No-Service Rail Surface Defect Segmentation via Normalized Attention and Dual-Scale Interaction

Gongyang Li, Chengjun Han, and Zhi Liu, *Senior Member, IEEE*

***Abstract*— No-service rail surface defect (NRSD) segmentation is an essential way for perceiving the quality of no-service rails. However, due to the complex and diverse outlines and low-contrast textures of no-service rails, existing natural image segmentation methods cannot achieve promising performance in NRSD images, especially in some unique and challenging NRSD scenes, such as low illumination, chaotic background, multiple/tiny defects, and inconsistent defects. To this end, in this article, we propose a novel segmentation network for NRSDs based on normalized attention and dual-scale interaction, named *NaDiNet*. Specifically, NaDiNet follows the enhancement-interaction paradigm. The *normalized channel-wise self-attention module* (NAM) and the *dual-scale interaction block* (DIB) are two key components of NaDiNet. NAM is a specific extension of the channel-wise self-attention mechanism (CAM) to enhance features extracted from low-contrast NRSD images. The softmax layer in CAM will produce very small correlation coefficients which are not conducive to low-contrast feature enhancement. Instead, in NAM, we directly calculate the normalized correlation coefficient between channels to enlarge the feature differentiation. DIB is specifically designed for the feature interaction of the enhanced features. It has two interaction branches with dual scales, one for fine-grained clues and the other for coarse-grained clues. With both branches working together, DIB can perceive defect regions of different granularities. With these modules working together, our NaDiNet can generate an accurate segmentation map. Extensive experiments on the public NRSD-MN dataset with man-made and natural NRSDs demonstrate that our proposed NaDiNet with various backbones (i.e., VGG, ResNet, and DenseNet) consistently outperforms ten state-of-the-art methods. The code and results of our method are available at https://github.com/monxxcn/NaDiNet.**

***Index Terms*— Dual-scale interaction, no-service rail surface defect (NRSD), normalized attention, segmentation.**

Manuscript received 23 April 2023; revised 4 June 2023; accepted 24 June 2023. Date of publication 10 July 2023; date of current version 19 July 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62171269 and in part by the China Postdoctoral Science Foundation under Grant 2022M722037. The Associate Editor coordinating the review process was Dr. Qiang Miao. *(Gongyang Li and Chengjun Han contributed equally to this work.) (Corresponding author: Zhi Liu.)*

Gongyang Li and Zhi Liu are with the Key Laboratory of Specialty Fiber Optics and Optical Access Networks, Joint International Research Laboratory of Specialty Fiber Optics and Advanced Communication, Shanghai Institute for Advanced Communication and Data Science, and the School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China, and also with the Wenzhou Institute of Shanghai University, Wenzhou 325000, China (e-mail: ligongyang@shu.edu.cn; liuzhisjtu@163.com).

Chengjun Han is with the Key Laboratory of Specialty Fiber Optics and Optical Access Networks, Joint International Research Laboratory of Specialty Fiber Optics and Advanced Communication, Shanghai Institute for Advanced Communication and Data Science, and the School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China (e-mail: monxxcn@gmail.comp).

Digital Object Identifier 10.1109/TIM.2023.3293561

I. Introduction

SEGMENTION is a foundational and critical task in the computer vision community, aiming at assigning a category label to each pixel [1], [2], [3], [4]. As a pixel-wise classification task, segmentation can accurately locate the target objects in a scene. Therefore, segmentation systems are very popular in industrial society, especially in steel companies, to determine the exact location of defects [5], [6], [7], [8]. With the help of automated and precise defect segmentation systems, the industrial company can rapidly locate and repair defects to ensure the quality of its products. Nowadays, with the popularization of high-speed rail, the reliability of railway tracks is very important. Thus, in this article, we focus on the no-service rail surface defect (NRSD) segmentation [9], [10], [11], [12], [13] to ensure the reliability of the no-service rail and improve production efficiency.

As we all know, NRSD images are taken in the production workshop, usually with low lighting and chaotic environments. In addition, the defect region also has unique characteristics such as various shapes and scales, high similarity to the non-defect region, complex, and extremely irregular contours, and significant differences between different defects. The above factors result in significant differences between natural images and NRSD images. For example, the former has obvious textures, rich colors, and clear boundaries, while the latter only has low-contrast textures, a single color, and illegible boundaries. Obviously, the advanced natural images segmentation methods [1], [2], [3], [14] are not suitable for NRSD images.

Currently, there are relatively few specialized NRSD segmentation methods available. According to the input data, these NRSD segmentation methods can be divided into two categories, including the single image-based method [9], [10] and the RGB-D-based method [11], [12], [13]. The former follows a traditional input mode, which is the category we focus on in this article, while the latter follows a rising input mode inspired by multimodal data processing methods [15], [16], [17], [18]. Among these methods, MCnet [9] is a pioneer in convolutional neural networks (CNNs) [19]-based NRSD segmentation methods. It collects a big dataset for NRSD segmentation, named NRSD-MN, containing 4101 NRSD images. This dataset provides a data foundation for the development of CNN-based NRSD segmentation methods. Besides, MCnet also provides a model benchmark. The performance of various models in the benchmark indicates that there is significant room for performance improvement in NRSD segmentation and that there is an urgent need to develop specialized CNN-based NRSD segmentation methods.

Motivated by the above observations, in this article, we propose a novel specialized network for NRSD segmentation, termed *NaDiNet*, which explores the enhancement-interaction paradigm to adapt to the unique characteristics of NRSD images. Our main idea is to first enhance the basic features (i.e., intra-level enhancement) and then interact with cross-level features (i.e., inter-level interaction). The self-attention mechanism [20], [21] is a commonly used feature enhancement manner. However, we found that the classical channel-wise self-attention mechanism (CAM) does not work well in NRSD images. So, we analyze the presentation form of NRSD images in detail, and extend CAM to enhance features extracted from NRSD images more effectively in NaDiNet. Furthermore, based on the enhanced features, we focus on capturing the cross-level context to perceive defect regions from multiple perspectives, which is conducive to segmenting defects completely.

Specifically, we achieve our NaDiNet in the classic encoder-decoder architecture for image segmentation [2]. Our NaDiNet consists of four components, including the feature extractor, the normalized channel-wise self-attention module (NAM), the dual-scale interaction block (DIB), and the segmentation head (SegHead). Notably, NAM is tailored for NRSD images to achieve effective feature enhancement. We argue that the softmax layer in CAM significantly reduces the value of correlation coefficients in the attention map, which is detrimental to low-contrast NRSD images. Thus, in NAM, we remove the original softmax layer of CAM and directly normalize the dependency relationships between channels to obtain the normalized correlation coefficients. DIB is connected after NAM, and is proposed for cross-level feature interaction. We design two interaction branches in DIB, one for a large-scale interaction and the other for a small-scale interaction, and capture the common knowledge of both branches for perceiving defect regions comprehensively. In this way, our NaDiNet can segment NRSD more accurately than all the compared methods and generate satisfactory segmentation maps.

Our main contributions are in three aspects.

1) We explore the enhancement-interaction paradigm, and propose a novel specialized solution, *NaDiNet*, for NRSD segmentation. In NaDiNet, we develop an extended normalized attention and an effective dual-scale interaction manner for enhancement and interaction, respectively, resulting in our method exhibiting an excellent performance on the NRSD-MN dataset, even with different backbones.
2) We extend the vanilla CAM, and propose the NAM to directly model the normalized long-range dependencies between channels, i.e., without using the softmax layer for relative importance measurement, to effectively adapt to the unique scenes of NRSDs for feature enhancement.
3) We propose the DIB to perform a cross-level feature interaction at dual scales and extract the common knowledge at both scales, comprehensively capturing the fine-grained and coarse-grained clues of NRSDs for subsequent segmentation.

