applied sciences

MDPI

Article

AFFNet: An Attention-Based Feature-Fused Network for Surface Defect Segmentation

Xiaodong Chen [1], Chong Fu [1,2,3,*], Ming Tie [4], Chiu-Wing Sham [5] and Hongfeng Ma [6]

[1] School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China; 2001672@stu.neu.edu.cn
[2] Engineering Research Center of Security Technology of Complex Network System, Ministry of Education, Northeastern University, Shenyang 110819, China
[3] Key Laboratory of Intelligent Computing in Medical Image, Ministry of Education, Northeastern University, Shenyang 110819, China
[4] Science and Technology on Space Physics Laboratory, Beijing 100076, China; tming7611@sina.com
[5] School of Computer Science, The University of Auckland, Auckland 1010, New Zealand; b.sham@auckland.ac.nz
[6] Dopamine Group Ltd., Auckland 1542, New Zealand; wind_ma621@hotmail.com
* Correspondence: fuchong@mail.neu.edu.cn

Abstract: Recently, deep learning methods have widely been employed for surface defect segmentation in industrial production with remarkable success. Nevertheless, accurate segmentation of various types of defects is still challenging due to their irregular appearance and low contrast with the background. In light of this challenge, we propose an attention-based network with a U-shaped structure, referred to as AFFNet. In the encoder part, we present a newly designed module, Residual-RepGhost-Dblock (RRD), which focuses on the extraction of more representative features using CA attention and dilated convolution with varying expansion rates without a concomitant increase in the parameters. In the decoder part, we introduce a novel global feature attention (GFA) module to selectively fuse low-level and high-level features, suppressing distracting information such as background. Moreover, considering the imbalance of the dataset sampled from actual industrial production and the difficulty of training samples with small defects, we use the online hard sample mining (OHEM) cross-entropy loss function to improve the learning ability of hard samples. Experimental results on the NEU-seg dataset demonstrate the superiority of our method over other state-of-the-art methods.

Keywords: deep CNN; surface defect segmentation; U-shape architecture; feature attention; feature fusion

Citation: Chen, X.; Fu, C.; Tie, M.; Sham, C.-W.; Ma, H. AFFNet: An Attention-Based Feature-Fused Network for Surface Defect Segmentation. *Appl. Sci.* **2023**, *13*, 6428. https://doi.org/10.3390/app13116428

Academic Editor: Byung-Gyu Kim

Received: 13 April 2023
Revised: 8 May 2023
Accepted: 18 May 2023
Published: 24 May 2023

1. Introduction

The process of identifying and inspecting abnormalities or irregularities in industrial products is known as surface defect segmentation. The purpose is to ensure that the product meets certain quality standards and is free from defects that could affect its function, performance, or appearance. Current inspection methods can be performed manually by skilled workers, or using automated inspection systems that utilize various techniques such as image processing, computer vision, and machine learning. These systems need to detect various surface defects such as cracks, scratches, dents, and other surface irregularities and provide accurate, fast, and reliable results, thereby increasing the efficiency and reliability of the production process. Unfortunately, this task remains challenging owing to the following concerns: (1) In actual industrial production, the contrast between defective and non-defective parts is often low due to factors such as the environment and production process. (2) Even the same type of defects can have an irregular surface and vary significantly in terms of its area. Figure 1 shows several examples of distinct types of defects commonly observed in industrial production.

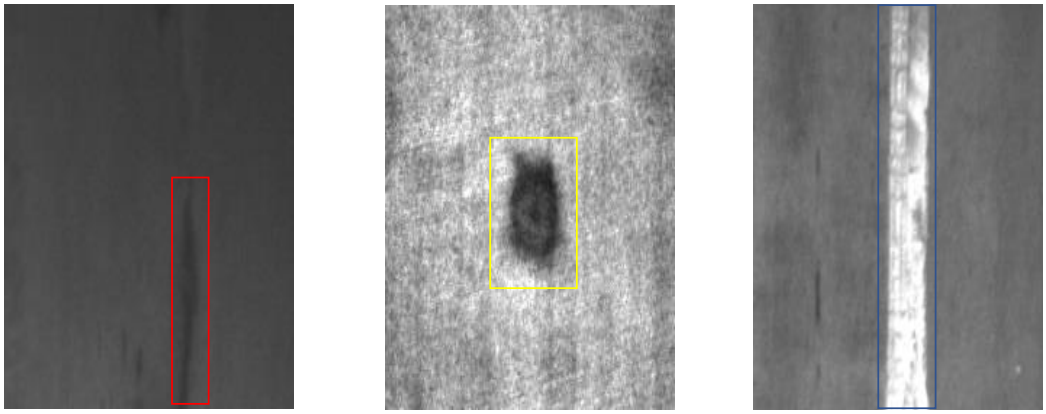


Figure 1. Several examples of surface defects. The red box, yellow box, and blue box show inclusions, patches, and scratches, respectively.

Traditional surface defect detection methods mainly include non-automated detection and automatic detection. Non-automated detection mainly adopts the method of manual participation, but the detection standards of this method cannot be unified, and intense eye use can cause fatigue, so this kind of method cannot adapt to the requirements of industrial development. As computer vision improves by leaps and bounds, automatic surface defect detection is gradually being applied in the industrial field. Zhang et al. [1] employed a Gaussian mixture model for detecting defects on rail surfaces. Additionally, some other hand-crafted feature-based methods, including metals [2] and steel [3], are also employed in surface defect detection and have yielded favorable outcomes in recent years. However, these traditional methods need to manually extract features from images, and they cannot provide satisfactory generalization performance.

Over the past several years, the field of deep learning has witnessed significant progress. The successful application of deep learning methods in diverse domains, including computer vision and multimedia, has been widely recognized and has garnered substantial attention from researchers and practitioners alike. Deep learning methods [4] can extract distinguishing features by constructing convolutional neural networks (CNN) and various nonlinear combinations, leading to a reduction in incompleteness resulting from human-designed features. At present, segmentation methods based on CNNs have been universally employed in surface defect detection by designing different neural networks. The FCN (fully convolutional network) [5] is the most primitive structure that adopts semantic segmentation for network training. It changes the final layer of the model by transforming it into a convolutional layer, enabling the transformation of an arbitrary-sized input into a pixel-level output. Based on FCN, Chen et al. [6] introduced an NB-fully convolutional network, achieving good results in the real-time segmentation of defects. Following these pioneering works, various networks for semantic segmentation have emerged, including UNet [7], SegNet [8], and PSPNet [9], and they exhibit excellent performance on many datasets. In addition, the proposal of networks such as ResNet [10] and DeepLab [11] has also promoted the development of semantic segmentation. For surface defect segmentation, the model needs to effectively use the characteristics of objects to distinguish them from the background, different layers of a CNN model have different sensitivities to objects. Fusion with different features is a crucial measure for improving segmentation performance. Low-level features in a network possess a higher spatial resolution and comprise more precise and detailed information, however, they also tend to obtain lower semantic understanding and are more susceptible to noise due to undergoing fewer convolutional operations. Conversely, high-level features provide more semantic information, but they tend to have a lower spatial resolution and limited perception of fine details.

