c/2, \end{aligned} \tag{3}$$

where x_i' represents the i-th feature map in the intrinsic feature X'. G_i represents the i-th linear transformation that is used to produce the i-th ghost feature map y_i' in Y'. The ghost feature Y' concatenates with the identity intrinsic feature X' to generate the output $Y \in \mathbb{R}^{c\times w\times h}$ of the ghost module:

$$Y = concat(X', Y'), \tag{4}$$

where $concat$ represents the concatenation operation. Since the output feature Y builds on the identity mapping and linear transformation of the intrinsic feature, the ghost module aggregates more informative features than the traditional convolution operation. These informative features will be further processed by the following soft-attention module.

Specifically, given an input $Y \in \mathbb{R}^{c\times w\times h}$, a global average pooling (GAP) is employed to squeeze the input feature into a channel descriptor $U \in \mathbb{R}^{c\times1\times1}$.

$$U_l = \frac{1}{w * h}\sum_{j=1}^{w}\sum_{k=1}^{h} Y_l(j,\ k),\ l = 1, \ldots, c. \tag{5}$$

The global spatial information is effectively aggregated by the GAP operation. Two convolution operations with a hard-swish activation function ($conv - hard\_swish - conv$) are then used to fully exploit the spatial information.

$$V = conv_{1\times1}(hard\_swish(conv_{1\times1}(U))). \tag{6}$$

After the hard-sigmoid activation, the activated channel descriptor $V \in \mathbb{R}^{c\times1\times1}$ is multiplied with the input Y. And then the output Z' of the soft-attention module is gained:

$$Z' = Y * hard\_sigmoid(V), \tag{7}$$

$$hard\_sigmoid(x) = \begin{cases} 0, & if\ x \leqslant -2.5 \\ 1, & if\ x \geqslant 2.5 \\ 0.2x + 0.5, & otherwise \end{cases}. \tag{8}$$

The final output $Z \in \mathbb{R}^{c\times w\times h}$ of the GA module is defined as:

$$Z = X + Z', \tag{9}$$

As shown in Fig. 2, the GA module is directly embedded into the RPN for the proposal prediction. Due to the inhibitory effect of the GA module on the complex background, the novel GARPN can aggregate more informative features and extract more refined proposals than the traditional RPN. A detailed ablation study is conduct in Section IV to demonstrate its effectiveness.

2) Proposal Feature Metric Module: The flowchart of the proposed PFM module is presented in Fig. 2. The PFM is employed to measure similarities between the proposal features extracted by RoI feature extractor. This module is based on the multiplication of the scaled cosine similarity matrix with the original input feature to obtain a small intra-class variance and larger inter-class difference of proposal features. In detail, the proposal feature $P \in \mathbb{R}^{N\times d}$ is multiplied with its transposition to calculate the similarity matrix between different proposal features, where N denotes the number of proposals and d represents the dimension of a proposal feature. The cosine similarity with a learnable factor δ is innovatively employed to calculate the proposal-similarity matrix $A \in \mathbb{R}^{N\times N}$:

$$A(r, s) = \delta\frac{P_r \cdot P_s^T}{||P_r||||P_s||}, \tag{10}$$

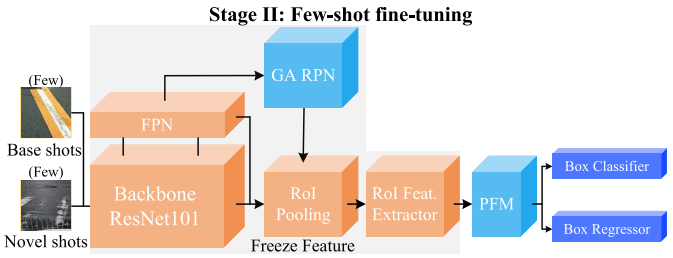

Fig. 3. The second stage: the few-shot fine-tuning stage of our FSRDD. The feature extraction modules are frozen and only the last layers of the base detector are fine-tuned with a balanced few-shot dataset from base and novel classes.

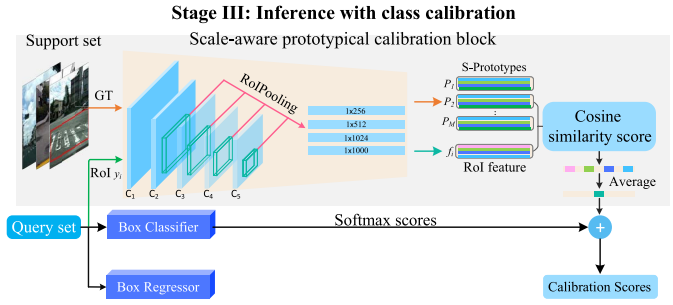

Fig. 4. The third stage is the inference stage of our FSRDD. The support set represents the few training shots in stage II, and scale-aware prototypes (S-prototypes) are the representative vectors for each class. The query set is the test data.

where δ denotes the learnable scale factor and $||\cdot||$ represents the L2 norm. The $r \in \{1,\ldots,N\}$ and $s \in \{1,\ldots,N\}$ denotes the index of a proposal-feature vector, and $A(r,s)$ denotes the similarity between the r-th proposal feature P_r and the s-th proposal feature P_s. The matrix operation symbol $\cdot$ represents the dot product. The similarity matrix $A(r,s)$ is normalized by the following $softmax$ transformation. It is performed on each row of the similarity matrix and defined as:

$$f_{r,s} = \frac{exp(A(r,\ s))}{\sum_{r=1}^{N} exp(A(r,\ s))}, \tag{11}$$

where $f_{r,s}$ measures the class-aware similarity between the r^{th} proposal and s^{th} proposal. The normalized similarity map is multiplied with the proposal features P, followed by an element-wise summation with P to obtain the final output $Q \in \mathbb{R}^{N\times d}$ of the PFM:

$$Q = f * P + P. \tag{12}$$

The metric-proposal feature Q will be applied to predict the final category and position of road damage. The PFM focuses on instance-level similarity measurement rather than image level, which can be generalized to novel classes with few training instances.