The rest of this article is organized as follows. In Section II, we review the related work of CNN-based natural image segmentation and NRSD segmentation. In Section III, we elaborate our method in detail. In Section IV, we present the experiments, ablation studies, and analysis. In Section V, we summarize the conclusion.

II. Related Work

In this section, we review the representative works of CNN-based natural image segmentation and NRSD segmentation.

A. CNN-Based Natural Image Segmentation

With the use of CNNs in natural image segmentation, the segmentation performance has been greatly improved. As a pioneer work in the field of segmentation based on CNNs, Long et al. [1] proposed first an end-to-end CNN-based segmentation method, namely fully convolutional network (FCN). FCN included an effective skip architecture for integrating multi-level information for accurate segmentation. As a contemporaneous work, Noh et al. [22] proposed the deconvolution network (DeconvNet), which includes not only a convolution network, but also an ingenious DeconvNet. The DeconvNet of DeconvNet gradually inferred the object through the unpooling and deconvolution operations. Similarly, Badrinarayanan et al. [2] proposed the famous SegNet, which is an encoder-decoder architecture, for image segmentation.

The above three efforts have laid the foundation for CNN-based natural image segmentation. Subsequent segmentation methods are mainly developed from multi-scale perception and context information extraction. For example, Chen et al. [3], [23], [24], [25] proposed the Deeplab series of segmentation methods, and perceived multi-scale information with atrous convolutions [i.e., atrous spatial pyramid pooling (ASPP)]. Yang et al. [26] introduced the densely connected structure to ASPP, proposing DenseASPP to achieve the multi-scale perception with a very large-scale range. Zhao et al. [27], proposed the pyramid parsing module to perceive different sub-region representations using pooling layers. For context information extraction, Yuan and Wang [28] adopted the self-attention mechanism to extract the object context. Moreover, Fu et al. [21] applied the self-attention mechanism not only in the position domain but also in the channel domain, proposing DANet to integrate local features with global contextual dependencies. Differently, Huang et al. [29] harvested the context information in a crisscross manner in their proposed crisscross attention. Driven by transformers [20], [30], Xie et al. [31] and Strudel et al. [32] modeled and captured the contextual information more powerfully, and proposed SegFormer and Segmenter, respectively.

Due to differences in the photography environment and shooting targets, the above CNN-based segmentation methods for natural images are not suitable for the NRSD images we care about. But these methods give us some inspiration. Concretely, our solution is based on the famous encoder-decoder architecture [2]. And we extend the traditional self-attention

mechanism [20], [21], which is widely used in natural image segmentation, to NAM for feature enhancement in our solution. Our NAM is designed specifically for NRSD images and can alleviate the shortcoming of CAM's inability to effectively enhance low-contrast NRSD images.

B. NRSD Segmentation

NRSD segmentation can achieve pixel-level defect localization, which can significantly improve the quality of no-service rail. With the development of technology, the segmentation system has also been upgraded from a single camera to a complex system composed of multiple data acquisition devices, resulting in two categories of methods for NRSD segmentation. The first one is based on a single NRSD image. To solve the problem of data shortage, Zhang et al. [9] constructed a big dataset for NRSD segmentation. They explored the multiple types of context information for NRSD segmentation with the pixel-wise contextual attention [33]. Moreover, to alleviate the problem of expensive pixel-level annotations, Zhang et al. [10] also explored NRSD segmentation with image-level annotations. The second one handles RGB-D data, and aims to achieve accurate segmentation with the help of the depth map. Wang et al. [11] explored NRSD segmentation for the first time, and mined the multi-modal complementarity of multiple levels in a multi-modal attention block. Wu et al. [12] repeatedly used multi-modal features to locate defects and refine their boundaries. Zhou and Hong [13] handled multi-modal and multi-level feature integration via a hierarchical exploration strategy.

In addition to the above NRSD segmentation methods, we also introduce segmentation methods for other materials similar to no-service rail, such as steel [34], [35], [36], steel sheet [37], [38], and strip steel [39], [40], [41], [42], [43], [44]. For steel, Qian [34] adopted the model ensemble strategy to achieve good performance in both time and accuracy. Pan and Zhang [35] integrated the position and channel self-attention mechanisms into Deeplabv3+, proposing DAN-Deeplabv3+ for steel defect segmentation. Zhao et al. [36] captured the intrinsic relationship of features at different levels to adapt to the complexity of steel defects. For steel sheet, Zhou et al. [37] regarded defect regions as the salient part of image, and proposed a saliency detection method based on double low-rank and sparse decomposition to segment defects. In [38], they implemented the approach of classification first and then segmentation. For strip steel, similar to [37], many researchers also regarded the surface defects of strip steel as salient regions, and proposed a series of saliency detection methods for strip steel, which explore the prediction-refinement strategy [39], [43], the edge information [40], [42], [43], and the multi-resolution inputs [41]. Moreover, Feng et al. [44] achieved few-shot strip steel surface defect segmentation to improve the performance of defects with insufficient samples and unseen defects.

In general, the specialized methods for NRSD segmentation are quite scarce, especially for the single image-based NRSD segmentation. For MCnet, the only method of this category, it employs the existing pixel-wise contextual attention module to handle the unique NRSD images, which is a suboptimal solution. In our NaDiNet, based on the characteristics of NRSD images, we propose the NAM to enlarge the feature differentiation of the low-contrast features for effective feature enhancement. Our NAM is superior to the contextual attention module (proposed for natural images) used in MCnet. In addition to the feature enhancement module, we propose a tailored DIB instead of the simple feature summation used in MCnet to achieve cross-level feature interaction. Our DIB can capture multi-scale information of defect regions, which is beneficial for handling scenes with multiple defects and defects of various sizes in NRSD images. The above efforts result in our NaDiNet substantially ahead of state-of-the-art methods, including MCnet.

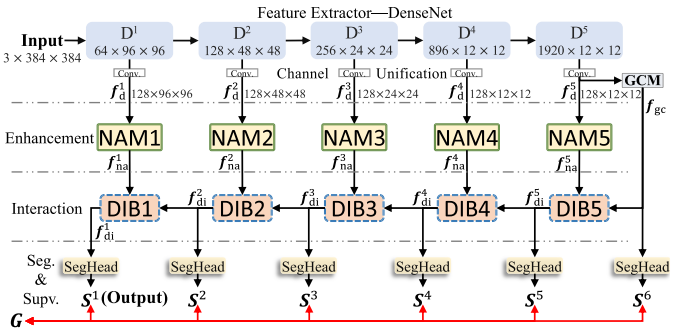

Fig. 1. Pipeline of the proposed NaDiNet for NRSD segmentation. NaDiNet follows the enhancement-interaction paradigm, and consists of the feature extractor, the NAM, the DIB, and the SegHead. The feature extractor is comprised of the DenseNet-201 [45] and a GCM [18], producing five-level basic features and the global context features, respectively. Then, NAM is connected after the feature extractor, and enhances the basic features with normalized attention. DIB is connected after NAM, and performs cross-level feature interaction at dual scales. Finally, the SegHead is connected after the DIB/GCM, and generates the segmentation map. Notably, there are six SegHeads in our NaDiNet. The SegHead after DIB1 generates the output segmentation map of our NaDiNet, while the other SegHeads are used for deep supervision during the network training phase.