Most of the existing methods directly upsample high-level features and subsequently fuse them with corresponding low-level ones, leading to redundant information and inefficiency. Xu et al. [12] proposed an attention fusion network for multi-spectral images, incorporating a co-attention module to enhance the correlation in relation between the different features. Yan et al. [13] performed a feature fusion with spatial regions over features at various scales. The feature fusion module implemented three learnable spatial attention masks, and then attached them to different features. Wang et al. [14] introduced a lightweight attention module to fuse the spectral features and spatial features of different layers. Liang et al. [15] fused features of different layers, and then applied an attention module such as SENet to re-weight the channels of mixed features. Although these methods take into account the relationship between different features, they only utilize spatial or channel information to assign weights to the features and fail to fully capture the complex relationships between different features. Furthermore, compared with traditional object segmentation, surface defect segmentation has the following challenges: (1) In real-world industry scenarios, it is usually hard to collect the large amount of training data needed by a deep model. This small sample size problem makes it very challenging to train an accurate defect segmentation model. (2) Intra-class imbalance can cause inaccurate classification of different defect subtypes, while inter-class imbalance can lead to a model biased towards more common defects and poor segmentation performance on rarer defects. (3) Different types of defect often display a great similarity in shape and size, making it difficult for the model to distinguish between them. Meanwhile, even the same type of defects can have irregular surfaces and vary significantly in terms of their area, making it challenging to find a unified feature representation for defect segmentation. (4) Industrial production images collected from industrial sites usually have complex background noise and a low-contrast environment, making it difficult to distinguish defects from the background. Additionally, some other negative factors, including light changes, shadows, and blurring may also have a significant impact on the performance of the defect segmentation model, resulting in a failed segmentation.

To address these issues mentioned above, we propose a novel attention-based feature-fused network (AFFNet) with a U-shaped structure for surface defect segmentation. In detail, a Residual-RepGhost-Dblock (RRD) module is presented in the decoder stage to extract rich detailed feature information. Compared to traditional convolutional layers, the RRD module utilizes residual decoupling in combination with an attention mechanism. It also employs paralleled convolutional kernels with varying expansion rates to enable the module's feature extraction capacity for multiscale targets. Meanwhile, we propose a global feature attention (GFA) module to fuse adjacent resolution features. This module extracts global information from high-level features, and then utilizes the global features to selectively fuse low-level ones, refining the spatial location of category pixels while optimizing the feature channel dimensions. This not only contributes to effective feature propagation but also minimizes the additional computational costs.

Our major contributions are summarized as follows.

- A Residual-RepGhost-Dblock (RRD) module is presented to replace the simple CNN convolution layer, it could be used as a flexible module to enable the network to perform multiscale feature extraction.
- A global feature attention (GFA) module is proposed to selectively fuse feature maps with different resolutions, so as to fuse more contextual semantic information and improve network segmentation performance.
- We adopt the OHEM cross-entropy loss to address the issue of imbalanced samples. As a result, our framework exhibits exceptional performance on the NEU-seg defect dataset, particularly in enhancing the segmentation accuracy of challenging samples.
- A novel attention-based feature-fused network for surface defect segmentation is proposed, our proposed method achieves an 80.94% mIoU on the NEU-seg dataset, outperforming other state-of-the-art methods.

The remainder of this paper is organized as follows. Section 2 presents related work on surface defect detection. Section 3 introduces the proposed AFFNet thoroughly. Section 4 presents the evaluation and the experimental results on the NEU-seg dataset and compares them with classical and other state-of-the-art methods. Finally, in Section 5, a conclusion of our work is presented.

2. Related Work

The existing methods for surface defect segmentation can be roughly classified into traditional approaches and deep learning segmentation methods. Traditional machine learning methods rely on manually designed feature extractors to perform surface defect detection, while deep learning methods are capable of automatically learning a hierarchical representation of features from defect images, enabling them to adapt more easily to various domains and applications.

2.1. Traditional Detection Approaches

Approaches based on computer vision are widely utilized in the inspection of various material surfaces due to their high efficiency and accuracy. For instance, Chu et al. [16] presented a method for extracting features using smoothed LBF. Truong et al. [17] proposed an automatic thresholding strategy that improves upon Otsu's approach, the suggested method enables the detection of incredibly small defect areas. Su et al. [18] proposed the CPICS-LBP method to detect solar cell defects. In order to classify strip surface defects, Luo et al. [19] presented a novel descriptor based on selective LBP, which was combined with the nearest neighbor classifier (NNC) to improve the overall performance. Zhao et al. [20] presented a novel approach that incorporates a training mechanism into the generation of local descriptors, thereby enhancing the accuracy of defect classification. Liu et al. [21] introduced an enhanced algorithm for multi-block LBP feature extraction. By adjusting the block sizes, the method determines the appropriate scale to depict the faulty features in addition to having the simplicity and effectiveness of the LBP algorithm, ensuring high recognition accuracy. Navarro et al. [22] suggested a wavelet reconstruction-based approach for defect detection in various texture images. Refs. [23,24] presented the characteristics and usage of various machine learning methods applied in renewable energy systems. Liu et al. [25] improved the existing fault diagnosis methods for multiphase drive systems by introducing adaptive secondary sampling filtering based on machine learning. Ren et al. [26] suggested a method that used the conventional photometric stereo and broadened its scope to enable the accurate and efficient inspection of non-Lambertian surfaces. Despite their effectiveness, these detection techniques have trouble in picking up small defects or flaws that have textural characteristics blending in with the background. Moreover, several of these techniques lack generality and are difficult to adapt to different kinds of materials since they are restricted to particular materials or types of defects.

2.2. Deep Learning Segmentation Approaches

Deep learning methods are currently widely employed in the segmentation of surface defects. Li et al. [27] presented a coarse-to-fine defect detection framework for detecting surface defects on aero-engine blades. Ju et al. [28] proposed a photometric stereo network guided by the Lambertian model to improve the handling of non-Lambertian surfaces. Ren et al. [29] introduced a method for constructing an inverse reflectance model that characterizes the nonlinear reflectance behavior of non-Lambertian surfaces. Lu et al. [30] introduced a new network, MRD-Net, for detecting surface defects, that utilizes a combination of multiscale feature enhancement fusion and reverse attention. Zhang et al. [31] introduced the FDSNet network on a two-branch architecture for real-time segmenting of surface defects, this network incorporates two auxiliary tasks, aimed at capturing additional boundary details and semantic context. Tian et al. [32] presented a deep adversarial model for surface defect segmentation. To identify the flaws on the strip steel surface, Zhou et al. [33] suggested a unique saliency model to detect salient strip defects.

Li et al. [34] devised a method of integrated coordinate attention mechanism to detect hot-rolled defects. Tang et al. [35] used the spatial attention bilinear CNN to detect casting defects. Pan et al. [36] introduced a pixel-level method for surface defect segmentation that integrates the Deeplabv3+ model with dual parallel attention. Wu et al. [37] suggested a method with boundary guidance for salient rail surface defects detection. Although these segmentation-based surface defect detection models are capable of more accurately segmenting the location and boundaries of defects than traditional methods, further improvement in segmentation accuracy is still necessary.

