mathematics

MDPI

Article

CASI-Net: A Novel and Effect Steel Surface Defect Classification Method Based on Coordinate Attention and Self-Interaction Mechanism

Zhong Li [1], Chen Wu [1,*], Qi Han [1], Mingyang Hou [1], Guorong Chen [1] and Tengfei Weng [2]

[1] College of Intelligent Technology and Engineering, Chongqing University of Science and Technology, Chongqing 401331, China; 2006055@cqust.edu.cn (Z.L.); hanqicq@163.com (Q.H.); hmy394481125@163.com (M.H.); cgr@cqust.edu.cn (G.C.)
[2] College of Electrical Engineering, Chongqing University of Science and Technology, Chongqing 401331, China; wengtf_cq@163.com
* Correspondence: cwcqust@163.com

Abstract: The surface defects of a hot-rolled strip will adversely affect the appearance and quality of industrial products. Therefore, the timely identification of hot-rolled strip surface defects is of great significance. In order to improve the efficiency and accuracy of surface defect detection, a lightweight network based on coordinate attention and self-interaction (CASI-Net), which integrates channel domain, spatial information, and a self-interaction module, is proposed to automatically identify six kinds of hot-rolled steel strip surface defects. In this paper, we use coordinate attention to embed location information into channel attention, which enables the CASI-Net to locate the region of defects more accurately, thus contributing to better recognition and classification. In addition, features are converted into aggregation features from the horizontal and vertical direction attention. Furthermore, a self-interaction module is proposed to interactively fuse the extracted feature information to improve the classification accuracy. The experimental results show that CASI-Net can achieve accurate defect classification with reduced parameters and computation.

Keywords: hot-rolled steel strip; defect classification; convolutional neural network; attention mechanism; visual interaction mechanism

MSC: 68T01; 68T07

Citation: 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. https://doi.org/10.3390/math10060963

Academic Editor: Andrea Prati

Received: 23 February 2022
Accepted: 15 March 2022
Published: 17 March 2022

1. Introduction

As the most important product in iron and steel enterprises, the steel strip has become an irreplaceable raw material in automobile manufacturing, aerospace, mechanical processing, and other fields [1]. However, in the actual production process of the hot-rolled strip, due to the imperfect manufacturing process, the surface of the strip usually contains different types of defects, such as scratches, surface cracks, and rolling marks [2–4]. These defects not only affect the appearance of the product but also reduce the quality of the finished product [1,2]. Traditionally, the classification of steel surface defects is checked manually by experts [2–4].

However, the manual detection process is subjective, fatigued, and the work speed is slow, which is not conducive to the completion of real-time detection tasks [4]. Therefore, in order to improve the recognition efficiency and accuracy, it is essential to develop an accurate automatic detection solution. In the past decades, machine vision technology as a safe, non-contact, and automatic solution has been widely used in material surface detection [5]. Machine vision detection is mainly composed of image acquisition and defect detection [6]. With the increasingly complex industrial environment, machine vision detection technology faces many challenges, such as low universality of equipment, high

requirements for the light-source environment, and expensive costs of production and maintenance of the machine vision detection device [7]. In this case, machine vision detection technology is often inefficient and makes it difficult to achieve better detection results. In order to overcome the shortcomings of machine vision, researchers considered that deep learning has good performance compared with traditional machine vision. They applied deep learning to defect detection and achieved great improvement.

In recent years, due to the outstanding performance of deep learning in comparison to machine vision, deep learning has developed rapidly in computer vision applications [8–12]. Deep learning can solve the problem whereby different tasks need different image-processing algorithms in traditional machine vision. AlexNet [8] was proposed in 2012 and made a huge impact on the development of deep learning. Compared with traditional machine learning methods, deep learning uses convolution, pooling, and other operations for feature extraction to obtain the abstract feature information of the image. The convolution neural network (CNN) uses convolution operation for feature extraction of input images, which can learn local features and capture different degrees of semantic information so as to effectively learn feature expression from a large number of samples, and the model has a stronger generalization ability. Compared with traditional machine vision methods [7], CNN adopts a pooling layer and sparse connection to reduce model parameters while ensuring the efficiency of computing resources and network performance [13,14]. Deep learning combines the full connection layer to achieve high-precision detection and classification, which promotes the further development of deep learning in the field of image processing. Therefore, in order to achieve better classification accuracy, a deeper learning architecture is needed. However, deeper learning architectures [15,16] contain a large number of parameters and require a large amount of computation load.

In order to overcome the above problems, we propose a lightweight convolutional neural network called CASI-Net, which combines channel attention, location information, and a self-interaction module based on the biological vision to achieve fast and accurate classification of steel surface defects. In the feature extraction stage, inspired by [17], we use 3×1 convolution kernels and 1×3 convolution kernels to replace 3×3 convolution kernels, aiming at reducing network parameters. Then, in order to help CASI-Net more accurately locate and identify the region of interest, a coordinate attention (CA) block [18] is introduced and a self-interaction module based on biological vision is constructed. The self-interaction module can improve the richness of the extracted features. CASI-Net is compared with the typical surface defect identification methods and can use a small number of parameters to achieve more accurate identification results. Overall, the contributions of this paper are summarized as follows:

- An end-to-end CASI-Net model is proposed, which combines location information and channel attention to locate defects more accurately. In addition, we construct a self-interaction module based on the biological visual interaction mechanism to learn more detailed feature information. Finally, CASI-Net can use very few parameters to achieve accurate classification.
- We introduce the CA block to CASI-Net. The CA block can not only capture cross-channel information but also capture location information, which can help CASI-Net to locate and identify targets of interest more accurately.
- The self-interaction module based on biological mechanisms is constructed to enrich the representation of feature maps, which is helpful for better recognition and classification.
- To evaluate the performance of the CASI-Net for real industrial data, we use the NEU dataset provided by Northeastern University to validate the performance of CASI-Net. The classification results on NEU will verify the effectiveness of our proposed network.

The remainder of our paper is organized as follows. Section 2 introduces the related work, and two improved techniques are introduced in Section 3. Section 4 provides an evaluation of our method and experimental by comparison with state-of-the-art methods. In Section 5, the conclusion is provided.