III. Proposed Method

In this section, we present our NaDiNet in detail. In Section III-A, we introduce the network overview of our NaDiNet. In Section III-B, we elaborate the proposed NAM. In Section III-C, we elaborate the proposed DIB. In Section III-D, we formulate the loss function.

A. Network Overview

As illustrated in Fig. 1, the proposed NaDiNet follows the enhancement-interaction paradigm, and includes a feature extractor, five NAMs, five DIB, and six SegHeads. For feature extractor, we take the DenseNet-201 [45] as an example here, which can be replaced by other backbones, such as VGG [46] and ResNet [47]. We arrange a convolutional layer with ReLU activation function after each block (denoted as D^i $(i = 1, 2, 3, 4, 5)$) of DenseNet-201 for feature channel unification, extracting five-level basic features $\boldsymbol{f}_d^i \in \mathbb{R}^{c_i \times h_i \times w_i}$ where c_i is 128. Since the size of input NRSD image is 384×384, h_i and w_i belong to $\{96, 48, 24, 12, 12\}$. And the feature extractor also includes a global contextual module (GCM) [18], which contains four parallel convolutional

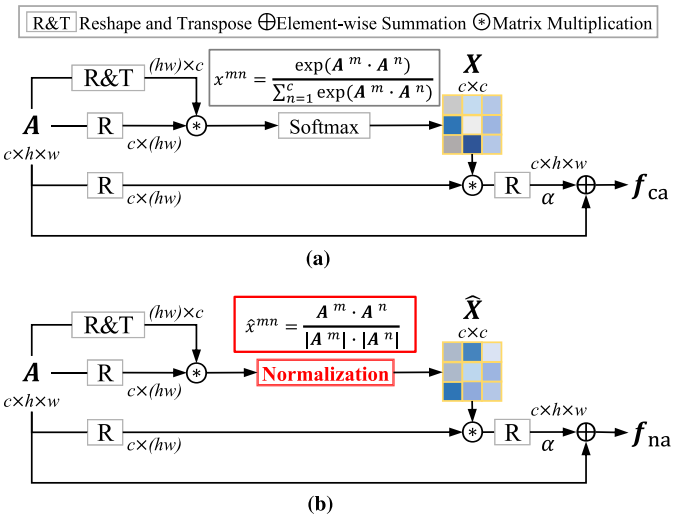

Fig. 2. Illustration of vanilla CAM and our proposed NAM. (a) Vanilla CAM. (b) Our NAM.

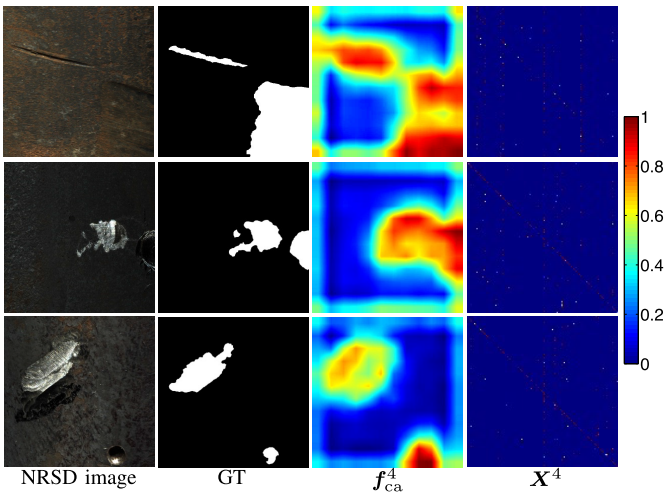

Fig. 3. Feature visualization of the enhanced feature f_{ca}^4 and the channel attention map X^4 in vanilla CAM4. Zoom-in for viewing details.

layers with progressive dilation rates, to extract global context features $f_{gc} \in \mathbb{R}^{c_5 \times h_5 \times w_5}$. Then, in the NAM, we perform feature enhancement on f_d^i by the normalized attention, which is tailored for the low-contrast textures of NRSD images, generating $f_{na}^i \in \mathbb{R}^{c_i \times h_i \times w_i}$. Next, in DIB, we carry out the dual-scale cross-level feature interaction, i.e., we integrate cross-level features on a large scale and a small scale to perceive defect regions of different granularities, generating $f_{di}^i \in \mathbb{R}^{c_i \times h_i \times w_i}$. Except for DIB5 which deals with f_{na}^5 and f_{gc}, the rest of DIBs are used to interact f_{na}^i and f_{di}^{i+1}. Finally, the SegHead consisting of a convolutional layer without the activation function and an upsampling operation is used to generate the segmentation map and restore its size to 384 × 384. In our NaDiNet, the segmentation map generated by the SegHead after DIB1, i.e., $S^1 \in \mathbb{R}^{1 \times 384 \times 384}$, is the output segmentation map of our NaDiNet.

B. Normalized Channel-Wise Self-Attention Module

In this section, we first elaborate the vanilla CAM of DANet [21], which is the basis of our NAM. We illustrate the detailed structure of CAM in Fig. 2(a). Here, for simplicity, we take $A \in \mathbb{R}^{c \times h \times w}$ as the input feature of CAM. Specifically, we reshape A to $\mathbb{R}^{c \times (hw)}$, and then perform the matrix multiplication between the reshaped A and the transpose of the reshaped A. Next, we adopt a softmax layer to produce the channel attention map $X \in \mathbb{R}^{c \times c}$ of CAM, which can be formulated as

$$x^{mn} = \frac{\exp(A^m \cdot A^n)}{\sum_{n=1}^{c} \exp(A^m \cdot A^n)} \tag{1}$$

where x^{mn} measures the correlation between the nth channel and the mth channel, and $x^{mn} \in (0, 1)$ and $\sum_{n=1}^{c} x^{mn} = 1$. In addition, we transfer the correlation of X to the reshaped A by the matrix multiplication, and reshape the generated features to $\mathbb{R}^{c \times h \times w}$. Finally, we fuse the reshaped features and the input A via an element-wise summation with a coefficient (i.e., α) to generate the enhanced feature $f_{ca} \in \mathbb{R}^{c \times h \times w}$.

CAM is designed for natural images with rich texture and color cues. Since NRSD images are usually with low-contrast textures and the single dull tone, CAM is a suboptimal feature enhancement method for NRSD images. For intuitive understanding, we use the CAM in our NaDiNet for feature enhancement, and visualize the output feature f_{ca}^4 of CAM4 in Fig. 3. We observe that defect regions are only highlighted roughly in f_{ca}^4, and even the two defect regions in the second example are connected together, which is not accurate and fine enough. We argue that the generation approach of the channel attention map makes CAM not suitable for NRSD images with low-contrast textures. Concretely, the softmax layer used for relative importance measurement will make $\sum_{n=1}^{c} x^{mn}$ to 1, making the value of each pixel in X very small. Such small values are unfriendly to the low-contrast NRSD images, and cannot clearly highlight the valuable channels. The above inference is also consistent with the corresponding channel attention map X^4 of CAM4 shown in the last column of Fig. 3. For these three examples, X^4 is almost all blue, which indicates that the values of X^4 are close to 0, i.e., X^4 cannot effectively model the relationship between channels for NRSD images. Thus, f_{ca}^4 is not effectively enhanced.