2.3. Attention Mechanism

The use of attention mechanisms in different scenarios, including image classification, object detection, scene segmentation, and natural language processing has proven to be very beneficial. For computer vision, the attention mechanism can capture the salient regions of an image and selectively attend to them to refine the accuracy of the model. Meanwhile, it can assist the model in dynamically focusing on the information that is of interest, thereby lowering the interference from noise and redundant information and enhancing the model's resilience and generalizability. Hu et al. [38] devised a channel attention mechanism called squeeze-and-excitation (SE), which uses global pooling and nonlinear transformation on the input feature map to learn the correlation among channels. To address the limitations of channel attention, Woo et al. [39] presented the convolutional block attention module. This module can collect the spatial and channel correlations in the input feature map and use them to enhance model performance by combining spatial attention and channel attention. Wang et al. [40] proposed a model based on a non-local attention mechanism, namely non-local neural networks (NLNN), aiming at improving the performance of multiple computer vision tasks by introducing non-local information. Fu et al. [41] suggested suggested a dual attention network for scene segmentation tasks. By integrating an attention mechanism in both space and channel, the DAN module may dynamically choose the regions of interest and learn the correlation between regions, boosting the accuracy and robustness of the segmentation results.

3. Proposed Method

Making use of the effective U-shaped structure, we propose an attention-based feature-fused network for surface defect segmentation, referred to as AFFNet, as illustrated in Figure 2. The architecture comprises two key components: the Residual-RepGhost-Dblock (RRD) module and the global feature attention (GFA) module. The network takes an original image of three channels as input and outputs a segmentation mask for the designated category after post-processing. During the encoding stage, five RRD modules are employed for feature extraction, enabling the network's ability to efficiently extract multiscale features at a low computation cost. Meanwhile, we use a Maxpool2d layer between two RRD modules for down-sampling operations. In the decoder stage, four GFA modules are employed to selectively fuse high-level features and low-level ones for processing distinct levels of output. In the GFA module, we use a bilinear upsampling operation to scale up the feature map, combined with an attention mechanism to selectively highlight important low-level features from both the spatial and channel dimensions, facilitating the effective propagation of informative features from lower to higher levels. After the GFA module, a 3×3 convolution followed by BN layer and ReLU activation is used to reduce the channel dimensions by half. Finally, a 3×3 convolution is employed to produce the output, ensuring the channel numbers correspond to the predefined categories.

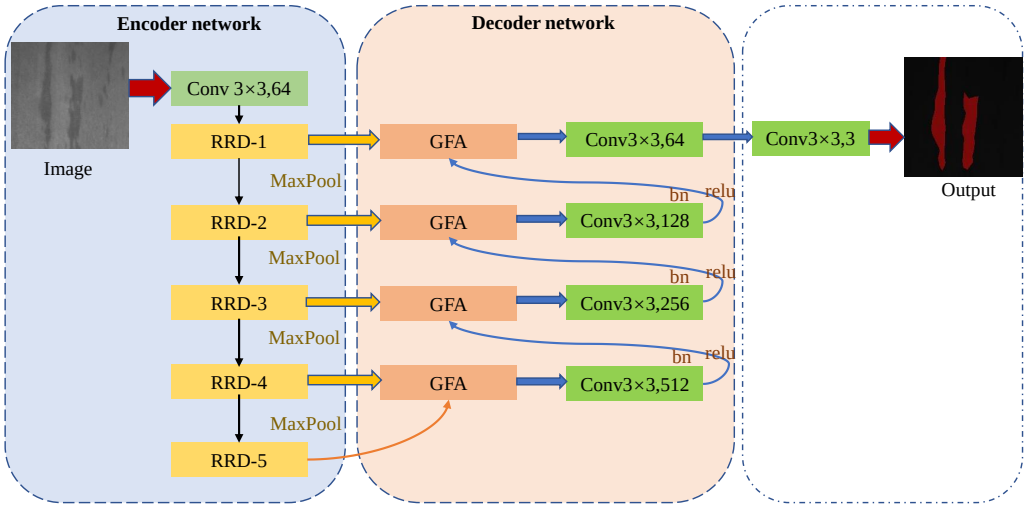


Figure 2. The overview of our proposed AFFNet. MaxPool denotes the maxpool2d layer with a kernel size of 2.

3.1. Residual-RepGhost-Dblock Module

The RepGhost [42] introduced a bottleneck structure that utilizes the RepGhost module to generate and merge various feature maps through reparameterization, thereby eliminating the inefficient concatenate operation of Ghostnet and saving reasoning time. Inspired by the RepGhost bottleneck module, we propose a Residual-RepGhost-Dblock (RRD) module, which also adopts a bottleneck structure, as illustrated in Figure 3. Concretely, first, the input feature map is fed into the RepGhost module, reducing the channel numbers to half of the input feature map, the purpose of this process is to decrease the module's computational complexity and make the module more efficient. Secondly, the resulting feature map is fed into two paralleled depth-wise (DW) convolution blocks for feature extraction. The two depth-wise convolutions, both followed by BN and ReLU, use different dilation rates, one with a 3×3 kernel and a dilation rate of 1, and the other with a 3×3 kernel and a dilation rate of 2. By utilizing different dilation rates, the module can obtain multiscale features without incurring additional computational complexity that would have been associated with using larger convolutional kernels. The two convolutional results are then concatenated to generate a feature map with dimensions equivalent to the original feature map. Subsequently, the feature map is fed into a 3×3 convolutional layer to extract more abstract image features. To enhance the network's sensitivity to both channel-wise and spatial-wise relationships, a coordinate attention (CA) [43] module is added to our proposed module. The CA attention module can focus the network's attention on the specific channels and spatial locations that are most informative, thereby improving the network's performance. After the CA module, a RepGhost module is employed to the features. Finally, the module adopts a residual structure where the input and output features are added to generate the output result. This residual structure enables the network to learn residual mappings and helps to prevent the problem of vanishing gradients, resulting in better network performance.

3.2. Global Feature Attention Module

As aforementioned, many of the existing methods suffer from redundant information or increasing numbers of parameters in the upsampling stage. To address this problem, we introduce a global feature attention module (GFA) that utilizes high-level features to select and fuse relevant low-level features, as shown in Figure 4. The GFA module operates in the following steps: First, a 3×3 convolution followed by a BN layer adjusts the channel dimension of the high-level feature map to make it the same as the low-level feature map. Then, a 3×3 convolution followed by a BN is applied to extract low-level information. Secondly, by utilizing the channel attention module, we can obtain the channel attention

vector from the processed high-level feature map, and the results are then multiplied with the low-level feature map. By assigning distinct weights to the channels of feature maps, the module's capacity to represent features is improved. After that, the size of the high-level features is doubled by using bilinear upsampling. By incorporating the spatial attention module with the previous high-level feature map, we can obtain a spatial vector that has the same dimensions as the low-level feature map. Then, this spatial vector is multiplied with the low-level feature map to extract more relevant contour features about the target. Finally, the processed high-level features and low-level ones are concatenated as the output result.

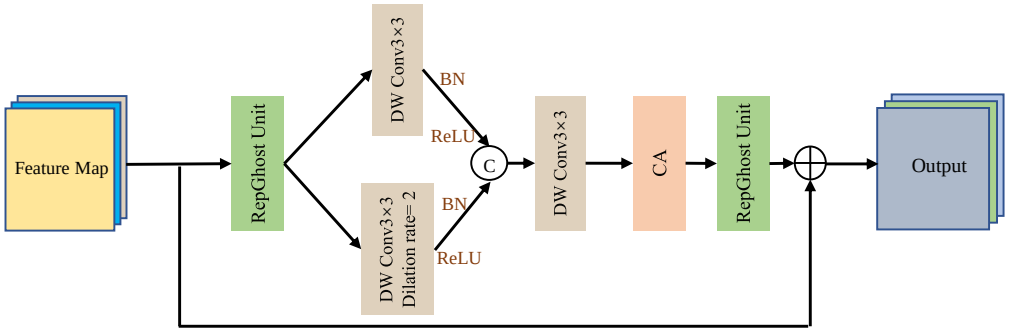


Figure 3. Details of the RRD module, where ⓒ and ⊕ are concatenate and add operations, respectively. DW denotes depth-wise convolution with a kernel size of 3 × 3 and a padding size of 1.