C. Few-Shot Road Damage Detector

Our FSRDD includes three stages: base-training, fine-tuning, and inference stages.

1) Base Model Training Stage: In the base training stage, the feature extractor and the box predictor are trained on base classes C_b with abundant samples. The loss function L used to optimize the base detector is referred from [17]:

$$L = L_{cls}^{rpn} + L_{loc}^{rpn} + L_{cls}^{rcnn} + L_{loc}^{rcnn}, \tag{13}$$

where L_{cls}^{rpn} and L_{loc}^{rpn} are used to divide the foreground and background of the extracted proposal by RPN. For the R-CNN branch, L_{cls}^{rcnn} represents the cross-entropy loss for the instance-level classification, and L_{loc}^{rcnn} denotes the smooth-L1 loss for bounding box prediction. Moreover, the parameters of the proposed GA module are also optimized during the training process, which does not require an additional loss function for optimization.

2) Few-Shot Fine-Tuning Stage: In the few-shot fine-tuning stage, a small balanced training set including base and novel classes is adopted to fine-tune the pre-trained base model of the first stage. As shown in Fig. 3, a novel-class weighting parameter for the prediction network is randomly initialized. Then, we fine-tune the last layers of the detection model, while the other parameters are frozen. The loss function is similar to Eq. 13, and the learning rate is less than the pre-defined value in the first stage. In addition, following [10], a *cosine* classifier is adopted to predict the classes in last layer of the detector, which is expressed as:

$$logit_{c,d} = \varepsilon \frac{Q_c^T \cdot w_d}{||Q_c||\ ||w_d||}, \tag{14}$$

where $logit_{c,d}$ denotes the similarity value between the c-th proposal feature Q_c and the d-th weight vector w_d of the class d. Following [10], the parameter ε is set to 20 in all experiments.

3) Inference Stage With Class Calibration: Based on the fine-tuned model obtained in the second stage, an offline SPCB is proposed in the inference stage to calibrate the class of predicted box. The existing method [28] limits to a single-scale calibration, resulting in the scarcity of scale robustness problem for the few-shot detector. Furthermore, due to insufficient training data, this problem is magnified in FSOD. To tackle this issue, an SPCB is proposed to rectify the classification, which can effectively narrow such a gap between the multi-scale distributions of objects.

As shown in Fig. 4, SPCB encompasses a ImageNet pre-trained Resnet101 model and a RoIPooling layer. Specifically, given an M-category E-shot support set S with the ground-truth boxes $\{\{b_{m,e}^{S}\}_{m=1}^{M}\}_{e=1}^{E}$, the SPCB first employs the ImageNet pre-trained Resnet101 to extract feature maps, and then applies RoIPooling with the ground truths to the multiple scale-aware layers C_2, C_3, C_4, C_5 to generate the instance-level representations $r_{m,e}^{C_2}$, $r_{m,e}^{C_3}$, $r_{m,e}^{C_4}$, and $r_{m,e}^{C_5}$, respectively.

After averaging according to the category label m, the scale-aware prototype $\{\{P_m^{C_l}\}_{l=2}^{5}\}_{m=1}^{M}$ is obtained, which is a multi-scale representation for each class:

$$P_m^{C_l} = \frac{1}{|S_m|}\sum_{e=1}^{E} r_{m,e}^{C_l}, \tag{15}$$

where C_l represents the l-th scale layer and S_m denotes a subset containing samples with the same label m in S. SPCB

Algorithm 1 Scale-Aware Prototypical Calibration Block

#Step 1: Build scale-aware prototypes.

Input: Support set S, ground-truth (gt) boxes $\{\{b^S_{m,e}\}^M_{m=1}\}^E_{e=1}$.

Extract multi-level RoI features $\{\{\{r^{C_l}_{m,e}\}^5_{l=2}\}^M_{m=1}\}^E_{e=1}$ of all gt boxes b^S.

Average RoI features by category label m to obtain scale-aware prototypes $\{\{P^{C_l}_m\}^5_{l=2}\}^M_{m=1}$ using Equation 15.

Output: Scale-aware prototypes $\{\{P^{C_l}_m\}^5_{l=2}\}^M_{m=1}$.

#Step 2: Execute calibration.

Input: a query image x_i, a RoI $y_i = (m_i, b_i, \widehat{s_i})$ for x_i.

Extract multi-level RoI features $\{f^{C_l}_i\}^5_{l=2}$.

Calculate auxiliary calibration score $\widehat{s}^{\cos}_i$ using Equation 16.