Accordingly, we propose NAM to directly model the normalized relationship between two channels to enlarge the feature differentiation and achieve effective enhancement for NRSD features, generating the enhanced feature $f_{na} \in \mathbb{R}^{c \times h \times w}$. The detailed structure of NAM is illustrated in Fig. 2(b). Specifically, in NAM, we propose to replace the softmax layer with the normalization operation to produce an updated channel attention map $\hat{X} \in \mathbb{R}^{c \times c}$

$$\hat{x}^{mn} = \frac{A^m \cdot A^n}{|A^m| \cdot |A^n|} \tag{2}$$

where $\hat{x}^{mn}$ measures normalized correlation the nth channel and the mth channel. Notably, in our NaDiNet, the input feature A of NAM is actually f_d generated by a convolutional layer with ReLU activation function, resulting in A being non-negative. So $\hat{x}^{mn}$ belongs to [0, 1] without the restriction

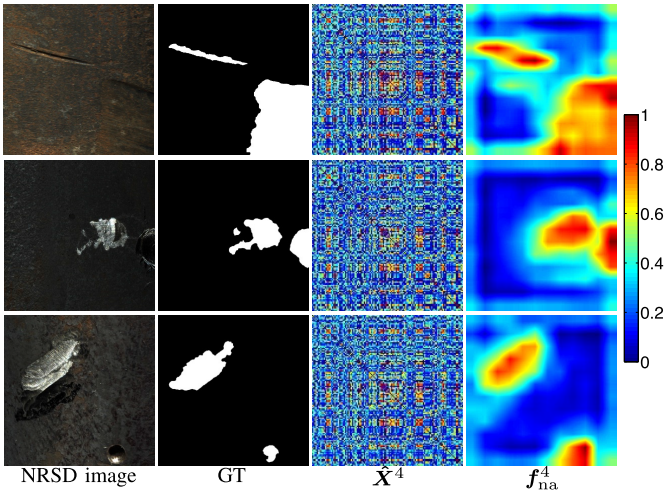

Fig. 4. Feature visualization of the channel attention map $\hat{X}^4$ and the enhanced feature f^4_{na} in the proposed NAM4. Zoom-in for viewing details.

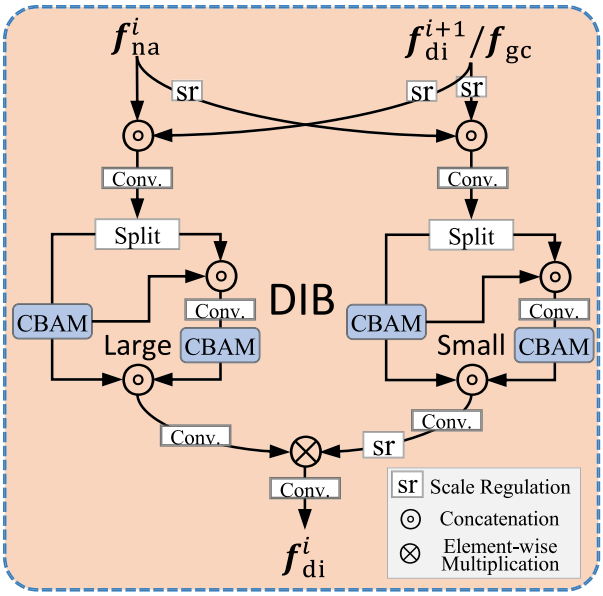

Fig. 5. Illustration of the DIB. CBAM is the classic CBAM [48] with channel attention and spatial attention.

of the sum. We present the updated channel attention map $\hat{X}^4$ of NAM4 in Fig. 4. Compared with X^4, the values of our $\hat{X}^4$ are significantly large. The attention map with large values is indeed suitable for NSRD features. It can enlarge feature channel differences using large values to handle the unique low-contrast characteristics of NSRD features, enabling effective feature enhancement. We also show the enhanced f^4_{na} generated from our NAM4 in the last column of Fig. 4. We can clearly observe that the defect regions of all three examples are highlighted more accurately and finely than those in CAM4.

C. Dual-Scale Interaction Block

NAM only performs the intra-level enhancement, which is not enough for NSRDs with various sizes and shapes. We propose DIB to explore the information between different levels (i.e., inter-level interaction). Moreover, we achieve the inter-level interaction at dual scales to perceive defect regions of different granularities. As presented in Fig. 5, DIB contains a large-scale interaction branch (i.e., the left branch in Fig. 5) and a small-scale interaction branch (i.e., the right branch in Fig. 5). According to the input, we generally formulate the processing of DIB as F(·)

$$f^i_{di} = \begin{cases} \mathrm{F}\left(f^i_{na}, f^{i+1}_{di}\right), & i = 1, 2, 3, 4 \\ \mathrm{F}\left(f^i_{na}, f_{gc}\right), & i = 5 \end{cases} \tag{3}$$

where f^i_{di} is the output feature of DIB. In the following, we introduce these two interaction branches in turn.

1) Large-Scale Interaction Branch: In the large-scale interaction branch of DIBi, we operate features in the size of $h_i \times w_i$. For DIB4 and DIB5, their inputs are with the same size of $h_i \times w_i$, so we forego the scale regulation operation on f^5_{na} and f_{gc}. For the other DIBs, we adopt the upsampling operation as the scale regulation operation, resizing f^{i+1}_{di} to the size of $h_i \times w_i$. Then, we concatenate the inputs of this large-scale interaction branch, and use a convolutional layer to fuse them, generating $f^i_b \in \mathbb{R}^{2c_i \times h_i \times w_i}$. Here, we perform a progressive ensemble manner to group and asymptotically modulate f^i_b with the classic convolutional block attention module (CBAM) [48]. Concretely, we firstly split f^i_b into $f^i_{b1} \in \mathbb{R}^{c_i \times h_i \times w_i}$ and $f^i_{b2} \in \mathbb{R}^{c_i \times h_i \times w_i}$. f^i_{b1} is modulated through the CBAM to highlight defect regions, getting $\hat{f}^i_{b1} \in \mathbb{R}^{c_i \times h_i \times w_i}$. We concatenate f^i_{b2} and $\hat{f}^i_{b1}$, and further modulate their combinations through the CBAM, producing $\hat{f}^i_{b2} \in \mathbb{R}^{c_i \times h_i \times w_i}$. Next, we integrate $\hat{f}^i_{b1}$ and $\hat{f}^i_{b2}$ through the concatenation operation and a convolutional layer, generating the output of the large-scale interaction branch $\hat{f}^i_b \in \mathbb{R}^{c_i \times h_i \times w_i}$. We formulate the above operations as follows:

$$\begin{cases} \hat{f}^i_{b1} = \mathrm{CBAM}\left(f^i_{b1}\right) \\ \hat{f}^i_{b2} = \mathrm{CBAM}\left(f^i_{b2} \circledcirc \hat{f}^i_{b1}\right) \\ \hat{f}^i_b = \hat{f}^i_{b1} \circledcirc \hat{f}^i_{b2} \end{cases} \tag{4}$$

where $\circledcirc$ contains a concatenation operator and a convolutional layer. In this way, we adopt two CBAMs to achieve the gradual expansion of the receptive field in a series-like manner, which can capture multi-scale information of NSRDs.