In the GFA module, we have utilized two attention modules, one is the channel attention module, and the other is the spatial attention module. Subsequently, we will present an in-depth exposition of the two modules.

3.2.1. Channel Attention Module

The main idea of this module is to establish correlations between channels and dynamically adjust the channel weight of the input features, resulting in enhanced model performance and improved generalization capability.

First, formally, given an input feature map $f \in \mathbb{R}^{C\times H\times W}$, where H, W, and C denote its height, width, and the channel numbers, respectively, we propose a Channel Attention module to calculate the channel attention vector. The structure of this module is show in Figure 5a.

Secondly, we perform a 1×1 convolution to reduce the channel numbers of the input feature map to $f_{c1} \in \mathbb{R}^{C/4\times H\times W}$, and next we reshape f_{c1} to $f_{c1} \in \mathbb{R}^{C/4\times (HW)}$. Meanwhile, we can obtain $f_{c2} \in \mathbb{R}^{1\times H\times W}$ by performing a 1×1 convolution to reduce the channel numbers of f. Next, we reshape and transpose f_{c2} to $f_{c2} \in \mathbb{R}^{(HW)\times 1\times 1}$. After that, we perform a matrix multiplication between f_{c1} and f_{c2} to obtain $f_c \in \mathbb{R}^{C/4\times 1\times 1}$. The process can be described by

$$f_c = RS(\mathcal{W}_{1\times 1} \odot f + b) \otimes \sigma(RT(\mathcal{W}_{1\times 1} \odot f + b)), \tag{1}$$

where $\otimes$ and $\odot$ indicate element-wise multiplication and the convolution operation, respectively, and $RS(\cdot)$, $RT(\cdot)$, and $\sigma(\cdot)$ denote the reshape operation, reshape and transpose operation, and sigmoid activation, respectively. $W_{1\times 1}$ denotes the filter value, and b indicates the bias value.

Finally, the input feature map is calculated according to Equation (1), then, we can obtain the result $f_c \in \mathbb{R}^{C/4\times 1\times 1}$. Next, we use a 1×1 convolution to make the channel number of f_{s1} the same as the input feature map, then, we employ a sigmoid activation to make it nonlinear. After that, we can obtain the output $y_c \in \mathbb{R}^{C\times 1\times 1}$, which can be described by

$$y_c = \sigma(\mathcal{W}_{1\times 1} \odot f_c + b). \tag{2}$$

3.2.2. Spatial Attention Module

This module can selectively focus on important spatial locations in the input feature map, while suppressing or ignoring less important regions.

First, for a given input feature map $f \in \mathbb{R}^{C\times H\times W}$, where H, W, and C denote its height, width, and the channel numbers, respectively, we can obtain a spatial vector through this attention module, as illustrated in Figure 5b.

Secondly, we can obtain two vectors by calculating the input feature map f with two different pathways, one is f_{s1}, the other is f_{s2}. Specifically, we can obtain the $f_{s1} \in \mathbb{R}^{C/4\times H\times W}$ with a 1×1 convolution, and then we reshape it to $f_{s1} \in \mathbb{R}^{C/4\times HW}$. Meanwhile, we can obtain the $f_{s2} \in \mathbb{R}^{C\times 1\times 1}$ by using global pooling, and next we reduce its channel numbers through a 1×1 convolution, and then we reshape and transpose it to $f_{s2} \in \mathbb{R}^{1\times C/4}$. After that, a sigmoid activation is employed to make it nonlinear.

Thirdly, we apply a matrix multiplication between f_{s1} and f_{s2} to obtain the $f_s \in \mathbb{R}^{1\times HW}$. This process can be described by

$$f_s = RS(\mathcal{W}_{1\times 1} \odot f + b) \otimes \sigma(RT(\mathcal{W}_{1\times 1} \odot (\mathbb{G}(f) + b)), \tag{3}$$

where $\otimes$ and $\odot$ represent element-wise multiplication and the convolution operation, respectively, and $RS(\cdot)$, $RT(\cdot)$, $\mathbb{G}(\cdot)$, and $\sigma(\cdot)$ denote the reshape operation, reshape and transpose operation, global pooling, and sigmoid activation, respectively. $W_{1\times 1}$ denotes the filter value, and b indicates the bias value.

Finally, with the reshaping and sigmoid operations, we can obtain the output $y_s \in \mathbb{R}^{1\times H\times W}$. The process can be described by

$$y_s = \sigma(RS(f_s)). \tag{4}$$

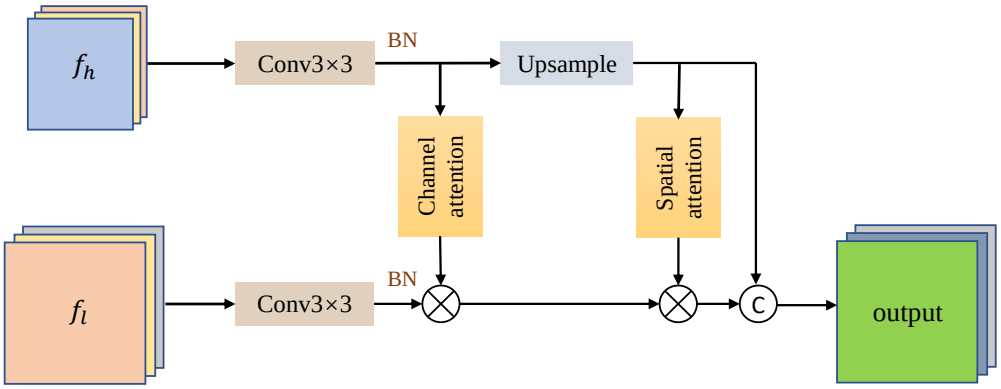


Figure 4. Details of the GFA module, where $\otimes$, ©, f_l, and f_h denote multiplication, concatenate operation, low-level feature map, and high-level feature map, respectively.

Overall, our proposed GFA module addresses the limitations of existing approaches by selectively using both high-level and low-level feature maps and producing more diverse feature representations for improved segmentation performance.

3.3. Loss Function

For semantic segmentation tasks, cross-entropy [44] is generally used as the loss function. Mathematically, it is described by

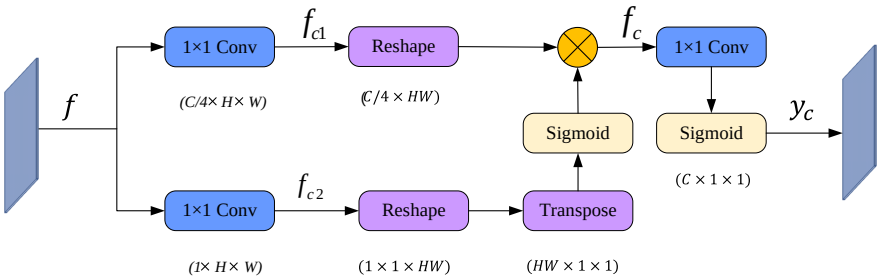
$$L(y, \widehat{y}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}[y_{ik}log(\widehat{y}_{ik}) * (1 - y_{ik})log(1 - \widehat{y}_{ik})], \tag{5}$$

where N refers to the total number of pixels present in a batch, K denotes the category numbers, and y and $\widehat{y}$ indicate the defect labels and prediction results, respectively.