2. Related Work

2.1. Convolutional Neural Networks

CNN was proposed as early as 1989 [19]. Yann et al. proposed the first classical architecture LeNet-5 [20] of CNN in 1998. LeNet-5 contains six hidden layers, mainly by convolution and pooling operations stacking to extract image features, which can achieve good results on the MNIST dataset [20]. Krizhevsky proposed AlexNet in 2012, which has five convolution layers and three fully connected layers, except the pooling layer [8]. After that, SimonYan and Zisserman proposed to use 3×3 filters to construct a deeper network, VGG [15], which promoted the development of computer vision tasks in 2015. However, when only using the convolution layer stacking method, when the depth reaches a certain degree, it will not improve the effect but will rather deteriorate the effect. Therefore, He et al. designed the residual learning method and proposed ResNet [16] to solve the degradation problem of the deep network and realized a significant improvement in the deep network. However, the above studies placed too much emphasis on deepening the network depth to improve accuracy, without considering the calculation of the model. In order to achieve fewer parameters and lower degradation of the network performance, Iandola et al. proposed architecture to generate high-precision identification with significantly fewer parameters, called SqueezeNet [21]. Later, researchers proposed other representative lightweight networks such as ShuffleNet [22], MobileNet [23], and MobileNet V2 [24].

2.2. Attention Mechanisms

In recent years, the attention mechanism [25] has been widely used in various computer vision tasks, such as image classification [26–29] and image segmentation [30,31]. One successful example is SE [26], which squeezes each two-dimensional feature map to efficiently construct the interdependence between channels. However, SE [26] only considered channel information, ignoring the importance of location information. However, the spatial information of the object is also important in computer vision [28]. BAM [32] and CBAM [28] tried to use the channel domain and spatial domain for feature extraction, but BAM [32] and CBAM [28] only captured local information and could not obtain long-term dependence [15]. In order to solve the above problems, we introduce the CA block [15] into CASI-Net. In the coordinate attention [18], the channel attention is decomposed into two one-dimension feature coding processes, in which information for different directions is aggregated. Different from SE [26], the CA block [18] can not only capture the correlation dependence of feature maps but also retain accurate location information along the spatial direction.

2.3. Biological Visual Interaction Mechanism

The interaction mechanism of biological vision refers to that in visual information processing, where visual information interacts to a certain extent and, finally, completes the storage and recovery of back flow and abdominal flow [33–36]. In addition, when visual information is transmitted in the dorsal or ventral stream, the self-interaction behavior will be triggered [35]. This form of feature interaction can enrich the information of features and enhance the expression ability of information in the cerebral cortex [37,38]. For deep learning, when the feature map contains less effective information than the original map, the classification accuracy of the deep neural network is not excellent. On the contrary, when the feature map contains more effective information than the original map, the representation of the feature map can be expanded, thereby enhancing the expression ability of the CNN model. Based on the above research, a self-interaction module is constructed, inspired by the biological visual interaction mechanism, which enables CASI-Net to obtain more abundant original image information and enhance the characterization of CASI-Net features, thus further improving the classification accuracy of CASI-Net.

3. Proposed Method

The proposed CASI-Net architecture consists of a lightweight basic feature extractor (BLFE), a CA block [18], and a self-interaction module. The CASI-Net architecture is shown in Figure 1.

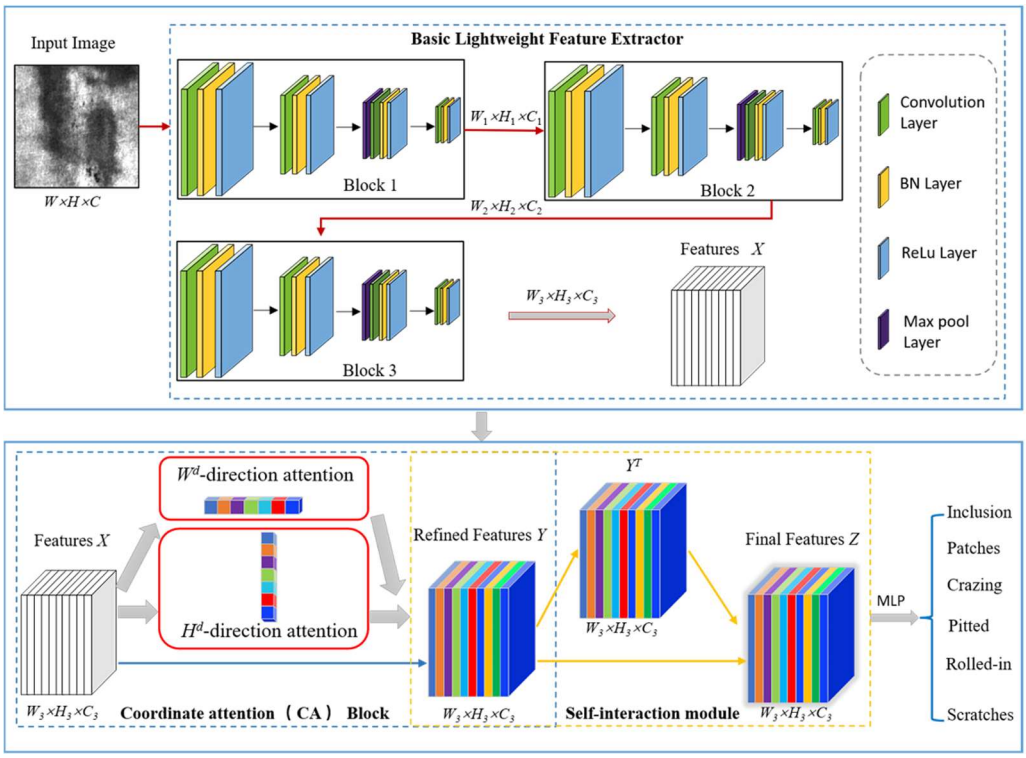


Figure 1. Overview of the CASI-Net framework.

In CASI-Net, the input image is a $W \times H \times C$ steel surface image with defects, and the output of CASI-Net is defect category confidence. Here, W, H, and C denote the width, height, and channel numbers of the input image, respectively. In the basic lightweight feature extractor, the output dimension of block i is $W_i \times H_i \times C_i$. Here, W_i, H_i, and C_i (i = 1, 2, 3) denote the width, height, and channel of the output of feature maps in block i, respectively. In order to ensure CASI-Net focuses on the defect area, we introduce the CA block [18] into our constructed network to obtain refined feature maps through the attention of W^d and H^d directions. Then we construct a self-interaction module based on the interaction mechanism of biological vision to enrich the feature maps information and enhance the characterization of CASI-Net. Finally, CASI-Net connects to the Multilayer Perceptron (MLP) to obtain the category of an input defect image.