2) Small-Scale Interaction Branch: In the small-scale interaction branch, we operate features on the size of $(h_i/2) \times (w_i/2)$. For DIB4 and DIB5, we adopt the downsampling operation as the scale regulation operation, resizing the size of both inputs from $h_i \times w_i$ to $(h_i/2) \times (w_i/2)$. For the other DIBs, we only apply the scale regulation operation which is the downsampling operation to f^i_{na}, resizing its size to $(h_i/2) \times (w_i/2)$. And we forego the scale regulation operation on f^{i+1}_{di}. Then, the subsequent operations of this branch are the same as those of the large-scale interaction branch, as formulated in (4). In this way, this branch generates the output $\hat{f}^i_s \in \mathbb{R}^{c_i \times (h_i/2) \times (w_i/2)}$.

3) Branch Integration: Benefitting from the progressive ensemble manner in both interaction branches, the large one can capture the multi-scale fine-grained information, while the small one can capture the multi-scale coarse-grained information. With the above information of different granularities, we adopt the element-wise multiplication operation to

extract the common knowledge of $\hat{\boldsymbol{f}}_b^i$ and the upsampled $\hat{\boldsymbol{f}}_s^i$, generating the output of DIB $\boldsymbol{f}_{\text{di}}^i$. In this way, $\boldsymbol{f}_{\text{di}}^i$ enables a complete characterization of defect regions, which facilitates the segmentation process in the SegHead.

D. Loss Function

As shown in Fig. 1, there are six SegHeads attached after DIBs and GCM. As mentioned in Section III-A, the SegHead after DIB1 generates the output segmentation map of our NaDiNet, while the other SegHeads generate the side segmentation maps which are only used for deep supervision [49], [50]. And all segmentation maps are supervised by the ground truth (GT) binary map. Therefore, the total loss L_{total} of our NaDiNet in the training phase can be calculated as

$$L_{\text{total}} = \sum_{i=1}^{6} \ell_{\text{seg}}^i(\boldsymbol{S}^i, \boldsymbol{G}) \tag{5}$$

where $\boldsymbol{S}^i$ is the predicted segmentation map generated from SegHead, and $\boldsymbol{G}$ is the binary GT map. These two maps have the same resolution of $1 \times 384 \times 384$. To effectively train our NaDiNet and improve the segmentation performance, ℓ_{seg} includes not only the general pixel-level binary cross-entropy loss function ℓ_{bce} but also the map-level intersection over union (IoU) loss function ℓ_{iou}. Thus, L_{total} is updated as

$$L_{\text{total}} = \sum_{i=1}^{6} \left(\ell_{\text{bce}}^i(\boldsymbol{S}^i, \boldsymbol{G}) + \ell_{\text{iou}}^i(\boldsymbol{S}^i, \boldsymbol{G})\right). \tag{6}$$

When performing segmentation reasoning after the network training is completed, we will discard the last five SegHeads and only keep the first one, whose output $\boldsymbol{S}^1$ is the segmentation result of our NaDiNet.

IV. Experiments

A. Experimental Protocol

1) Datasets: We conduct all experiments on the public NRSD-MN dataset [9]. NRSD-MN dataset contains 4101 NRSD images and corresponding pixel-level GTs, of which 3936 images are man-made NRSD images and 165 images are natural NRSD images. For the man-made NRSD images in MCnet [9], 2086 images are used for training, 885 for validation, and 965 images for testing. We only use the above training set of 2086 images and not the validation set during network training. After training, we test the model on the man-made NRSD testing set of 965 images and the natural NRSD set of 165 images.

2) Implementation Details: We achieve our proposed NaDiNet on the PyTorch platform [51], and perform all experiments using an NVIDIA RTX 3090 GPU (24GB memory). The backbone is initialized with the pre-trained parameters, and the other parts are initialized using the "Kaiming" method [52]. For all experiments, the input NRSD image and GT are resized to 384×384. We set the batch size, base learning rate, betas, and eps to 12, $1e^{-3}$, (0.9, 0.999), and $1e^{-8}$, respectively. We train our NaDiNet using the Adam optimizer [53] without data augmentation for 100 epochs until the loss converges.

3) Evaluation Metrics: We use two quantitative evaluation metrics to fairly evaluate the segmentation performance of our NaDiNet and all compared methods, including pixel accuracy (PA), IoU, and F1 score.

PA can represent the percentage of correctly classified pixels in the image, and it can be calculated as follows:

$$\text{PA} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

where TP and FN represent the true positives and false negatives, respectively.

IoU can evaluate the similarities and the differences between the predicted segmentation map and the GT, and it can be calculated as follows:

$$\begin{aligned} \text{IoU} &= \frac{|\boldsymbol{S}^1 \cap \boldsymbol{G}|}{|\boldsymbol{S}^1 \cup \boldsymbol{G}|} \\ &= \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \end{aligned} \tag{8}$$

where FP represents the false positives.

F1 score is the harmonic mean of precision and recall, and it can be calculated as follows:

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{9}$$

The higher these three metrics are, the better the performance.

B. Comparison With State-of-the-Arts

We compare our NaDiNet with a total of ten state-of-the-art CNN-based methods in three categories. The first one is the image segmentation method, including SegNet [2], FCN [1], UNet [14], and Deeplabv3 [3]. The second one is the saliency detection method for strip steel, including CSEPNet [42], DACNet [41], TSERNet [43], EDRNet [39], and EMINet [40]. The last one is the specialized NRSD segmentation method, i.e., MCnet [9]. Except for MCnet dedicated for NRSD segmentation, we retrain the other compared methods with the same dataset and input size (i.e., 384×384) as our method using their default parameter settings, and test them on the man-made NRSD testing set and the natural NRSD set to get the segmentation maps. The above compared methods use different backbones for feature extraction, including VGG [46], ResNet [47], and DenseNet [45]. Here, for a fair comparison, we also provide three variants of our NaDiNet with backbones of VGG, ResNet, and DenseNet, termed Ours-V, Ours-R, and Ours-D, respectively.

1) Quantitative Comparison on Man-Made and Natural NRSDs: We report the quantitative performance of all methods in Table I. Overall, Ours-V, Ours-R, and Ours-D outperform methods with the same backbone as theirs, except for the F1 score of Ours-R on the man-made NRSDs. Our NaDiNet with different backbones consistently outperforms all the compared methods in terms of PA and IoU. This illustrates the effectiveness of our proposed modules and the flexible adaptability of our methods to different backbones. Ours-D improves the specialized MCnet by 4.8% in PA, 3.2% in IoU, and 2.8% in F1 score on man-made NRSDs, and by 3.0% in PA, 2.6% in IoU, and 2.7% in F1 score on natural

TABLE I

QUANTITATIVE RESULTS (%) OF TEN STATE-OF-THE-ART METHODS WITH DIFFERENT BACKBONES AND OUR METHOD WITH THREE DIFFERENT BACKBONES (I.E., VGG, RESNET, AND DENSENET) ON THE NRSD-MN DATASET, INCLUDING THE MAN-MADE NRSDS AND THE NATURAL NRSDS. ↑ MEANS THAT THE BIGGER THE BETTER. THE TOP TWO RESULTS ARE HIGHLIGHTED IN **BOLD** AND UNDERLINED