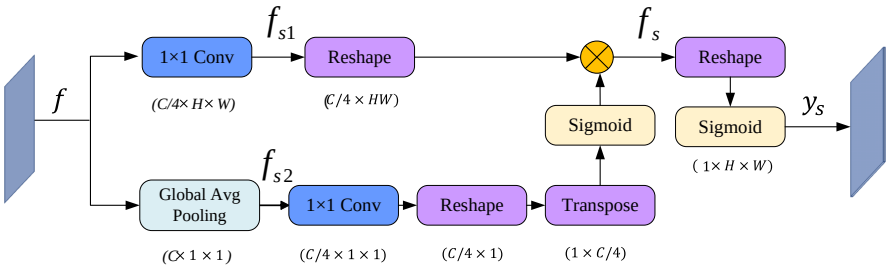
The online hard example mining (OHEM) strategy consists of two phases. First, the model is trained using the standard cross-entropy loss function, which evaluates the difference between the predicted class probabilities and the actual labels. Second, the loss values of all the examples in a mini-batch are sorted from large to small, and the most difficult examples, which have the highest loss values, are chosen for retraining using a modified loss function that has a higher priority for these examples. The last loss, L, can be described by

$$\begin{cases} L = \dfrac{\sum_{k=1}^{10000} l_k}{10000} & M \leq 10000, \\ L = \dfrac{\sum_{k=1}^{M} l_k}{M} & M > 10000, \end{cases} \tag{6}$$

where l_k is the new sorted sequence and M is the total number satisfying $l_k >$ threshold, respectively. The default threshold value is 0.7. In summary, the total loss, L, is calculated with the values of l_k and M.



(a)



(b)

Figure 5. Details of the two attention modules: (**a**) the channel attention module; and (**b**) the spatial attention module.

4. Experiments and Results

4.1. Dataset

The NEU [45] surface defect dataset contains 1800 grayscale images of hot-rolled steel strips, each with a resolution of 200 × 200 pixels. Six common steel surface defects are presented in the dataset, namely rolled-in scale, plaque, crack, pitting surface, inclusions, and scratches. Unfortunately, the bounding box annotations provided in the original dataset are not applicable for the task of semantic segmentation. To address this issue, the NEU-seg dataset was created by [45], which provides pixel-level annotations for three typical defects (inclusions, patches, and scratches) using the labeling tool LabelMe. The NEU-seg dataset consists of 3630 training images and 840 testing images.

In a real-world production environment, it is often challenging to acquire plenty of diverse defect images and their corresponding labels due to operational and equipment

constraints. To address the over-fitting issue caused by limited training data, we employ a data augmentation operator during the training process. To align with the network's training requirements, we transform each original image to 256 × 256 pixels, and then augment the processed image by rotating it by 90°, 180°, and 270°. At the same time, we use the same augmentation to process the ground truth.

4.2. Evaluation Metrics

The evaluation of the segmentation performance on the NEU-seg dataset involves the use of pixel-based metrics. These metrics are primarily used to assess the integrity of segmented objects. To assess the segmentation performance of our network, we use various metrics, including the mean intersection over union (mIoU), Dice coefficient, mean average precision (mAP), and recall, where mIoU is considered as the primary indicator.

The mIoU and Dice coefficient are employed as evaluation metrics to quantify the degree of similarity between predicted segmentations and their respective ground truth labels, while precision and recall are used to represent the number of targets correctly identified based on the prediction results and the ground truth, respectively. These metrics are described by

$$\text{mIoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}, \tag{7}$$

$$\text{Dice} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}, \tag{8}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{9}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{10}$$

where TP, FN, and FP represent the number of pixels that are correctly identified as defective, the number of unidentified defective pixels, and the number of false positive pixels identified as defective, respectively.

4.3. Implementation Details

We implemented our model using the Baidu PaddlePaddle [46] library with CUDA 11.6 and cuDNN 8.5. We ran our model on an NVIDIA GeForce RTX 3060 GPU (with 12 GB memory) and Ubuntu 20.04.

During the training phase, we chose the SGD [47] optimizer with momentum of 0.9 and a weight decay of 4.0×10^{-5}. We adopted the warm-up strategy and the poly-learning rate scheduler to optimize the training process, the warm-up initial learning rate was set to 1.0×10^{-4} and increased to 1.0×10^{-2} at 1200 iteration steps, and then, the learning rate was decreased by 0.9 at every epoch until it reached 1.0×10^{-4}. The training process converged after around 110 epochs with a batch size of 8. When training and testing, we resized each image to $3 \times 256 \times 256$ as the input size.

4.4. Comparative Experiments with State-of-the-Art Methods

The proposed AFFNet was compared with some classical and state-of-the-art segmentation networks, such as UNet, UNet++ [48], Deeplabv3+, FCN-hrnetw18 [49], PSPNet, HardNet [50], and EMANet [51]. All of the comparative experiments were conducted on the same machine and environment. Additionally, these methods were all run on the same training set using their default settings. The comparison results are shown in Table 1, and Figure 6 shows comparison results of different indicators with other models.

As can be seen from Table 1, the results indicate that our proposed AFFNet outperforms other methods in terms of all of the four metrics. The bold text in the table indicates that the value is the best indicator. Specifically, compared with FCN-hrnetw18, our pro-

posed AFFNet showed increases of 1.12%, 0.86%, 0.15%, and 1.73% in the mIoU, Dice coefficient, mAP, and recall, respectively, on the test dataset. Meanwhile, compared with FC-HardNet70, our method showed increases of 0.89%, 0.70%, 0.11%, and 0.42% in the mIoU, Dice coefficient, mAP, and recall, respectively, on the test dataset.

Table 1. Comparison results of the surface defects segmentation performance of different networks on the NEU-seg dataset.

Method	mIoU (%)	Dice (%)	mAP	Recall
FCN-8s [5]	77.25	87.07	0.9707	0.8758
FCN-hrnetw18 [49]	79.82	88.70	0.9748	0.8712
Deeplabv3+ [11]	78.76	88.02	0.9736	0.8739
UNet [7]	77.40	87.17	0.9700	0.8786
UNet++ [48]	77.76	87.41	0.9714	0.8669
PSPNet [9]	78.68	87.41	0.9738	0.8603
EMANet [51]	78.56	87.25	0.9733	0.8544
FC-HardNet70 [50]	80.05	88.86	0.9747	0.8853
AFFNet (ours)	**80.94**	**89.56**	**0.9758**	**0.8895**

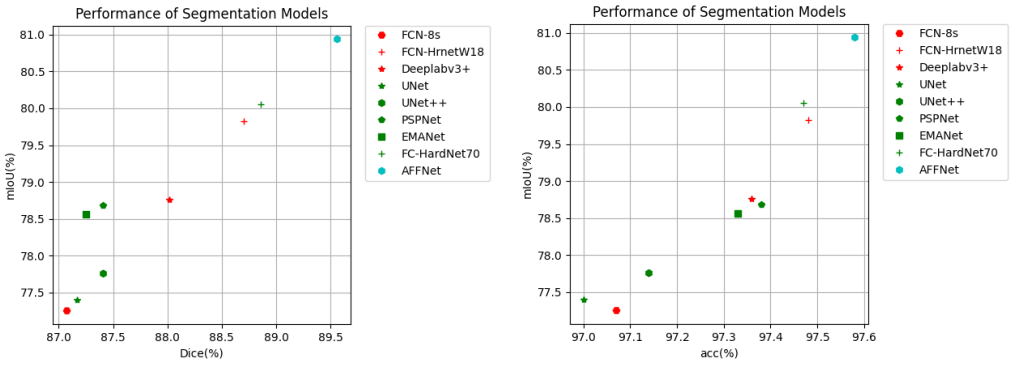


Figure 6. Comparison results of indicators with other models.