Weighted aggregation between the original softmax score $\widehat{s_i}$ and $\widehat{s}^{\cos}_i$ to obtain the calibrated score s_i using Equation 17.

Output: Calibrated score s_i

first performs RoIPooling on the predicted box b_i to generate the multi-scale RoI feature $f_i = [f^{C_2}_i; f^{C_3}_i; f^{C_4}_i; f^{C_5}_i]$, when an object proposal $y_i = (m_i, b_i, \widehat{s_i})$ of a query image x_i is input into the offline SPCB, where m_i is the predicted category, b_i is the boundary of the predicted box, and $\widehat{s_i}$ is the classification score. The cosine similarity score is then computed between f_i and $P_{m_i} = [P^{C_2}_{m_i}; P^{C_3}_{m_i}; P^{C_4}_{m_i}; P^{C_5}_{m_i}]$ on different scales. After averaging, the auxiliary calibration score $\widehat{s}^{\cos}_i$ is obtained:

$$\widehat{s}^{\cos}_i = \sum_{l=2}^{L=5} \frac{f^{C_l}_i \cdot P^{C_l}_{m_i}}{||f^{C_l}_i|| \, ||P^{C_l}_{m_i}||} / (L-1), \tag{16}$$

where L denotes the number of scale layers. Finally, the weighted aggregation is performed between the $\widehat{s}^{cos}_i$ from the SPCB and $\widehat{s_i}$ from the few-shot detector to output the calibrated classification score s_i:

$$s_i = \lambda \widehat{s_i} + (1-\lambda)\widehat{s}^{cos}_i, \tag{17}$$

where λ represents the balance factor. Our SPCB is summarized in Algorithm 1. The SPCB loads ImageNet pre-training weights for class-aware inference and does not participate in training. The difference between SPCB and PCB [28] is that our SPCB is a multi-scale calibration method that aggregates the scale-aware features from the multi-scale layers to calculate the calibration scores. It enhances the robustness of the network to scale variation. Using SPCB for the category-auxiliary prediction can greatly improve the accuracy of rare road damage detection.

IV. Experiments

In this section, extensive experiments are conducted on the RDD dataset [5] and our CNRDD dataset. For a fair comparison, we run 10 trials and report the average performance.

RDD	Damage type	Annotation	Number
Base classes C_b	Linear crack, longitudinal, wheel-marked part	D00	6592
	Linear crack, lateral, equal interval	D10	4446
	Alligator crack, partial pavement, overall pavement	D20	8381
	Pothole	D40	5627
	White line blur	D44	5057
	Utility hole	D50	3581
Novel classes C_n	Linear crack, longitudinal, construction joint part	D01	179
	Linear crack, lateral, construction joint part	D11	45
	Cross walk blur	D43	793

Fig. 5. The data distribution of the RDD dataset. The base classes include six categories: D00, D10, D20, D40, D44, D50. The novel classes include three categories: D01, D11, D43.

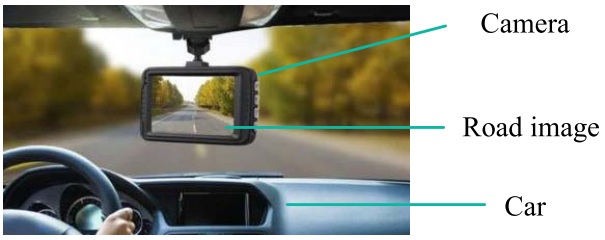

Fig. 6. The designed vehicle-mounted city road image acquisition system.

China-RDD	Damage type	Annotation	Number
Base classes C_b	Crack	sickness_1	2566
	Longitudinal Crack	sickness_3	3185
	Lateral Crack	sickness_4	2970
	Subsidence	sickness_5	1396
	Rutting	sickness_6	4841
	Strengthening	sickness_11	5294
Novel classes C_n	Pothole	sickness_8	471
	Looseness	sickness_9	716
	Uncertain	sickness_-1	114

Fig. 7. The data distribution of the our CNRDD dataset. The base classes include six categories: sickness_1,3,4,5,6,11. The novel classes include three categories: sickness_8,9,-1.

A. Experimental Setup

1) Dataset: The RDD dataset [5] contains 14.5k RGB images with nine categories. The weather is sunny and overcast during the day. The number of lanes is one or two, and the road colors are black, gray, or dark gray. As shown in Fig. 5, the base classes include six categories, and the few-shot novel classes include three categories. E shots mean E instances used for training. As presented in Fig 6, our CNRDD dataset[1] is constructed using a designed vehicle-mounted city road image acquisition system, which consists of a car and a camera. CNRDD includes 4319 RGB images with a resolution of 1600×1200 pixels. The weather is sunny during the day, and the number of the lanes is two. The road colors are black,

[1]https://transport.ckcest.cn/CatsCategory/asphaltRoadDiseases/1

TABLE I

FEW-SHOT ROAD DAMAGE DETECTION PERFORMANCE FOR NOVEL/BASE CLASSES ON RDD DATASET

Method (RDD)	Novel set (mAP50)				Base set (mAP50)			
	1	3	5	10	1	3	5	10
Meta YOLO [20]	3.1	8.6	11.4	12.6	40.2	42.2	43.3	44.1
CME [24]	4.2	10.3	11.8	13.2	41.5	42.2	43.5	44.1
TFA w/fc [10]	5.5	15.4	18.2	19.8	**46.3**	**46.6**	48.4	51.0
TFA w/cos [10]	6.6	15.6	18.3	19.7	44.5	45.2	46.8	48.9
DeFRCN [28]	10.6	19.9	23.9	28.2	45.0	45.3	46.3	50.0
Our FSRDD	**14.1**	**24.6**	**27.8**	**33.4**	45.8	46.5	**48.5**	**51.7**