3.1. Basic Lightweight Feature Extractor

BLFE consists of three depth-wise separable convolution modules shown in Figure 1. Each module consists of four convolution layers, four ReLu layers, four batch normalization (BN) layers, and a Max pool layer, which is shown in Figure 2. A convolution layer is the basis of the image feature extraction process, while the core is the convolution operation. The convolution layer at the lowest level extracts low-level features such as edges and lines, and the higher convolution layer extracts the more complex features such as object color and contour. The BN [39] layer can speed up the training process and greatly solve the problem of gradient disappearance and improve the performance of the CNN [40]. Max pooling is used to reduce the dimension of features, compress the number of data and parameters, and effectively reduce the overfitting phenomenon [40]. A ReLU [41] as a nonlinear activation layer can aptly solve the overfitting problem.

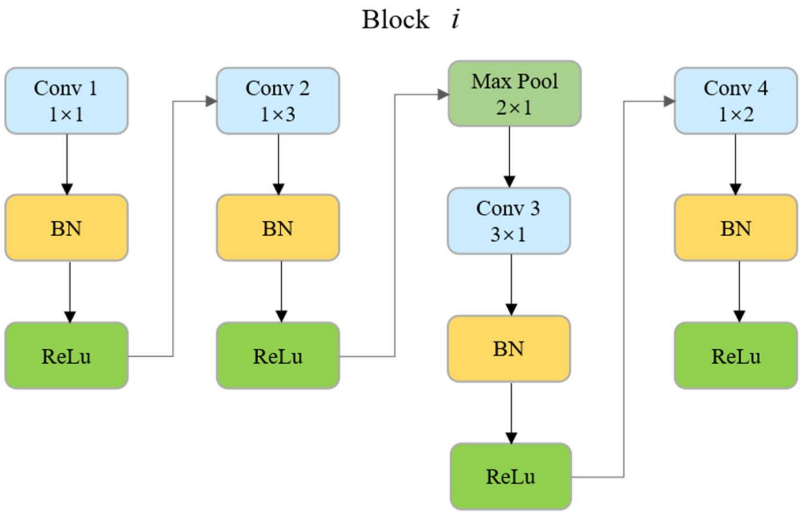


Figure 2. Depth-wise separable convolution block.

In block i, the traditional 3×3 deep convolution is decomposed into 1×3 convolution kernel Conv2 and 3×1 convolution kernel Conv3, and finally, the feature map X is obtained by 1×2 convolution kernel Conv4.

3.2. Coordinate Attention

In order to focus on the defect areas and suppress the unimportant areas to achieve more accurate identification, CASI-Net combines the channel attention mechanism and the location information to obtain more accurate defect areas. Attention modules such as SE [26] and CBAM [28] can improve network performance in image classification. Traditional attention modules such as SE [26] only considered the channel information of the image and ignored the spatial information. In addition, SE [26] lost too much primitive information via global pooling. To solve these problems, we integrate the CA block [18] into CASI-Net to improve the accuracy of classification. In the CA block, feature tensors $X = [x_1, x_2 \ldots, x_n]$ are obtained after BLFE as the input. Finally, CA outputs the re-weighted tensor $Y = [y_1, y_2 \ldots, y_n]$ [18]. The architecture of the CA block is shown in Figure 3, where 'W^d Avg Pool' and 'H^d Avg Pool' refer to 1D W^d Avg pooling and 1D 'H^d Avg pooling', respectively [18].

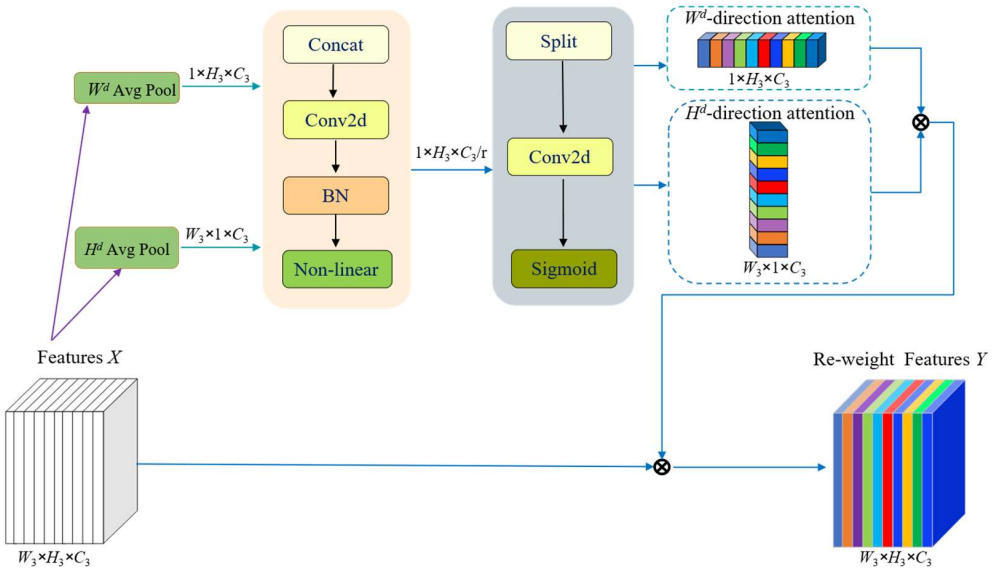


Figure 3. The architecture of coordinate attention.

For the input feature tensor $X = [x_1, x_2 \ldots, x_n]$, one-dimensional pooling operations in the first step generate feature descriptors in W^d and H^d directions in the CA block. Specifically, CA uses two different pooling kernels to encode the channel along the W^d direction and the H^d direction. The two different pooling kernels size are $(H_3, 1)$ and $(1, W_3)$, respectively [18]. Then the output of channel c, $c \in \{1, 2, \ldots, n\}$ at height h, $h \in \{1, 2, \ldots, H\}$ is expressed as follows

$$z_c^h(h) = \frac{1}{W_3} \sum_{0 \le i < W_3} x_c(h, i)$$

where $z_c^h(h)$ is the W^d directional awareness of x_c. x_c is the c-channel feature map of the feature tensor X. The output of channel $c, c \in \{1, 2, \ldots, n\}$ with width $w, w \in \{1, 2, \ldots, W\}$ is expressed as follows

$$z_c^w(w) = \frac{1}{H_3} \sum_{0 \le j < H_3} x_c(j, w)$$

where $z_c^w(w)$ is the H^d directional awareness of x_c. x_c is the c-channel feature map of the feature tensor X. $z_c^h(h)$ and $z_c^w(w)$ combine two different locations' information including the W^d direction and H^d direction, which allow CASI-Net to capture long-range dependencies along one spatial direction and preserve precise positional information along the other spatial direction, which helps CASI-Net more accurately locate the region of interest [18].