Several visualized segmentation results between our method and others for detecting surface defects on the NEU-seg dataset are shown in Figure 7. As can be seen from this figure, our algorithm can accurately predict the types of defect and the segmentation predictions are consistent with the ground truth labels. Specifically, the first two rows of the figure indicate that our method can more effectively suppress background interference and that it reduces false positives in non-defective areas. The last row shows that our proposed method can accurately identify defects of similar texture in different categories, even in small sizes. This capability is particularly valuable for analyzing the location and root causes of defects. In summary, our proposed method accurately segments defect areas and identifies the corresponding defect types, even in challenging scenarios such as low-contrast regions and intra-class differences.

To provide a comprehensive assessment of the computational complexity of our proposed method, we calculate the model parameters, the computational complexity in terms of FLOPs, and the inference time of the network.

The experimental results, which are shown in Table 2, include the number of trainable parameters, computational complexity (FLOPs), and inference time. It can be seen that AFFNet has a total of 11.42 MB trainable parameters and 25.16 GB FLOPs. Furthermore, the inference time per image of our model is 20.89 ms. These results imply that AFFNet has a superior segmentation performance and efficient inference speed.

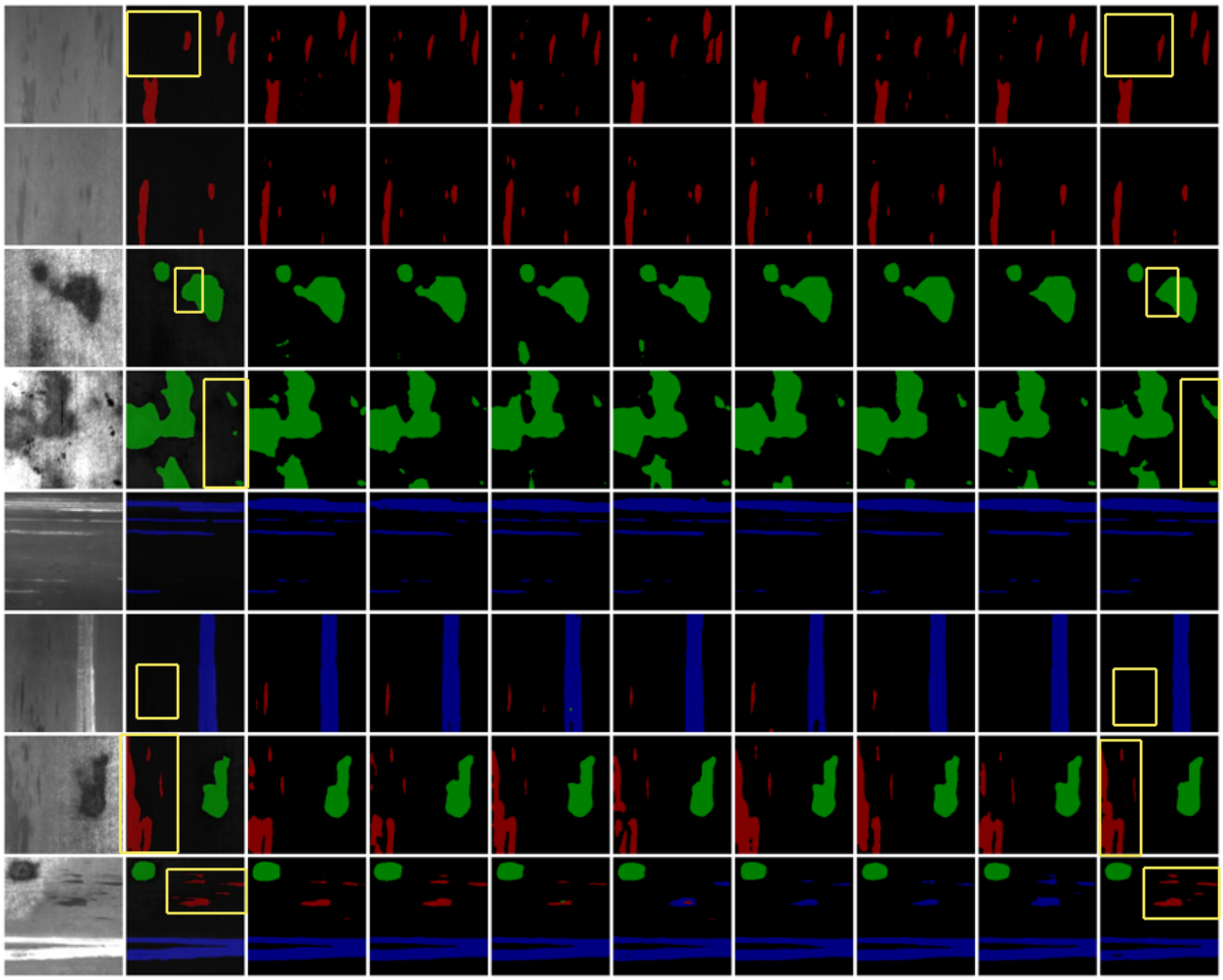


Figure 7. The visualized segmentation results of different networks on the test dataset. Red, green, and blue areas refer to inclusion, patches, and scratches defects, respectively. The yellow boxed areas indicate the consistency between our method's segmentation results and the ground truth. (**a**) Original image, (**b**) ground truth, (**c**) UNet, (**d**) UNet++, (**e**) FCN-hrnetw18, (**f**) Deeplabv3+, (**g**) PSPNet, (**h**) EMANet, (**i**) FC-HardNet70, and (**j**) AFFNet (ours).

Table 2. Comparison results in terms of the trainable parameters, FLOPs, and test time of different networks on the NEU-seg dataset.

Method	Params (MB)	FLOPs (GB)	Inference Time (ms/per Image)
FCN-8s [5]	18.6	25.5	**10.6**
FCN-hrnetw18 [49]	9.67	4.64	50.79
Deeplabv3+ [11]	26.79	28.55	18.03
UNet [7]	13.46	31.15	13.17
UNet++ [48]	8.37	30.06	16.46
PSPNet [9]	67.9	66.43	25.68
EMANet [51]	42.41	44.57	19.27
FC-HardNet70 [50]	**4.12**	**4.41**	19.7
AFFNet (ours)	11.42	25.16	20.89

4.5. Ablation Study

To provide evidence of the efficacy of the modules we propose in improving network performance, we conducted a series of ablation experiments, including assessing the effectiveness of the RRD module, the GFA module, and the OHEM training strategy. We set the UNet as the baseline. These ablation experiments were conducted on the NEU-seg dataset, and the corresponding performances are shown in Table 3. Meanwhile, Figure 8 shows some sample images and their corresponding predictions with different modules in the ablation experiments.