TABLE II

FEW-SHOT ROAD DAMAGE DETECTION PERFORMANCE FOR NOVEL/BASE CLASSES ON CNRDD DATASET

Method (CNRDD)	Novel set (mAP50)				Base set (mAP50)			
	1	3	5	10	1	3	5	10
Meta YOLO [20]	1.9	2.7	3.6	4.2	26.2	26.9	27.2	27.4
CME [24]	2.2	2.8	3.9	4.5	26.1	27.0	27.5	27.6
TFA w/fc [10]	2.9	3.2	4.3	5.1	30.4	**33.8**	**33.9**	34.5
TFA w/cos [10]	2.8	3.3	4.3	5.2	30.2	33.5	33.8	34.4
DeFRCN [28]	5.2	7.9	8.4	9.3	30.3	32.9	33.0	33.3
Our FSRDD	**7.7**	**9.4**	**11.3**	**12.9**	**30.8**	33.0	33.3	**34.6**

gray, or dark gray. As shown in Fig. 7, the base classes include six categories, and the few-shot novel classes include three categories. The above two datasets are divided into train and test sets in a ratio of 2:1. A validation set is not needed. Furthermore, 1, 3, 5, and 10 shots are selected from each novel class for training, and the rest is used for testing.

2) Evaluation Metric: As evaluation metrics, average precision (AP) and average recall (AR) are chosen to evaluate the detection performance of different algorithms. AP_M, AP_L, AR_M, and AR_L are selected to evaluate the performance of scale-variation damages, where M and L denote medium (32^2 <area< 96^2) and large (area> 96^2) damages. Note that there are no small damages (Area< 32^2) for the novel classes of the RDD dataset. mAP50 is the mean AP50 value of all classes, where the intersection of union (IoU) between the groundtruth and the predicted box > 50% is regarded as a true-positive (TP) box. Please see COCO metric[2] for better understanding.

3) Parameter Setting: Following [10], all the models are trained using a stochastic gradient descent optimizer with a mini-batch size, momentum, and weight decay of 8, 0.9, and 0.0001, respectively. Learning rate of 0.02 and 0.01 are used during base training and few-shot fine-tuning, respectively. The base training stage involves 15,000 steps. Moreover, 1-shot, 3-shot, 5-shot, and 10-shot fine-tuning stages involve 800, 1600, 2000, and 4000 steps, respectively.

B. Quantitative Evaluation

1) Results on RDD: Extensive experiments on the base and novel classes of the RDD dataset are presented in Table I. Our FSRDD is compared with Meta YOLO [20], CME [24], TFA [10], and DeFRCN [28] to explain its superiority in few-shot road damage detection. In detail, FSRDD outperforms DeFRCN in extremely low shot cases such as 1 shot and 3 shots. It demonstrates that our FSRDD exhibits high performance when the number of training samples is extremely small. Among all compared methods, the proposed FSRDD obtains the best few-shot road damage detection results (14.1%, 24.6%, 26.4%, and 30.9% for 1 shot, 3 shots, 5 shots, and 10 shots respectively). This highlights the effectiveness of our proposed extra blocks: GA, PFM, and SPCB. Regarding previous approaches [10], [20], [24], [28], the proposed FSRDD outperforms them by a large margin, which proves that the proposed FSRDD is effective to detect the rare road damages using a few training samples. Furthermore, as presented in Table I, our FSRDD can maintain the detection results of the base classes as high as possible, demonstrating that our method insists on the less forgetting attributes for the base classes.

2) Results on CNRDD: The results obtained on our CNRDD dataset are shown in Table II. Due to insufficient training data (≈2.9k), the detector cannot learn excellent transferred knowledge from the CNRDD dataset in the base training stage. However, although CNRDD is a challenging dataset, compared to the best method (DeFRCN), our FSRDD still achieves 2.5%, 1.5%, 2.9%, and 3.6% improvements on 1, 3, 5, and 10-shot respectively, verifying its effectiveness again. Comparing to these baselines, the proposed method surpasses them and thus is of merit, the absolute mAP values indicate that much room for improvement is still needed for the future works.

C. Ablation Study and Visual Analysis

Ablation studies are conducted on the RDD dataset to evaluate the effectiveness of the proposed modules.

1) Impact of Ghost Attention Module: Several ablation experiments are conduct to evaluate the effectiveness of the proposed GA module. As illustrated in Table III, FSRDD with a GA module outperforms the previous baseline (TFA w/cos) and FSRDD with only an attention (A) or a ghost (G) module. This demonstrates the effectiveness of the GA module to improve the detection performance of the base and novel classes. Fig. 8 depicts the deep features and proposals generated using PRN and GARPN. Comparing to the previous RPN, GARPN not only accurately focuses on the novel-class features (the third column of Fig. 8), but also extracts more refined proposals of the few-shot categories (the fifth column). This verifies that GARPN is substantially better at recommending object proposals.