Next, z^h and z^w are cascaded by the convolution transform and nonlinear activation and obtain the feature maps f [15]. The expression of f is as follows

$$\mathrm{f} = \delta\left(F_1\left(\left[z^h, z^w\right]\right)\right)$$

where f is the feature maps containing W^d and H^d directions, δ is the ReLU function, and F_1 is the 1×1 convolution operation. Next, f is decomposed into two feature tensors f^{h} and f^{w} by the spatial dimension, and then the feature maps are convoluted by two 1×1 convolution layers to form the attention weights g^h and g^w in W^d and H^d directions [18], which are described as follows

$$\mathrm{g}^h = \sigma\left(F_h\left(\mathrm{f}^{\mathrm{h}}\right)\right)$$

$$\mathrm{g}^w = \sigma\left(F_w(\mathrm{f}^{\mathrm{w}})\right)$$

where F_h and F_w are 1×1 convolution operations, and σ is the sigmoid function. Finally, the attention weights in W^d and H^d directions are weighted with the input of CA, and the final output is $y_c(i, j) \in$ Re-weight Features Y as follows

$$y_c(i, j) = x_c(i, j) \times \mathrm{g}_c{}^h(i) \times \mathrm{g}_c{}^w(j)$$

where $\mathrm{g}_c{}^h(i)$ and $\mathrm{g}_c{}^w(j)$ are the attention weights of the c-channel of X in W^d and H^d directions, respectively.

3.3. Self-Interaction Based on Biological Vision

In Section 3.2, through the CA block, CASI-Net obtains an enhancement feature map Y. Then we input the feature maps Y into the self-interaction module to enrich the effective information of feature maps. Inspired by the biological visual interaction mechanism, we design a novel feature augment extraction structure named self-interaction (SI). This interactive mechanism can enrich visual information and extract more discriminative feature information in deep learning, which can improve the results of defect classification. The specific structure of SI is shown in Figure 4.

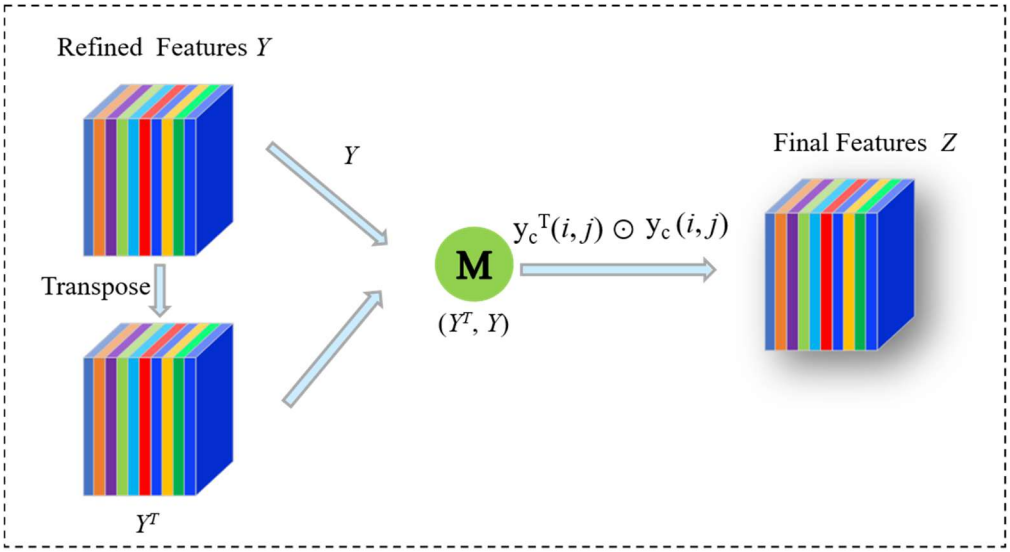


Figure 4. The architecture of self-interaction module structure.

In SI, the output Y of CA is used as the input. After transposing the feature maps Y and obtaining Y^T, the new feature map Z is obtained through interactive operation in the SI module constructed by us. The process of SI is described as follows

$$Z = M\left(Y^T, Y\right) = (y_c^T(i,j) \odot y_c(i,j))$$

where M represents the Hadamard product of Y and Y^T. Z is the final feature maps obtained after the interaction. y_c^T and y_c represent the c-channel feature map of the refined feature tensors Y^T and Y, respectively. (i,j) represent the coordinates of pixels of the feature map. The richness of the deep network in feature information processing is extended by the SI module. Interactive feature maps Z pay more attention to identifying regions and can obtain more detailed feature information in the original feature map, and Z is used for the final classification.

4. Experiments

4.1. Dataset

The dataset used in our work is NEU-CLS, which contains six types of surface defects of a hot-rolled steel strip, which are Crazing (Cr), Inclusion (In), Patches (Pa), Pitted Surface (PS), Rolled-in Scale (RS), and Scratches (Sc) [1]. Each type of sample has 300 grayscale images of which the size is 200 × 200. NEU-CLS has 1800 images. In our experiment, we resize the input images to 300 × 300 × 3 (width, height, channel). Figure 5 shows the samples of six types of typical surface defects images of steel strips.

Each type gives four sample images, and it can be clearly observed that there are great differences in the appearance of the same type of defects. In short, the challenges of the NEU-CLS dataset are the inter-class similarity, intra-class difference, and complex background interference [4].

4.2. Enhanced Dataset

The steel defect dataset is inevitably subjected to non-uniform illumination, noise, and motion blur in the process of industrial acquisition, which poses a certain challenge to defect recognition. In order to evaluate the robustness of CASI-Net, we adapt the enhanced dataset, which includes severe non-uniform illumination, camera noise, and motion blur [2]. The 2 and 5 represent length of camera motion. The enhanced dataset is shown in Figure 6.