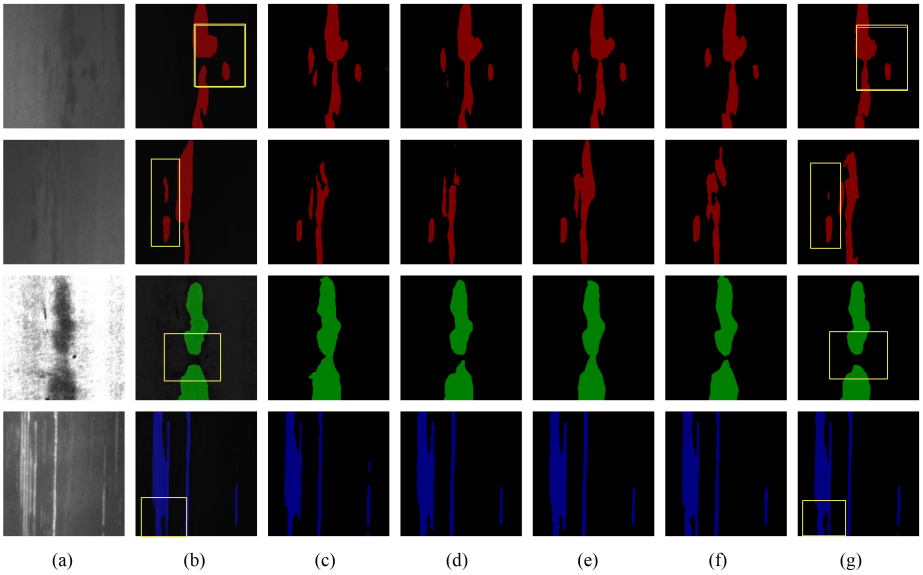


Figure 8. Comparison of ablation results on the NEU-seg dataset. The yellow boxed areas indicate the consistency between our method's segmentation results and the ground truth. (**a**) Original images, (**b**) ground truth, (**c**) UNet, (**d**) UNet+RRD, (**e**) UNet+GFA, (**f**) UNet+RRD+GFA, and (**g**) AFFNet (ours).

Table 3. The effects of different modules, including RRD, GFA, and OHEM strategy on network performance.

Method	mIoU (%)	Gain
UNet	77.40	-
UNet+RRD	79.93	+2.53
UNet+GFA	78.68	+1.28
UNet+RRD+GFA	80.54	+3.14
Ours (+OHEM)	**80.94**	+3.54

4.5.1. Effectiveness of the RRD Module

In the encoder module, we employ the Residual-RepGhost-Dblock as our feature extraction block. This approach leverages the RepGhost bottleneck module to reduce feature dimensionality while retaining feature information. Additionally, we utilized convolution blocks with varying expansion coefficients to extract features with receptive fields of different scales. The results, presented in Table 3, demonstrate a performance improvement from 77.40% to 79.93% in the mIoU indicator, which indicates that incorporating the RRD module has led to a significant improvement in the network's performance.

4.5.2. Effectiveness of the GFA Module

We have also incorporated a global feature attention (GFA) module into our proposed method to enhance the performance of the results. As presented in Table 3, combined with the GFA module, the mIoU has increased by 1.28%. This demonstrates the effectiveness

of GFA in enhancing the ability of the network to extract meaningful features for defect detection, resulting in improved accuracy and efficiency of industrial defect detection.

4.5.3. The Ablation Studies for the OHEM Strategy

To improve the convergence performance of our method, we incorporated the OHEM training strategy into our method. Table 3 shows the results, indicating that the mean intersection over union (mIoU) score of our method increased by 0.4%, demonstrating its effectiveness.

5. Conclusions

In this paper, a novel neural network framework for surface defect segmentation of industrial products is proposed. For the encoder part, the framework leverages a lightweight feature extraction module incorporating an attention module focusing more on representative features for defect segmentation. Additionally, during the decoding process, a global feature attention module is utilized to select and fuse high-level features and low-level ones. The experimental results show that the proposed framework outperforms the other state-of-the-art methods in segmenting surface defects of industrial products. Despite the encouraging results we have achieved on surface defect segmentation, there are still some challenges that need to be addressed in future work. For instance, although a data augmentation strategy is adopted in our method, it still suffers from the problem of insufficient labeled data. To solve this, we need to collaborate with other research teams and industrial organizations to expand the training dataset. Furthermore, we can use generative adversarial networks to generate plausible new samples to expand the dataset. In addition, the lightweight nature of the network is also important in practical applications of surface defect segmentation, so it is necessary to further reduce the number of computations and parameters of the segmentation network to make the model deployable on mobile and other portable devices.

Author Contributions: Conceptualization, C.F. and C.-W.S.; Methodology, X.C. and M.T.; Software, X.C. and H.M.; Validation, M.T. and H.M.; Formal analysis, M.T. and C.-W.S.; Data curation, H.M.; Writing—original draft, X.C.; Writing—review & editing, C.F. and C.-W.S.; Supervision, C.F.; Funding acquisition, C.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the National Natural Science Foundation of China (No. 62032013), and the Fundamental Research Funds for the Central Universities (No. N2224001-7).

Data Availability Statement: The data that support the findings of this study are openly available at https://github.com/DHW-Master/NEU_Seg (acessed on 17 May 2023).

Conflicts of Interest: The authors declare that they have no conflict of interest.

References

1. Zhang, H.; Jin, X.; Wu, Q.J.; Wang, Y.; He, Z.; Yang, Y. Automatic visual detection system of railway surface defects with curvature filter and improved Gaussian mixture model. *IEEE Trans. Instrum. Meas.* **2018**, *67*, 1593–1608. [CrossRef]
2. Wang, J.; Li, Q.; Gan, J.; Yu, H.; Yang, X. Surface defect detection via entity sparsity pursuit with intrinsic priors. *IEEE Trans. Ind. Inform.* **2019**, *16*, 141–150. [CrossRef]
3. Luo, Q.; Sun, Y.; Li, P.; Simpson, O.; Tian, L.; He, Y. Generalized completed local binary patterns for time-efficient steel surface defect classification. *IEEE Trans. Instrum. Meas.* **2018**, *68*, 667–679. [CrossRef]
4. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
5. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
6. Chen, F.C.; Jahanshahi, M.R. NB-FCN: Real-time accurate crack detection in inspection videos using deep fully convolutional network and parametric data fusion. *IEEE Trans. Instrum. Meas.* **2019**, *69*, 5325–5334. [CrossRef]
7. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
8. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]

9. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
10. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
11. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
12. Xu, J.; Lu, K.; Wang, H. Attention fusion network for multi-spectral semantic segmentation. *Pattern Recognit. Lett.* **2021**, *146*, 179–184. [CrossRef]
13. Yan, L.; Cui, Y.; Chen, Y.; Liu, D. Hierarchical attention fusion for geo-localization. In Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 6–12 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 2220–2224.
14. Wang, J.; Huang, R.; Guo, S.; Li, L.; Zhu, M.; Yang, S.; Jiao, L. NAS-Guided Lightweight Multiscale Attention Fusion Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 8754–8767. [CrossRef]
15. Liang, Y.; Qin, G.; Sun, M.; Yan, J.; Jiang, H. MAFNet: Multi-style attention fusion network for salient object detection. *Neurocomputing* **2021**, *422*, 22–33. [CrossRef]
16. Chu, M.; Gong, R. Invariant feature extraction method based on smoothed local binary pattern for strip steel surface defect. *ISIJ Int.* **2015**, *55*, 1956–1962. [CrossRef]
17. Truong, M.T.N.; Kim, S. Automatic image thresholding using Otsu's method and entropy weighting scheme for surface defect detection. *Soft Comput.* **2018**, *22*, 4197–4203. [CrossRef]
18. Su, B.; Chen, H.; Zhu, Y.; Liu, W.; Liu, K. Classification of manufacturing defects in multicrystalline solar cells with novel feature descriptor. *IEEE Trans. Instrum. Meas.* **2019**, *68*, 4675–4688. [CrossRef]
19. Luo, Q.; Fang, X.; Sun, Y.; Liu, L.; Ai, J.; Yang, C.; Simpson, O. Surface defect classification for hot-rolled steel strips by selectively dominant local binary patterns. *IEEE Access* **2019**, *7*, 23488–23499. [CrossRef]
20. Zhao, J.; Peng, Y.; Yan, Y. Steel Surface Defect Classification Based on Discriminant Manifold Regularized Local Descriptor. *IEEE Access* **2018**, *6*, 71719–71731. [CrossRef]
21. Liu, Y.; Xu, K.; Xu, J. An Improved MB-LBP Defect Recognition Approach for the Surface of Steel Plates. *Appl. Sci.* **2019**, *9*, 4222. [CrossRef]
22. Navarro, P.J.; Fernández-Isla, C.; Alcover, P.M.; Suardíaz, J. Defect detection in textures through the use of entropy as a means for automatically selecting the wavelet decomposition level. *Sensors* **2016**, *16*, 1178. [CrossRef]
23. Sharma, P.; Said, Z.; Kumar, A.; Nizetic, S.; Pandey, A.; Hoang, A.T.; Huang, Z.; Afzal, A.; Li, C.; Le, A.T.; et al. Recent advances in machine learning research for nanofluid-based heat transfer in renewable energy system. *Energy Fuels* **2022**, *36*, 6626–6658. [CrossRef]
24. Sharma, P.; Bora, B.J. A Review of Modern Machine Learning Techniques in the Prediction of Remaining Useful Life of Lithium-Ion Batteries. *Batteries* **2023**, *9*, 13. [CrossRef]
25. Liu, Z.; Fang, L.; Jiang, D.; Qu, R. A machine-learning-based fault diagnosis method with adaptive secondary sampling for multiphase drive systems. *IEEE Trans. Power Electron.* **2022**, *37*, 8767–8772. [CrossRef]
26. Ren, M.; Wang, X.; Xiao, G.; Chen, M.; Fu, L. Fast defect inspection based on data-driven photometric stereo. *IEEE Trans. Instrum. Meas.* **2018**, *68*, 1148–1156. [CrossRef]
27. Li, D.; Li, Y.; Xie, Q.; Wu, Y.; Yu, Z.; Wang, J. Tiny defect detection in high-resolution aero-engine blade images via a coarse-to-fine framework. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 3512712. [CrossRef]
28. Ju, Y.; Jian, M.; Guo, S.; Wang, Y.; Zhou, H.; Dong, J. Incorporating lambertian priors into surface normals measurement. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 5012913. [CrossRef]
29. Ren, M.; Xiao, G.; Zhu, L.; Zeng, W.; Whitehouse, D. Model-driven photometric stereo for in-process inspection of non-diffuse curved surfaces. *CIRP Ann.* **2019**, *68*, 563–566. [CrossRef]
30. Lu, P.; Jing, J.; Huang, Y. MRD-net: An effective CNN-based segmentation network for surface defect detection. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 2516812. [CrossRef]
31. Zhang, J.; Ding, R.; Ban, M.; Guo, T. FDSNeT: An Accurate Real-Time Surface Defect Segmentation Network. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 3803–3807.
32. Tian, S.; Huang, P.; Ma, H.; Wang, J.; Zhou, X.; Zhang, S.; Zhou, J.; Huang, R.; Li, Y. CASDD: Automatic surface defect detection using a complementary adversarial network. *IEEE Sens. J.* **2022**, *22*, 19583–19595. [CrossRef]
33. Zhou, X.; Fang, H.; Fei, X.; Shi, R.; Zhang, J. Edge-Aware Multi-Level Interactive Network for Salient Object Detection of Strip Steel Surface Defects. *IEEE Access* **2021**, *9*, 149465–149476. [CrossRef]
34. Li, Z.; Wu, C.; Han, Q.; Hou, M.; Chen, G.; Weng, T. CASI-Net: A Novel and Effect Steel Surface Defect Classification Method Based on Coordinate Attention and Self-Interaction Mechanism. *Mathematics* **2022**, *10*, 963. [CrossRef]
35. Tang, Z.; Tian, E.; Wang, Y.; Wang, L.; Yang, T. Nondestructive defect detection in castings by using spatial attention bilinear convolutional neural network. *IEEE Trans. Ind. Inform.* **2020**, *17*, 82–89. [CrossRef]
36. Pan, Y.; Zhang, L. Dual attention deep learning network for automatic steel surface defect segmentation. *Comput.-Aided Civ. Infrastruct. Eng.* **2022**, *37*, 1468–1487. [CrossRef]

37. Wu, Y.; Qin, Y.; Qian, Y.; Guo, F.; Wang, Z.; Jia, L. Hybrid deep learning architecture for rail surface segmentation and surface defect detection. *Comput.-Aided Civ. Infrastruct. Eng.* **2022**, *37*, 227–244. [CrossRef]
38. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
39. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
40. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
41. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
42. Chen, C.; Guo, Z.; Zeng, H.; Xiong, P.; Dong, J. RepGhost: A Hardware-Efficient Ghost Module via Re-parameterization. *arXiv* **2022**, arXiv:2211.06088.
43. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of theIEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
44. Zhang, Z.; Sabuncu, M. Generalized cross entropy loss for training deep neural networks with noisy labels. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montreal, Canada, 3–8 December 2018; pp. 8792–8802.
45. Bao, Y.; Song, K.; Liu, J.; Wang, Y.; Yan, Y.; Yu, H.; Li, X. Triplet-graph reasoning network for few-shot metal generic surface defect segmentation. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 5011111. [CrossRef]
46. Ma, Y.; Yu, D.; Wu, T.; Wang, H. PaddlePaddle: An Open-Source Deep Learning Platform from Industrial Practice. *Front. Data Comput.* **2019**, *1*, 105. [CrossRef]
47. Bottou, L. Large-scale machine learning with stochastic gradient descent. In Proceedings of the COMPSTAT'2010: 19th International Conference on Computational Statistics, Paris, France, 22–27 August 2010; Keynote, Invited and Contributed Papers 2010; Physica-Verlag HD: Heidelberg, Germany, 2010.
48. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Proceedings of the 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, 20 September 2018*; Proceedings 4; Springer: Berlin/Heidelberg, Germany, 2018; pp. 3–11.
49. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3349–3364. [CrossRef] [PubMed]
50. Chao, P.; Kao, C.Y.; Ruan, Y.S.; Huang, C.H.; Lin, Y.L. Hardnet: A low memory traffic network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3552–3561.
51. Li, X.; Zhong, Z.; Wu, J.; Yang, Y.; Lin, Z.; Liu, H. Expectation-maximization attention networks for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9167–9176.