2) Impact of Proposal Feature Metric Module: The PFM module is employed to measure the proposal features in the base detector. As presented in Table III, the detection model performs better than the baseline when a PFM is incorporated into FSRDD. For 1-shot novel set, FSRDD with a PFM achieves 7.5% mAP50 and has a 0.9% higher rate of accuracy than the baseline. A larger gap of 2.0% is obtained in 10-shot rare road damage detection. This proves that PFM promotes our FSRDD with a more discriminative feature representations. Furthermore, Fig. 9 shows the influence of scale factor δ on the detection performance of novel classes

[2]https://cocodataset.org/#detections-eval

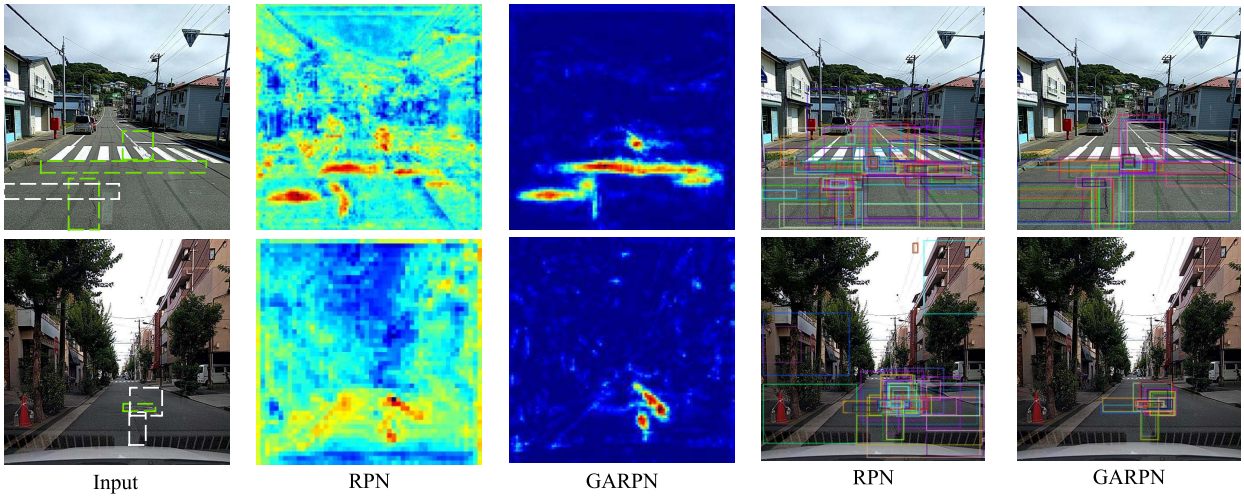

Fig. 8. Visualizations of features and proposals generated by PRN and GARPN. Ground truths of novel and base classes are shown in white and green dashed lines, respectively. The last two columns show the top 100 proposals of RPN and GARPN.

TABLE III

ABLATION STUDY ON NOVEL/BASE SET OF RDD TO EVALUATE THE PERFORMANCE OF DIFFERENT MODULES IN THE PROPOSED FSRDD

G	A	GA	PFM	PCB [28]	SPCB	Novel set (mAP50) 1	5	10	Base set (mAP50) 1	5	10
TFA w/cos (Baseline)						6.6	18.3	19.7	44.5	46.8	48.9
✓						6.8	18.5	19.9	44.8	47.1	49.6
	✓					6.8	18.6	19.8	44.9	47.0	49.8
		✓				6.9	18.9	20.0	45.2	47.2	50.0
			✓			7.5	19.1	21.7	45.1	46.8	49.2
				✓		9.8	22.1	26.4	45.7	46.3	49.0
					✓	12.9	25.0	30.1	45.8	46.4	49.2
		✓	✓		✓	**14.1**	**27.8**	**33.4**	**45.8**	**48.5**	**51.7**

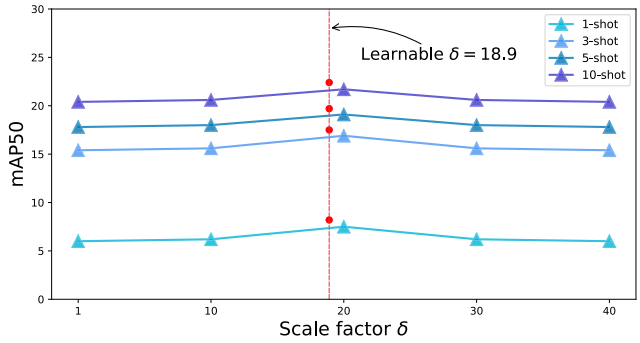

Fig. 9. Influence of Scale factor δ for novel classes of RDD.

in the RDD dataset. Obviously, PFM with a learnable δ outperforms other manual settings for any shot case, which verifies the significance of the learnable parameter.