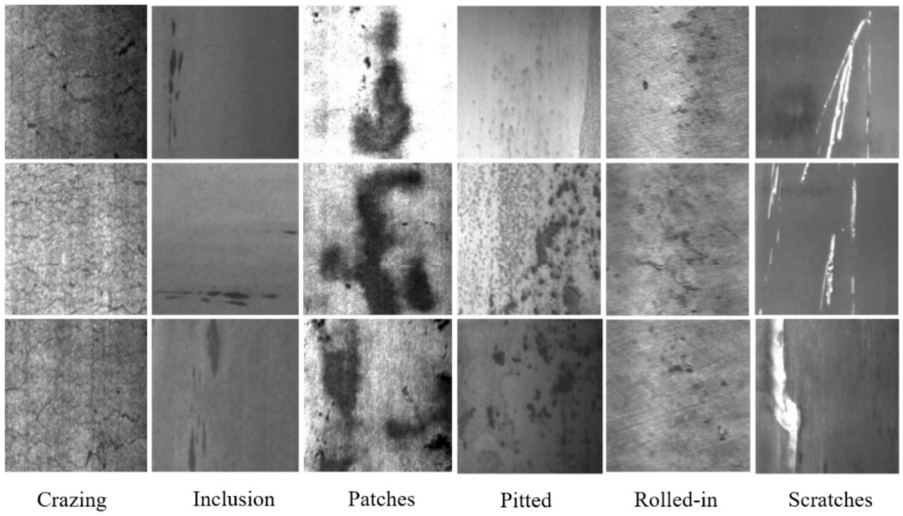


Figure 5. Sample images of six typical surface defects in NEU-CLS dataset.

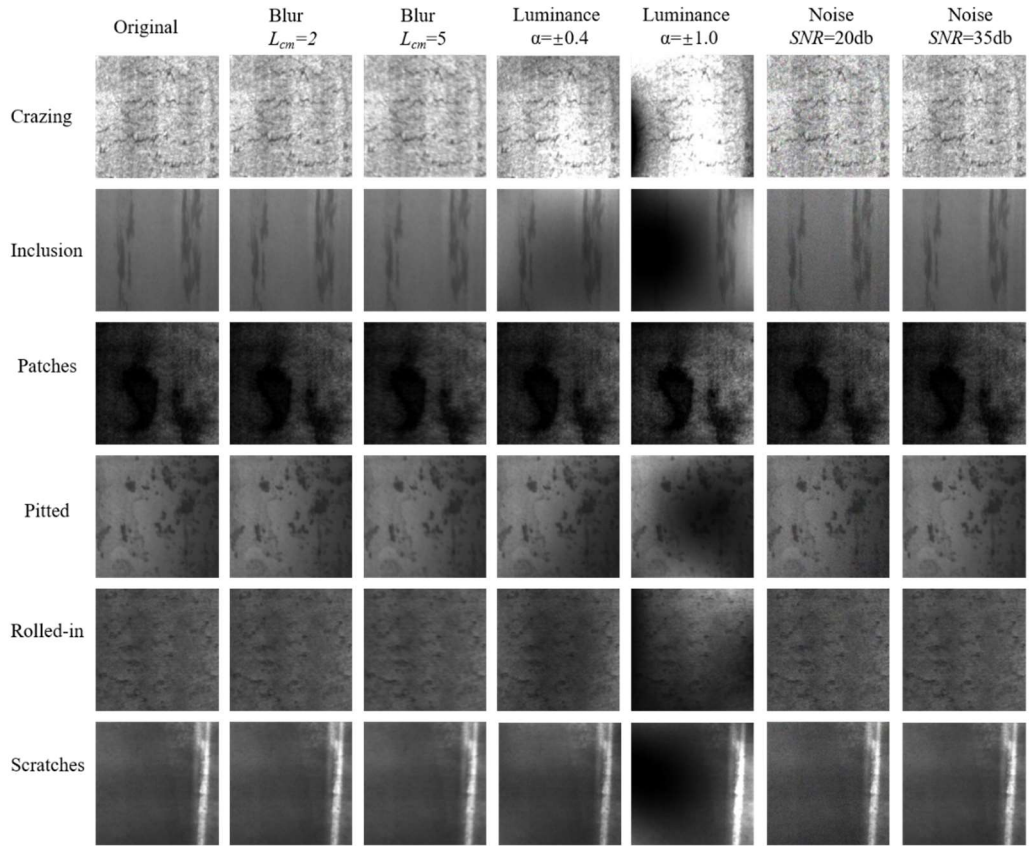


Figure 6. Sample images of six typical surface defects in NEU-CLS dataset after processing.

4.3. Implementation Details

All experiments are performed by Pytorch. We use 70% of the images as the training dataset and 30% of the images as the test dataset. Training is performed on GTX 1060 GPU, and we use SGD with a weight decay of 0.001, momentum of 0.9, and batch size of 16. In order to verify CASI-Net, we conduct experiments in the public surface defect database NEU released by Northeastern University of China [1].

Specifically, the input image size $W \times H \times C$ is $300 \times 300 \times 3$ (width, height, and channel). The dimension $W_1 \times H_1 \times C_1$ of the output in block 1 is $150 \times 150 \times 64$. The dimension

$W_2 \times H_2 \times C_2$ of the output in block 2 is $75 \times 75 \times 128$. The dimension $W_3 \times H_3 \times C_3$ of the output in block 3 is $37 \times 37 \times 256$.

4.4. Performance Analysis

In this section, we establish ablation experiments to evaluate the effectiveness of CASI-Net. The comparison results are shown in Table 1. Firstly, we use the baseline to classify the surface defects of NEU, and the accuracy reached 94.79%. Then, we add the self-interaction module based on the biological visual interaction mechanism to the baseline. The baseline combined with the self-interaction module reaches 95.22% on NEU-CLS. Next, we add the CA block to the baseline without the self-interaction module. The recognition accuracy rate of the baseline after adding the CA block reaches 95.47% on the NEU steel surface defect dataset. Finally, we add the self-interaction module constructed by the biological visual interaction mechanism to the baseline, where the performance of CASI-Net in NEU-CLS reaches 95.83%. After adding the CA module to the baseline, we visualize the sample data of NEU steel surface defects in Figure 7.

Table 1. The classification accuracy (%) of CASI-Net in NEU dataset.