Methods	Backbone	#Param (M)↓	Speed (*fps*)↑	Man-made (965)			Natural (165)		
				PA↑	IoU↑	F1↑	PA↑	IoU↑	F1↑
SegNet [2]	VGG	29.44	29	75.8	62.3	90.8	57.6	50.8	70.5
FCN [1]	VGG	18.64	31	79.3	64.8	92.0	59.9	53.3	72.5
UNet [14]	VGG	31.04	78	79.8	66.4	93.7	64.0	57.2	76.1
CSEPNet [42]	VGG	18.78	12	81.1	69.3	95.1	65.1	59.1	75.2
Ours-V	VGG	27.68	27	83.9	**71.4**	**97.5**	71.7	63.8	80.0
DACNet [41]	ResNet	98.39	17	77.0	66.3	93.7	61.7	55.4	72.7
Deeplabv3 [3]	ResNet	16.48	27	81.1	67.4	93.1	63.7	57.3	75.3
TSERNet [43]	ResNet	189.64	21	80.7	68.7	93.9	66.2	59.3	75.7
EDRNet [39]	ResNet	39.31	17	81.9	69.0	95.2	68.3	60.5	77.8
EMINet [40]	ResNet	99.13	22	81.7	69.6	96.2	68.6	60.7	79.0
Ours-R	ResNet	48.39	22	83.6	69.7	96.0	70.8	62.7	80.9
MCnet [9]	DenseNet	38.41	18	79.4	68.1	94.4	69.5	62.6	79.8
Ours-D	DenseNet	33.11	13	**84.2**	71.3	97.2	**72.5**	**65.2**	**82.5**

NRSDs. The performance gain of our method is significant. Besides, we found the following phenomena. The performance of the four image segmentation methods (i.e., SegNet, FCN, UNet, and Deeplabv3) is downstream of all compared methods, which means that none of the four image segmentation methods are well adapted to the unique scenes of NRSD images. The performance of saliency detection methods for strip steel is generally better than that of image segmentation methods, which is due to the similarity of strip steel and no-service rail.

2) Parameter and Speed Comparison: Moreover, we measure the network parameters and inference speed (without I/O time) of all methods using the same hardware, and report them in Table I. Overall, under different backbones, our method is among the three methods with the least number of parameters. This indicates that our method is also efficient, and the effectiveness of our method does not lie in blindly increasing the number of parameters. In terms of inference speed, Ours-V and Ours-R achieve near real-time inference speeds of 27 and 22 fps, respectively, which are at the mid-stream level among all comparison methods. Due to the deep network depth of DenseNet, the inference speeds of Ours-D and MCnet are only 13 and 18 fps, respectively.

3) Visual Comparison: In Fig. 6, we compare our NaDiNet equipped with the backbone of DenseNet (i.e., Ours-D) with all state-of-the-art methods visually on the man-made NRSDs and the natural NRSDs. There are several challenging scenes in Fig. 6, including multiple defects, tiny defects, the inconsistent defect, and irregular defects. We can clearly observe that our segmentation maps are close to GTs, while the segmentation maps of other compared methods have the situation of missing defects, incomplete segmentation of defects, or even incorrect segmentation. The satisfactory segmentation maps of our method are attributed to the enhancement-interaction paradigm we adopted, which first enhances features to make defect regions complete and avoid missing defects, and then performs the cross-level interaction to perceive defect regions of various granularities. The segmentation maps of image segmentation methods further demonstrate that these methods do not effectively handle NRSD images. As for the saliency detection methods of strip steel, there is still a certain gap between their segmentation maps and our segmentation maps.

C. Ablation Studies

We conduct exhaustive ablation studies on the NRSD-MN dataset to demonstrate the effectiveness of the proposed modules in our NaDiNet. Concretely, we analyze these modules from the following three aspects: 1) the contribution of NAM and DIB in NaDiNet; 2) the advantages of NAM over vanilla CAM (i.e., NAM versus CAM); and 3) the effectiveness of each part in DIB. Here, we adopt the DenseNet-201 as the backbone of the feature extractor, and perform all ablation experiments with the same parameter settings as in Section IV-A.

1) Contribution of NAM and DIB in NaDiNet: NAM and DIB are the keys in our NaDiNet. To evaluate the contribution of NAM and DIB, we design three variants: 1) Baseline-D, in which we remove all NAMs and DIBs and employ a concatenation operator and three convolutional layers to fuse the cross-level features; 2) Baseline-D + NAM; and 3) Baseline-D + DIB. We provide the quantitative results of the above variants in Table II.

At first glance, we can observe that both modules improve the performance of "Baseline-D" on two test sets. Through detailed comparison, NAM boosts the performance of "Baseline-D" by 0.6% and 1.6% in PA and 0.8% and 1.8% in IoU, and DIB boosts the performance of "Baseline-D" by 1.0% and 2.3% in PA and 1.2% and 2.0% in IoU. With both modules working together, the full NaDiNet outperforms "Baseline-D" by a large margin, especially for natural NSRDs, where its advantages reach 5.2% in PA and 4.8% in IoU. This above