3) Impact of Scale-Aware Prototypical Calibration Block: As illustrated in Table III, SPCB has greatly promoted the detection performance of rare road damages. Furthermore, as illustrated in Table IV, aggregating distinct scales (C_2, C_3, C_4, C_5) achieves a large improvement than using only the single-scale C_5 (PCB [28]) in the recognition of medium and large objects, respectively. Specifically, for 5-shot, our FSRDD with an SPCB (C_2, C_3, C_4, and C_5) outperforms FSRDD with

TABLE IV

THE PERFORMANCE OF FSRDD IN DISTINCT LAYERS

Method (RDD)	5-shot (novel set) AP_M	AP_L	AR_M	AR_L	10-shot(novel set) AP_M	AP_L	AR_M	AR_L
C_5 (PCB [28])	4.0	12.3	5.9	28.5	6.0	15.7	8.2	29.7
C_4, C_5	4.7	13.1	7.0	29.4	6.8	16.5	9.1	30.9
C_3, C_4, C_5	5.8	14.7	8.6	30.2	7.5	17.3	10.2	32.1
C_2, C_3, C_4, C_5	**6.2**	**15.1**	**9.4**	**31.9**	**8.4**	**18.5**	**11.9**	**33.4**

a PCB (C_5) by +2.2, +2.8, +3.5, and +3.4 in terms of AP_M, AP_L, AR_M, and AR_L, respectively. This demonstrates that our scale-aware calibration approach performs well with scale-variation objects. Fig. 10 explicitly shows the scale-aware embedding distribution of 10-shot support images in distinct layers. For the novel class id-6, the embedding space in layer C_5 is discrete, but it is compact in layer C_3. The multi-scale embedding space can compensate for the discrete embedding space of the single-scale caused by the lack of scale information. This proves that the multi-scale calibration (SPCB) outperforms the single-scale (PCB).

More visual analysis is shown in Fig. 11. Missing cases caused by the single-scale PCB are rescued by our scale-balance SPCB. Specifically, as shown in the green box of Fig. 11, the medium damages (D01 and D11) are discarded by the PCB [28]. However, our SPCB balances the decision information from different layers and obtains a fairer and better calibration effect than the single-scale PCB. This verifies the superiority of our scale-aware calibration approach. Simultaneously, as shown in the red boxes of Fig. 11, this judgment can also be obtained for the large-scale damages.

D. Failure Analysis

Fig. 12 depicts a confusion matrix of the 10-shot RDD set. The red boxes in Fig. 12 show that the novel classes D01, D11, and D43 can easily be mistaken as D00, D10, and D44, respectively. Referring to the damage type in Fig. 5, the attributes of D01, D11, and D43 are similar to D00, D10,

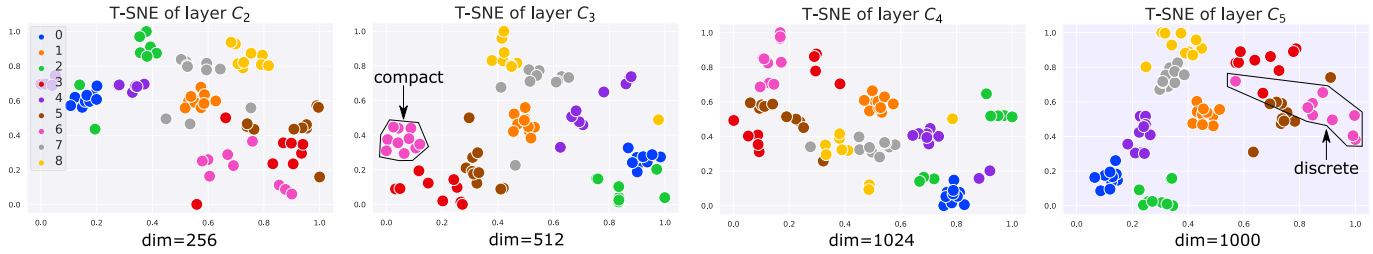

Fig. 10. T-SNE visualization of scale-aware embedding at distinct layers (C_2, C_3, C_4, C_5) on the 10-shot support images of RDD dataset. The *dim* denotes the dimension of the embedding. ID-6,7,8 are the novel classes.

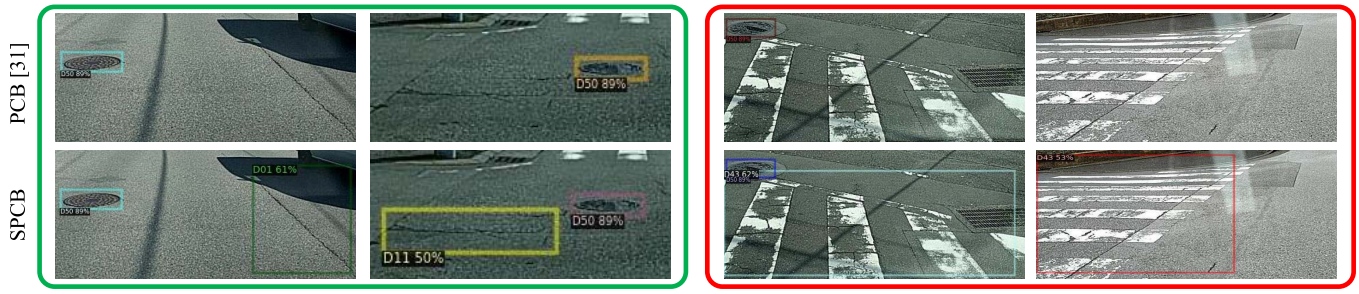

Fig. 11. Visualization of the detection results for the 10-shot RDD dataset. D01, D11, and D43 are novel classes, and we set the score threshold to 0.5. Green box is the medium object (32^2 <area< 96^2), and red box is the large object (area> 96^2).