Method	Original	Luminance ($\alpha \pm 0.4$)	Luminance ($\alpha \pm 1$)	Noise (20 db)	Noise (35 db)	Blur (2)	Blur (5)
BLFE + MLP	94.79	92.93	82.54	80.33	92.87	94.31	80.62
BLFE + SI + MLP	95.22	93.53	84.66	85.34	93.26	94.53	82.33
BLFE + CA + MLP	95.47	93.68	87.14	90.63	94.97	95.26	90.19
BLFE + CA + SI + MLP	95.83	94.21	92.56	94.71	95.26	95.66	91.62

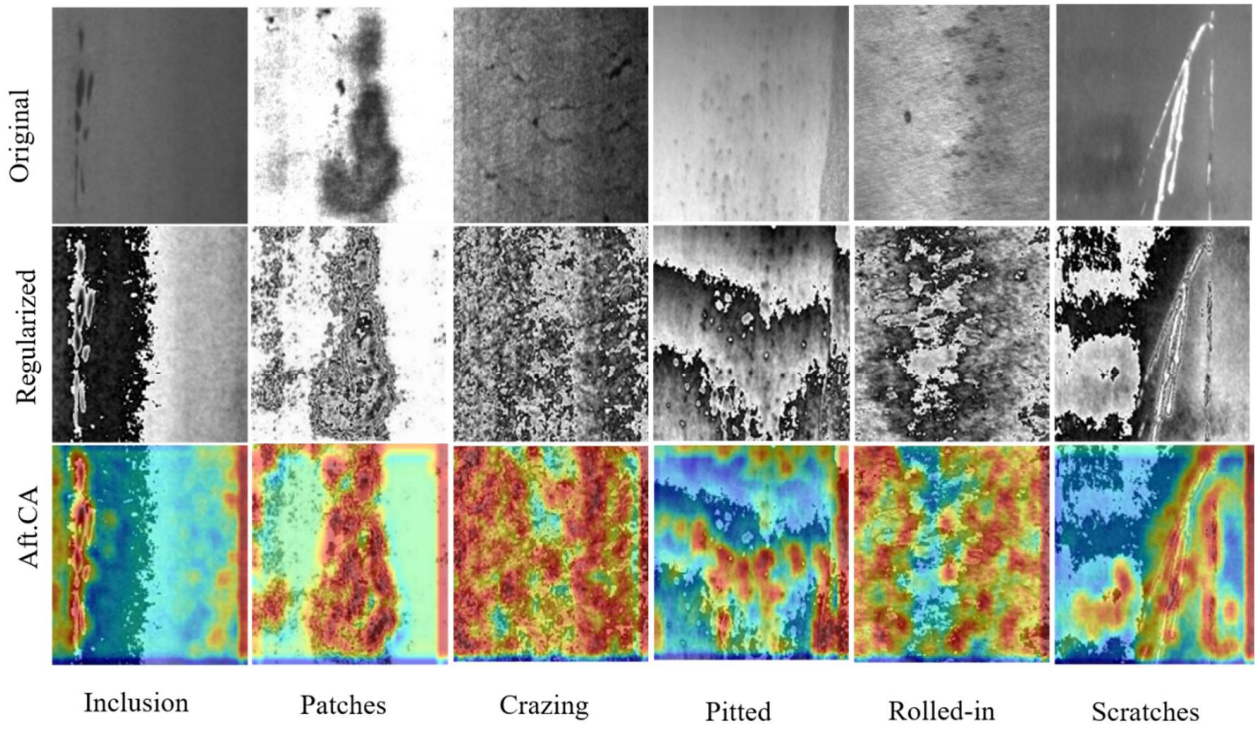


Figure 7. Visualization results of six typical surface defects of hot-rolled strip after being regularized and adding CA.

From Figure 7, we know that CASI-Net can concentrate more on the location of the defect and suppress the non-defect part.

4.5. Comparison with State-of-the-Art Methods

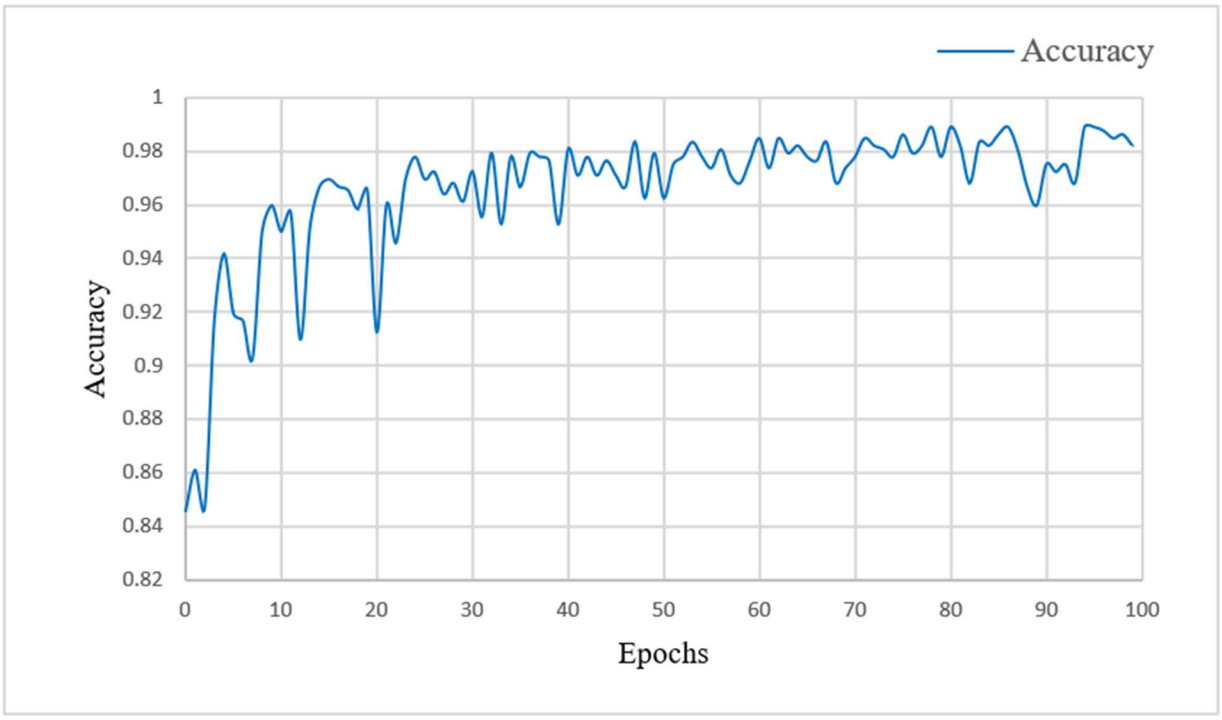
In addition, in order to verify the performance of CASI-Net, we compare the classification accuracy of various advanced steel surface defect classification models. The experimental results in Table 2 show that our proposed CASI-Net can achieve a higher classification accuracy of steel surface defects with fewer parameters. In the NEU public dataset, we evaluate and verify ResNet [16], MobileNet [23], EffNet [17], and CASI-Net. The experimental results show that compared with ResNet with 25.56 M parameters reaching 95.09%, CASI-Net achieves 95.83 % accuracy with much fewer parameters. In addition, compared with MobileNet [23] and EffNet [17], CASI-Net can achieve a higher classification accuracy with little overhead increase. Compared with the most advanced steel surface defect classification, CASI-Net can classify steel surface defects more accurately.