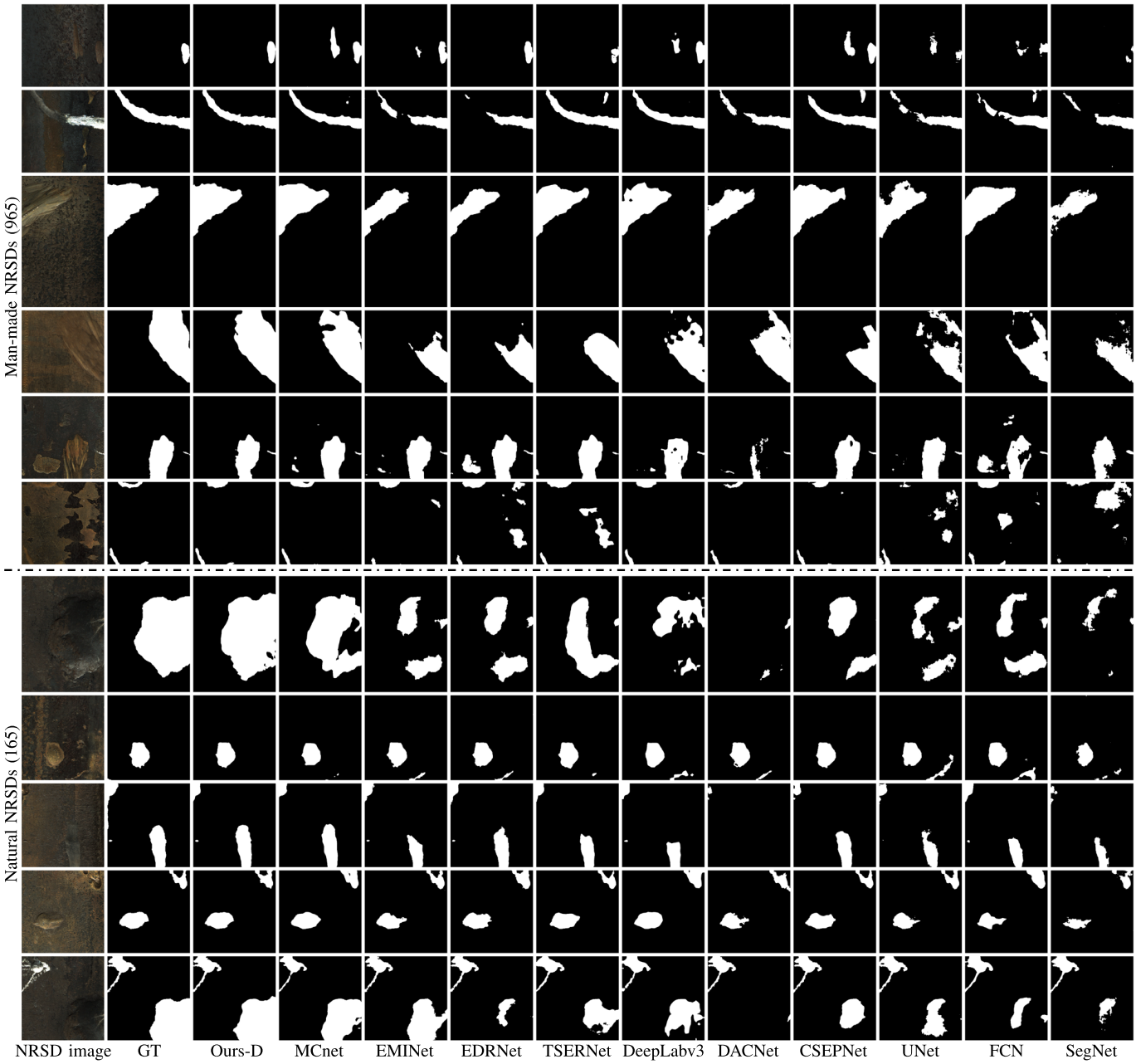

Fig. 6. Visual comparisons with ten state-of-the-art methods on two test sets, i.e., man-made NRSDs (965) and natural NRSDs (165). Please zoom-in for the best view.

analysis intuitively shows the strong contribution of NAM and DIB for feature enhancement and feature interaction, and proves the effectiveness of the enhancement-interaction paradigm in our NaDiNet.

2) Advantages of NAM Over Vanilla CAM (i.e., NAM Versus CAM): As presented in Section III-B, our NAM is proposed specifically for NSRD images. By comparing the channel attention maps and features of NAM and CAM in Figs. 3 and 4, we can visually observe the advantages of our NAM. In this section, we specifically evaluate the advantages of NAM over vanilla CAM from the quantitative perspective. To this end, we conduct a variant named *w/CAM*, which replaces all five NAMs with vanilla CAMs in our NaDiNet. As the quantitative results shown in the first row of Table III, we find that our NAM is indeed better for NSRD images than vanilla CAM, which shows that expanding the value of the channel attention map for low-contrast NSRD images can increase the difference in features to achieve effective feature enhancement. According to the thorough visual and quantitative results, our NAM is superior to vanilla CAM for NSRD images.

3) Effectiveness of Each Part in DIB: DIB is in charge of the interaction of the cross-level features, and contains several parts, such as the large-scale and small-scale interaction branches. In the following, we analyze their effectiveness one by one.

First, to prove the necessity of the scale interaction in both branches, we forego concatenating $\boldsymbol{f}_{\mathrm{na}}^{i}$ and $\boldsymbol{f}_{\mathrm{di}}^{i+1}/\boldsymbol{f}_{\mathrm{gc}}$, and

TABLE II

QUANTITATIVE RESULTS (%) OF EVALUATING THE CONTRIBUTION OF NAM AND DIB IN NADINET. THE SUBSCRIPT IS THE IMPROVED PERFORMANCE COMPARED TO BASELINE-D. THE BEST ONE IN EACH COLUMN IS **BOLD**

No.	Baseline-D	NAM	DIB	Man-made (965) PA↑	Man-made (965) IoU↑	Natural (165) PA↑	Natural (165) IoU↑
1	✓			82.0	69.3	67.3	60.4
2	✓	✓		82.6 +0.6	70.1 +0.8	68.9 +1.6	62.2 +1.8
3	✓		✓	83.0 +1.0	70.5 +1.2	69.6 +2.3	62.4 +2.0
4	✓	✓	✓	**84.2** +2.2	**71.3** +2.0	**72.5** +5.2	**65.2** +4.8

TABLE III

ABLATION STUDY (%) ON EVALUATING THE ADVANTAGES OF NAM AND THE EFFECTIVENESS OF EACH PART IN DIB. THE BEST RESULT IN EACH COLUMN IS **BOLD**

No.	Models	Man-made (965) PA↑	Man-made (965) IoU↑	Natural (165) PA↑	Natural (165) IoU↑
1	*w/ CAM*	83.6	70.6	71.4	64.0
2	*w/o SI*	83.4	70.9	70.0	63.4
3	*w/ large-scale*	83.6	70.7	70.9	63.3
4	*w/ small-scale*	83.7	70.4	70.4	62.2
5	*w/o PE*	83.5	70.6	69.7	62.4
6	**Ours-D**	**84.2**	**71.3**	**72.5**	**65.2**

perform the progressive ensemble on $\boldsymbol{f}_{\text{na}}^{i}$ and $\boldsymbol{f}_{\text{di}}^{i+1}/\boldsymbol{f}_{\text{gc}}$ individually rather than on their combination in *w/SI*. As results shown in the second row of Table III, the scale interaction is integral for cross-level fusion. Then, to prove the effectiveness of performing interactions at dual scales, we provide two variants. The first one *w/large scale* only retains the large-scale interaction branch (i.e., the left one in Fig. 5), while the second one *w/small scale* only retains the small-scale interaction branch (i.e., the right one in Fig. 5). The quantitative performance reported in Table III shows that two different scale interaction branches are indispensable, indicating that interaction at dual scales can indeed capture features of different granularities that are beneficial to NSRD segmentation. Last, we provide a variant *w/PE* to evaluate the superiority of the progressive ensemble manner in both branches, i.e., we directly modulate $\boldsymbol{f}_{b}^{i}$ and $\boldsymbol{f}_{s}^{i}$ with one CBAM. According to the results of the fifth row in Table III, we conclude that the progressive ensemble manner is good for NSRD segmentation, and it can capture multi-scale information to effectively characterize defect regions of different sizes. The above analysis proves that each part of our DIB is effective.

V. CONCLUSION

In this article, we propose a novel specialized NaDiNet for NSRD segmentation, following the enhancement-interaction paradigm. To address the unique low-contrast textures of NSRD images, we update vanilla CAM to NAM for intra-level feature enhancement, which directly calculates the normalized dependencies between channels to expand the value of attention maps. Furthermore, we achieve the inter-level feature interaction in DIB to capture the fine-grained and coarse-grained information of defects, thereby handling scenes with defects of various shapes and scales. In order to accelerate loss convergence and improve the representation ability of features, we adopt the deep supervision strategy during the network training phase. Comprehensive quantitative and visual comparisons, as well as ablation experiments, prove the superior performance of our NaDiNet and the flexibility and advantages of our NAM and DIB. In our future work, we plan to take the segmentation map generated by our NaDiNet as a prompt of the popular segment anything model (SAM) [54] and utilize the powerful segmentation ability of SAM to generate a more accurate segmentation map.

REFERENCES

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[3] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[4] G. Li, Y. Wang, Z. Liu, X. Zhang, and D. Zeng, "RGB-T semantic segmentation with location, activation, and sharpening," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 3, pp. 1223–1235, Mar. 2023.

[5] N. Neogi, D. K. Mohanta, and P. K. Dutta, "Review of vision-based steel surface inspection systems," *EURASIP J. Image Video Process.*, vol. 2014, no. 1, pp. 1–19, Nov. 2014.

[6] Q. Luo, X. Fang, L. Liu, C. Yang, and Y. Sun, "Automated visual defect detection for flat steel surface: A survey," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 3, pp. 626–644, Mar. 2020.