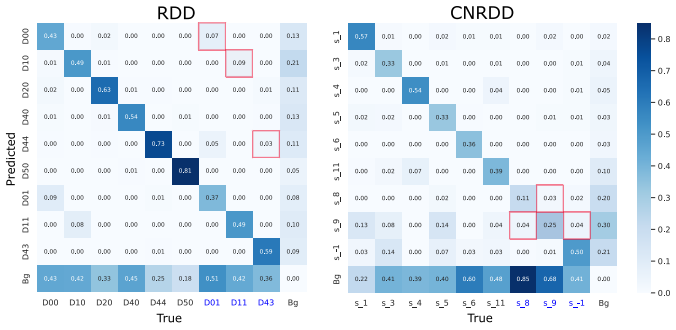

Fig. 12. Confusion matrices of the 10-shot setting in RDD and CNRDD datasets. Bg: background.

and D44, respectively, causing the failure cases. However, the similarity between the attributes is a minor reason for the detection errors. The main failure is the missed detection that damages are erroneously identified as background by our FSRDD, which is a common issue in RDD and CNRDD datasets. This is due to the insufficient training data for the new classes, which leads to a serious over-fitting problem.

E. Time Efficiency

The evaluation of the time efficiency is listed in Table V. The experimental environment is a server with an i7-10700 CPU and an RTX3090 GPU. Our FSRDD has a speed of 60.9 ms/img, which is moderate compared with other methods. The parameter number of our FSRDD is comparable to that of the baseline TFA [10], which presents that the proposed modules such as GA and PFM are lightweight.

TABLE V
EVALUATION OF TIME EFFICIENCY

Time efficiency	Meta YOLO [20]	CME [24]	TFA [10]	DeFRCN [28]	FSRDD
Speed (ms/img)	70.9	75.4	59.8	54.5	60.9
Parameter number (M)	64.7	66.8	60.3	52.0	60.4

V. CONCLUSION

In this paper, we have presented a few-shot road damage detector (FSRDD) to solve the rare road damage detection problem. First, GA module is designed to fully exploit the valuable information extracted by the network. Furthermore, a PFM module is proposed to adaptively measure the proposal features. During inference, an SPCB is proposed to boost the performance for rare road damage detection. The effectiveness of each component is verified through ablation studies. To the best of our knowledge, this is the first work to introduce a few-shot detection approach to detect rare road damages, and achieves impressive performance on the RDD and CNRDD datasets. Our future work will continue to focus on how to improve the detection performance of the model based on a few training samples.

REFERENCES

[1] P. Subirats, J. Dumoulin, V. Legeay, and D. Barba, "Automation of pavement surface crack detection using the continuous wavelet transform," in *Proc. Int. Conf. Image Process.*, Oct. 2006, pp. 3037–3040.
[2] R. Medina, J. Llamas, E. Zalama, and J. Gomez-Garcia-Bermejo, "Enhanced automatic detection of road surface cracks by combining 2D/3D image processing techniques," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 778–782.
[3] R. Kapela, P. Rydzewski, and M. Wyczaek, "Asphalt surfaced pavement cracks detection based on histograms of oriented gradients," in *Proc. 22nd Int. Conf. Mixed Des. Integr. Circuits Syst. (MIXDES)*, 2015, pp. 579–584.

[4] Y. Hu and C. Zhao, "A novel LBP based methods for pavement crack detection," *J. Pattern Recognit. Res.*, vol. 5, no. 1, pp. 140–147, 2010.

[5] H. Maeda, Y. Sekimoto, T. Seto, T. Kashiyama, and H. Omata, "Road damage detection and classification using deep neural networks with smartphone images," *Comput. Aided Civil Infrastruct. Eng.*, vol. 33, no. 12, pp. 1127–1141, Jun. 2018.

[6] H. Maeda, T. Kashiyama, and Y. Sekimoto, "Generative adversarial network for road damage detection," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 36, no. 1, pp. 47–60, 2021.

[7] F. Yang, L. Zhang, S. Yu, D. V. Prokhorov, X. Mei, and H. Ling, "Feature pyramid and hierarchical boosting network for pavement crack detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 4, pp. 1525–1535, Apr. 2020.

[8] W. Tang, S. Huang, Q. Zhao, R. Li, and L. Huangfu, "An iteratively optimized patch label inference network for automatic pavement distress detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 8652–8661, Jul. 2022, doi: 10.1109/TITS.2021.3084809.

[9] Y.-X. Wang, D. Ramanan, and M. Hebert, "Meta-learning to detect rare objects," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9924–9933.

[10] X. Wang, T. Huang, and T. Darrell, "Frustratingly simple few-shot object detection," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 9861–9870.

[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[12] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525.

[13] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, Jun. 2016, pp. 21–37.

[14] X. Gan, J. Qu, J. Yin, W. Huang, Q. Chen, and W. Gan, "Road damage detection and classification based on M2det," in *Proc. Adv. Artif. Intell. Secur.*, 2021, pp. 429–440.

[15] Q. Wang, J. Mao, and X. Zhai, "Improvements of YOLOv3 for road damage detection," *J. Phys., Conf. Ser.*, vol. 1903, no. 1, pp. 26–31, 2021.