Table 2. The accuracy (%) and the params of CASI-Net with state-of-the-art methods.

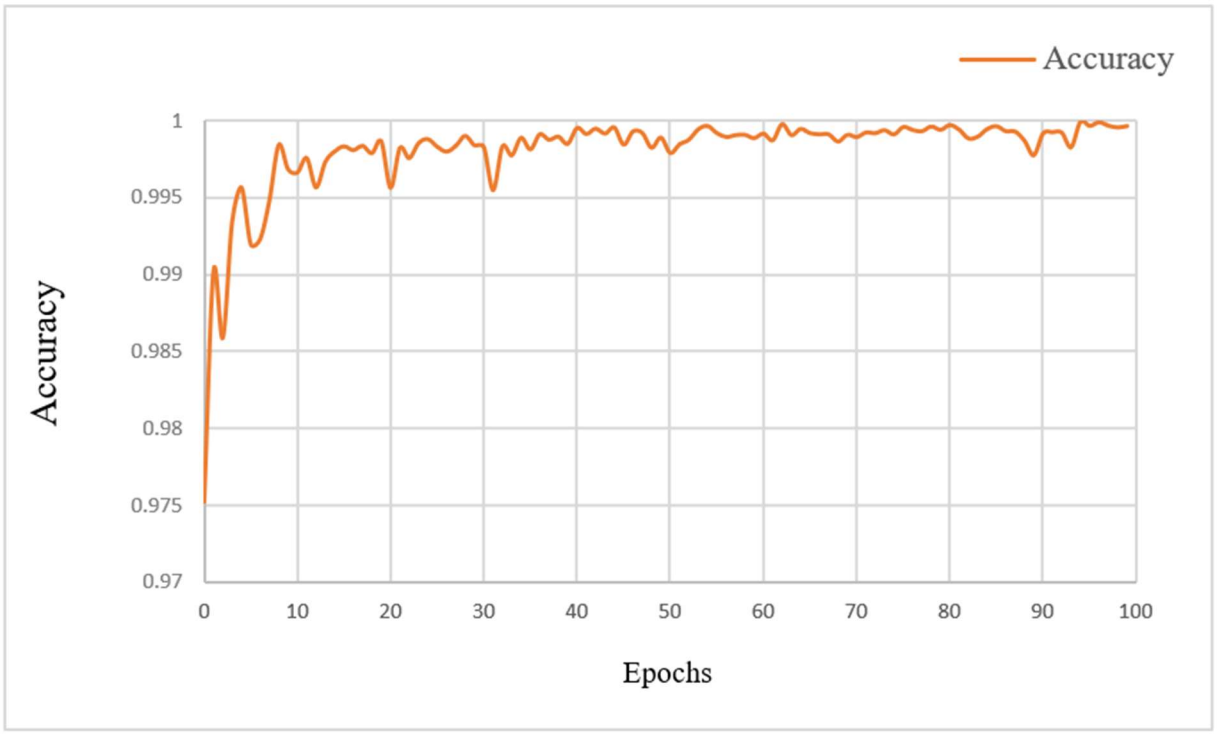
Method	Params	Accuracy
ResNet	25.56 M	95.09
EffNet	2.21 M	94.81
MobileNet	2.23 M	95.57
CASI-Net	2.22 M	95.83

5. Discussion

In this study, we demonstrated that compared with the traditional machine vision, the steel defect classification method based on deep learning can achieve higher classification accuracy. In this paper, we use the coordinated attention mechanism and the self-interaction module based on the biological vision to construct a lightweight convolutional neural network. By introducing the CA block, our network can concentrate more on defect areas. By constructing the SI module based on biological vision, the representation of the feature map is improved, so as to increase the recognition accuracy. In addition, compared with the depth network, our model can achieve a classification accuracy equivalent to that when the amount of parameters is reduced. In addition, we also discussed the impact of different dataset partitions on our construction method. We use 8:2 data division for the training network training and testing. The results show that CASI-Net can finally achieve 98.19% accuracy. We plotted the experimental results. The results and AUROC show that CASI-Net can accurately identify surface defects in Figure 8. Collectively, our data demonstrate that the recognition accuracy of CASI-Net verifies the applicability of our model in the task of surface defect recognition of a hot-rolled strip. However, there are some problems we have not taken into account. For example, for some defects in the dataset, there is a high degree of "inter class similarity and intra class diversity". For convolutional neural networks, it is difficult to distinguish them accurately. Therefore, in the next step, we will consider introducing fine-grained classification methods, such as bilinear pooling to improve the feature map of the extracted image or constructing high-order statistical features to model the channel to improve the feature map of the extracted image and capture the representative defect-recognition area, so as to improve the classification accuracy.



(**a**)



(**b**)

Figure 8. The accuracy variation with the change of epochs and AUROC of CASI-Net. (**a**) Accuracy variation with the change of epochs; (**b**) AUROC.

6. Conclusions and Future Work

This paper presents a light and effective classification network for steel surface defects called CASI-Net which adopts a new convolution block, which greatly reduces the computational burden and achieves high recognition accuracy. The proposed backbone

network can achieve accurate identification results of steel surface defects. We incorporate the attention mechanism and the self-interaction mechanism based on biological vision into CASI-Net to improve the defect recognition accuracy. Our experiments show that CASI-Net can achieve better performance than other models with fewer parameters. In Section 3, we considered using two different technologies to improve the defect recognition accuracy of the CASI-Net, including the CA block and a self-interaction module. In the CA block [18], the location information of feature maps is embedded into the channel attention and decomposed into two 1D feature encoding processes. Then the two 1D features are coded to form a pair of direction-aware and position-sensitive feature maps, which can be complementarily applied to the input feature maps to enhance the representation of the region of interest. Through the CA block, CASI-Net can capture correlation dependencies along the horizontal direction and retain accurate location information along the vertical direction. Inspired by the biological visual interaction mechanism, the self-interaction module is constructed. Through the self-interaction operation, the feature map contains more effective information from the original image, and the representation ability of features in the CNN model is further enhanced to improve the accuracy of defect classification. Overall, the recognition accuracy of CASI-Net is more than 95%, which verifies the applicability of our model in the task of surface defect recognition of a hot-rolled strip. In the future, our next work is to further verify the generalization performance of the model, and utilize optimization algorithms and adaptation equipment, so as to develop a complete steel surface defect diagnosis framework. Based on the needs of iron and steel enterprises, we aim to expand more actual functions, such as online help. In addition, the system can also provide users with more dynamic and beautiful interfaces.