[7] B. Tang, L. Chen, W. Sun, and Z. Lin, "Review of surface defect detection of steel products based on machine vision," *IET Image Process.*, vol. 17, no. 2, pp. 303–322, Feb. 2023.

[8] X. Jin et al., "DM-RIS: Deep multimodel rail inspection system with improved MRF-GMM and CNN," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 4, pp. 1051–1065, Apr. 2020.

[9] D. Zhang, K. Song, J. Xu, Y. He, M. Niu, and Y. Yan, "MCnet: Multiple context information segmentation network of no-service rail surface defects," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–9, 2021.

[10] D. Zhang, K. Song, J. Xu, H. Dong, and Y. Yan, "An image-level weakly supervised segmentation method for no-service rail surface defect with size prior," *Mech. Syst. Signal Process.*, vol. 165, pp. 1–14, Feb. 2022.

[11] J. Wang, K. Song, D. Zhang, M. Niu, and Y. Yan, "Collaborative learning attention network based on RGB image and depth image for surface defect inspection of no-service rail," *IEEE/ASME Trans. Mechatronics*, vol. 27, no. 6, pp. 4874–4884, Dec. 2022.

[12] J. Wu, W. Zhou, W. Qiu, and L. Yu, "Depth repeated-enhancement RGB network for rail surface defect inspection," *IEEE Signal Process. Lett.*, vol. 29, pp. 2053–2057, 2022.

[13] W. Zhou and J. Hong, "FHENet: Lightweight feature hierarchical exploration network for real-time rail surface defect inspection in RGB-D images," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–8, 2023.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Oct. 2015, pp. 234–241.

[15] G. Li, Z. Liu, and H. Ling, "ICNet: Information conversion network for RGB-D based salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 4873–4884, 2020.

[16] G. Li, Z. Liu, M. Chen, Z. Bai, W. Lin, and H. Ling, "Hierarchical alternate interaction network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 3528–3542, 2021.

[17] G. Li, Z. Liu, L. Ye, Y. Wang, and H. Ling, "Cross-modal weighting network for RGB-D salient object detection," in *Proc. ECCV*, Aug. 2020, pp. 665–681.

[18] D.-P. Fan, Y. Zhai, A. Borji, J. Yang, and L. Shao, "BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network," in *Proc. ECCV*, Aug. 2020, pp. 275–292.

[19] Y. LeCun et al., "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, Dec. 1989.

[20] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Dec. 2017, pp. 6000–6010.

[21] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3141–3149.

[22] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1520–1528.

[23] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2015, pp. 1–14.

[24] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[25] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with Atrous separable convolution for semantic image segmentation," in *Proc. ECCV*, Oct. 2018, pp. 833–851.

[26] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "DenseASPP for semantic segmentation in street scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3684–3692.

[27] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[28] Y. Yuan, L. Huang, J. Guo, C. Zhang, X. Chen, and J. Wang, "OCNet: Object context network for scene parsing," 2018, *arXiv:1809.00916*.

[29] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.

[30] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021, pp. 1–22.

[31] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. NeurIPS*, Dec. 2021, pp. 12077–12090.

[32] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7242–7252.

[33] N. Liu, J. Han, and M. Yang, "PiCANet: Learning pixel-wise contextual attention for saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3089–3098.

[34] K. Qian, "Automated detection of steel defects via machine learning based on real-time semantic segmentation," in *Proc. 3rd Int. Conf. Video Image Process.*, Dec. 2019, pp. 42–46.

[35] Y. Pan and L. Zhang, "Dual attention deep learning network for automatic steel surface defect segmentation," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 37, no. 11, pp. 1468–1487, Sep. 2022.

[36] X. Zhao, J. Zhao, and Z. He, "A multiple feature-maps interaction pyramid network for defect detection of steel surface," *Meas. Sci. Technol.*, vol. 34, no. 5, pp. 1–10, Jan. 2023.

[37] S. Zhou, S. Wu, H. Liu, Y. Lu, and N. Hu, "Double low-rank and sparse decomposition for surface defect segmentation of steel sheet," *Appl. Sci.*, vol. 8, no. 9, p. 1628, Sep. 2018.

[38] S. Zhou, H. Liu, K. Cui, and Z. Hao, "JCS: An explainable surface defects detection method for steel sheet by joint classification and segmentation," *IEEE Access*, vol. 9, pp. 140116–140135, 2021.

[39] G. Song, K. Song, and Y. Yan, "EDRNet: Encoder–decoder residual network for salient object detection of strip steel surface defects," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 12, pp. 9709–9719, Dec. 2020.

[40] X. Zhou, H. Fang, X. Fei, R. Shi, and J. Zhang, "Edge-aware multi-level interactive network for salient object detection of strip steel surface defects," *IEEE Access*, vol. 9, pp. 149465–149476, 2021.

[41] X. Zhou et al., "Dense attention-guided cascaded network for salient object detection of strip steel surface defects," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–14, 2022.

[42] T. Ding, G. Li, Z. Liu, and Y. Wang, "Cross-scale edge purification network for salient object detection of steel defect images," *Measurement*, vol. 199, pp. 1–11, Aug. 2022.

[43] C. Han, G. Li, and Z. Liu, "Two-stage edge reuse network for salient object detection of strip steel surface defects," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.

[44] H. Feng, K. Song, W. Cui, Y. Zhang, and Y. Yan, "Cross position aggregation network for few-shot strip steel surface defect segmentation," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–10, 2023.

[45] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

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

[48] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2018, pp. 3–19.

[49] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Proc. AAAI*, vol. 38, May 2015, pp. 562–570.

[50] G. Li, Z. Liu, D. Zeng, W. Lin, and H. Ling, "Adjacent context coordination network for salient object detection in optical remote sensing images," *IEEE Trans. Cybern.*, vol. 53, no. 1, pp. 526–538, Jan. 2023.

[51] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. NeurIPS*, Dec. 2019, pp. 8024–8035.

[52] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

[53] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2015, pp. 1–15.

[54] A. Kirillov et al., "Segment anything," 2023, *arXiv:2304.02643*.

Gongyang Li received the Ph.D. degree from Shanghai University, Shanghai, China, in 2022.

From July 2021 to June 2022, he was a Visiting Ph.D. Student with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He is currently a Post-Doctoral with the School of Communication and Information Engineering, Shanghai University. His research interests include saliency detection, multi-modal image processing, and image/video segmentation.

Chengjun Han received the M.E. degree from Shanghai University, Shanghai, China, in 2023.

His research interests include saliency detection and defect segmentation.

Zhi Liu (Senior Member, IEEE) received the B.E. and M.E. degrees from Tianjin University, Tianjin, China, and the Ph.D. degree from the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China, in 1999, 2002, and 2005, respectively.

From August 2012 to August 2014, he was a Visiting Researcher with the SIROCCO Team, IRISA/INRIA, Rennes, France, with the support by EU FP7 Marie Curie Actions. He is currently a Professor with the School of Communication and Information Engineering, Shanghai University, Shanghai. He has published more than 200 refereed technical papers in international journals and conferences. His research interests include image/video processing, machine learning, computer vision, and multimedia communication.

Dr. Liu is an Area Editor of *Signal Processing: Image Communication* and served as a Guest Editor for the special issue on *Recent Advances in Saliency Models, Applications and Evaluations* in *Signal Processing: Image Communication*.