[16] D. Ma, H. Fang, N. Wang, C. Zhang, J. Dong, and H. Hu, "Automatic detection and counting system for pavement cracks based on PCGAN and YOLO-MF," *IEEE Trans. Intell. Transp. Syst.*, early access, Mar. 31, 2022, doi: 10.1109/TITS.2022.3161960.

[17] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[18] A. Malini, P. Priyadharshini, and S. Sabeena, "An automatic assessment of road condition from aerial imagery using modified VGG architecture in faster-RCNN framework," *J. Intell. Fuzzy Syst.*, vol. 40, no. 6, pp. 11411–11422, Jun. 2021.

[19] X. Xu *et al.*, "Crack detection and comparison study based on faster R-CNN and mask R-CNN," *Sensors*, vol. 22, no. 3, pp. 1215–1232, 2022.

[20] B. Kang, Z. Liu, X. Wang, F. Yu, J. Feng, and T. Darrell, "Few-shot object detection via feature reweighting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8419–8428.

[21] X. Yan, Z. Chen, A. Xu, X. Wang, X. Liang, and L. Lin, "Meta R-CNN: Towards general solver for instance-level low-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9577–9586.

[22] X. Yang and M. Renaud, "Few-shot object detection and viewpoint estimation for objects in the wild," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 192–210.

[23] H. Hu, S. Bai, A. Li, J. Cui, and L. Wang, "Dense relation distillation with context-aware aggregation for few-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10180–10189.

[24] B. Li, B. Yang, C. Liu, F. Liu, R. Ji, and Q. Ye, "Beyond max-margin: Class margin equilibrium for few-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7363–7372.

[25] H. Dong, K. Song, Q. Wang, Y. Yan, and P. Jiang, "Deep metric learning-based for multi-target few-shot pavement distress Classification," *IEEE Trans. Ind. Informat.*, vol. 18, no. 3, pp. 1801–1810, Jun. 2022.

[26] K. J. Joseph, S. Khan, F. S. Khan, and V. N. Balasubramanian, "Towards open world object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 5826–5836.

[27] B. Sun, B. Li, S. Cai, Y. Yuan, and C. Zhang, "FSCE: Few-shot object detection via contrastive proposal encoding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7348–7358.

[28] L. Qiao, Y. Zhao, Z. Li, X. Qiu, J. Wu, and C. Zhang, "DeFRCN: Decoupled faster R-CNN for few-shot object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8661–8670.

[29] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1577–1586.

[30] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[31] S. R. Dubey, S. K. Singh, and B. B. Chaudhuri, "Activation functions in deep learning: A comprehensive survey and benchmark," 2021, *arXiv:2109.14545v2*.

[32] B. Su, H. Chen, P. Chen, G. Bian, K. Liu, and W. Liu, "Deep learning-based solar-cell manufacturing defect detection with complementary attention network," *IEEE Trans. Ind. Informat.*, vol. 17, no. 6, pp. 4084–4095, Jun. 2021.

[33] B. Su, H. Chen, Y. Zhu, W. Liu, and K. Liu, "Classification of manufacturing defects in multicrystalline solar cells with novel feature descriptor," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 12, pp. 4675–4688, Dec. 2019.

[34] B. Su, H. Chen, K. Liu, and W. Liu, "RCAG-Net: Residual channelwise attention gate network for hot spot defect detection of photovoltaic farms," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–14, 2021.

[35] B. Su, H. Chen, and Z. Zhou, "BAF-detector: An efficient CNN-based detector for photovoltaic cell defect detection," *IEEE Trans. Ind. Electron.*, vol. 69, no. 3, pp. 3161–3171, Mar. 2022.

[36] B. Su, Z. Zhou, and H. Y. Chen, "PVEL-AD: A large-scale open-world dataset for photovoltaic cell anomaly detection," *IEEE Trans. Ind. Informat.*, early access, Mar. 29, 2022, doi: 10.1109/TII.2022.3162846.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2016, pp. 770–778.

Binyi Su received the B.S. degree in intelligent science and technology and the M.S. degree in control engineering from the Hebei University of Technology, Tianjin, China, in 2017 and 2020, respectively. He is currently pursuing the Ph.D. degree in computer science and technology with Beihang University, Beijing, China.

His current research interests include computer vision and pattern recognition, intelligent transportation systems, and smart city.

Hua Zhang received the Ph.D. degree in computer science from the School of Computer Science and Technology, Tianjin University, Tianjin, China, in 2015.

He is currently an Associate Professor with the Institute of Information Engineering, Chinese Academy of Sciences. His research interests include computer vision, multimedia, and machine learning.

Zhaohui Wu received the Ph.D. degree in computer science and technology from Beihang University, Beijing, China, in 2015.

He is currently an Associate Researcher with the China Academy of Transportation Sciences. His current research interests include transportation virtual reality, transportation data virtualization, BIM, digital twins, and transportation simulation.

Zhong Zhou received the B.S. degree in material physics from Nanjing University in 1999 and the Ph.D. degree in computer science and technology from Beihang University, Beijing, China, in 2005.

He is currently a Professor and Ph.D. Advisor at the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University. His main research interests include virtual reality, augmented reality, computer vision, and artificial intelligence.