Author Contributions: Conceptualization, Q.H. and C.W.; methodology, C.W.; software, M.H.; validation, G.C.; analysis, Z.L.; investigation, T.W.; writing—original draft preparation, C.W.; writing—review and editing, Q.H.; visualization, C.W.; supervision, Z.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the West Light Foundation of the Chinese Academy of Science, in part by the Research Foundation of The Natural Foundation of Chongqing City (cstc2021jcyj-msxmX0146), in part by the Scientific and Technological Research Program of Chongqing Municipal Education Commission (KJZD-K201901504, KJQN201901537), in part by the humanities and social sciences research of the Ministry of Education (19YJCZH047), and in part by the Scientific and Technological Research Program of Luzhou City(2021-JYJ-92). The authors would like to thank the support of the China Scholarship Council.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kechen, S.; Yunhui, Y. A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects. *Appl. Surf. Sci.* **2013**, *285*, 858–864.
2. Di, H.; Ke, X.; Peng, Z.; Dongdong, Z. Surface defect classification of steels with a new semi-supervised learning method. *Opt. Lasers Eng.* **2019**, *117*, 40–48. [CrossRef]
3. Neogi, N.; Mohanta, D.K.; Dutta, P.K. Review of vision-based steel surface inspection systems. *EURASIP J. Image Video Process.* **2014**, *50*, 1–19. [CrossRef]
4. Fu, G.; Sun, P.; Zhu, W.; Yang, J.; Cao, Y.; Yang, M.Y.; Cao, Y. A deep-learning based approach for fast and robust steel surface defects classification. *Opt. Lasers Eng.* **2019**, *121*, 397–405. [CrossRef]
5. Bo, T.; Jianyi, K.; Shiqian, W. Review of surface defect detection based on machine vision. *J. Image Graph.* **2017**, *22*, 1640–1663.
6. Tao, X.; Hou, W.; Xu, D. A survey of surface defect detection methods based on deep learning. *Acta Autom. Sin.* **2021**, *47*, 1017–1034.

7. Li, H.; Fu, X.; Huang, T. Research on surface defect detection of solar pv panels based on pre-training network and feature fusion. *IOP Conf. Ser. Earth Environ. Sci.* **2021**, *651*, 022071. [CrossRef]
8. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
9. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceeding of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
10. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
11. Chen, S.; Wang, H.; Xu, F.; Jin, Y.Q. Target classification using the deep convolutional networks for sar images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 4806–4817. [CrossRef]
12. Hamdia, K.M.; Ghasemi, H.; Bazi, Y.; AlHichri, H.; Alajlan, N.; Rabczuk, T. A novel deep learning based method for the computational material design of flexoelectric nanostructures with topology optimization. *Finite Elem. Anal. Des.* **2019**, *165*, 21–30. [CrossRef]
13. Jeon, M.; Jeong, Y.-S. Compact and Accurate Scene Text Detector. *Appl. Sci.* **2020**, *10*, 2096. [CrossRef]
14. Vu, T.; Nguyen, C.V.; Pham, T.X.; Luu, T.M.; Yoo, C.D. Fast and Efficient Image Quality Enhancement via Desubpixel Convolutional Neural Networks. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018; p. 11133.
15. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
16. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
17. Freeman, I.; Roese-Koerner, L.; Kummert, A. Effnet: An efficient structure for convolutional neural networks. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 6–10.
18. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
19. Ma, S.; Ban, Y.; Daichen, Z. Survey of convolutional neural network. *Mod. Inf. Technol.* **2021**, *5*, 11–15.
20. Lecun, Y.; Bottou, L. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
21. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5 mb model size. *arXiv* **2016**, arXiv:1602.07360.
22. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856.
23. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Hartwig, A. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
24. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
25. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
26. Jie, H.; Li, S.; Gang, S.; Albanie, S. Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023.
27. Chen, Y.; Kalantidis, Y.; Li, J.; Yan, S.; Feng, J. a2-nets: Double attention networks. *arXiv* **2018**, arXiv:1810.11579.
28. 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.
29. Bello, I.; Zoph, B.; Le, Q.; Vaswani, A. Attention augmented convolutional networks. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seul, Korea, 27 October–2 November 2019; pp. 3286–3295.
30. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
31. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. CCNet: Criss-cross attention for semantic segmentation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 603–612.
32. Hou, Q.; Zhang, L.; Cheng, M.M.; Feng, J. Strip pooling: Rethinking spatial pooling for scene parsing. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 4002–4011.
33. Rosa, M.; Palmer, S.M.; Gamberini, M.; Burman, K.J.; Yu, H.-H.; Reser, D.H.; Bourne, J.A.; Tweedale, R.; Galletti, C. Connections of the dorsomedial visual area: Pathways for early integration of dorsal and ventral streams in extrastriate cortex. *J. Neurosci.* **2009**, *29*, 4548–4563. [CrossRef] [PubMed]
34. Milner, D.A. How do the two visual streams interact with each other? *Exp. Brain Res.* **2017**, *235*, 1297–1308. [CrossRef] [PubMed]
35. Wei, B.; He, H.; Hao, K.; Gao, L.; Tang, X.S. Visual interaction networks: A novel bio-inspired computational model for image classification. *Neural Netw.* **2020**, *130*, 100–110. [CrossRef] [PubMed]

36. Van Polanen, V.; Davare, M. Interactions between dorsal and ventral streams for controlling skilled grasp. *Neuropsychologia* **2015**, *79*, 186–191. [CrossRef] [PubMed]
37. Holtzman, J.D. Interactions between cortical and subcortical visual areas: Evidence from human commissurotomy patients. *Vis. Res.* **1984**, *24*, 801–813. [CrossRef]
38. Das, A.; Gilbert, C.D. Topography of contextual modulations mediated by short-range interactions in primary visual cortex. *Nature* **1999**, *399*, 655. [CrossRef] [PubMed]
39. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 448–456.
40. Chen, F.C.; Jahanshahi, R.M.R. NB-CNN: Deep learning-based crack detection using convolutional neural network and Naïve bayes data fusion. *IEEE Trans. Ind. Electron.* **2018**, *65*, 4392–4400. [CrossRef]
41. Glorot, X.; Bordes, A.; Bengio, Y. Deep sparse rectifier neural networks. *J. Mach. Learn. Res.* **2011**, *15*, 315–323.