symmetry

MDPI

Article

Visual Inspection Method for Metal Rolls Based on Multi-Scale Spatial Location Feature

Degang Xu [1,2], Hao Li [1], Ruirui Wu [3], Yizhi Wang [1], Yonghao Huang [1] and Yaoyi Cai [1,4,*]

[1] School of Automation, Central South University, Changsha 410083, China; dgxu@csu.edu.cn (D.X.); 204611105@csu.edu.cn (H.L.); wangyizhi@csu.edu.cn (Y.W.); 204611051@csu.edu.cn (Y.H.)
[2] State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China
[3] China Nonferrous Metals Processing Technology Co., Ltd., Luoyang 410005, China; 204611078@csu.edu.cn
[4] College of Engineering and Design, Hunan Normal University, Changsha 410081, China
* Correspondence: cyy@hunnu.edu.cn

Abstract: Metal rolls in a non-ferrous-metal manufacturing workshop manifest the characteristics of symmetry, multiple scales and mutual covering, which poses great challenges for metal roll detection. To solve this problem, firstly, an efficient attention mechanism algorithm named ECLAM (efficient capture location attendant model) is proposed for capturing spatial position features efficiently, to obtain complete location information for metal rolls in a complex environment. ECLAM can improve the ability to extract the spatial features of backbone networks and reduce the influence of the non-critical background. In addition, in order to give feature maps a larger receptive field and improve the weight of location information in multi-scale feature maps, a nonlinear feature fusion module named LFFM (location feature fusion module) is used to fuse two adjacent feature images. Finally, a multi-scale object detection network named L-MSNet (location-based multi-scale object detection network) based on the combination of ECLAM and LFFM is proposed and used to accurately detect multi-scale metal rolls. In the experiments, multi-scale metal roll images are collected from an actual non-ferrous-metal manufacturing workshop. On this basis, a pixel-level image dataset is constructed. Comparative experiments show that, compared with other object detection methods, L-MSNet can detect multi-scale metal rolls more accurately. The average accuracy is improved by 2% to 5%, and the average accuracy of small and medium-sized objects is also significantly improved by 3% to 6%.

Keywords: warehouse management; deep learning; object detection; attention mechanisms; feature fusion

Citation: Xu, D.; Li, H.; Wu, R.; Wang, Y.; Huang, Y.; Cai, Y. Visual Inspection Method for Metal Rolls Based on Multi-Scale Spatial Location Feature. *Symmetry* **2022**, *14*, 1291. https://doi.org/10.3390/sym14071291

Academic Editor: Jan Awrejcewicz

Received: 20 May 2022
Accepted: 20 June 2022
Published: 22 June 2022

1. Introduction

Warehouse management plays a vital role in the production and management of manufacturing enterprises. New intelligent material management systems based on computer vision can detect materials automatically and determine their quantity and location immediately using a monitoring camera. This type of management has attracted extensive attention due to its real-time performance and applicability.

Traditional material visual inspection algorithms use the designed operator to extract image features and classify the image, in order obtain the category and position of the material. However, it is very difficult to use some previously designed operators to detect materials in a variety of complex industrial scenes. With the rapid development of artificial intelligence technology, deep learning algorithms have achieved a significant breakthrough. Compared with the traditional detection methods, object detection algorithms based on deep learning usually have not only stronger feature extraction ability but also the advantages of more robust generalization and higher accuracy. Therefore, a considerable amount of research has been undertaken on making full use of deep-learning-based detection methods to detect materials in a complex industrial scene. Sun et al. [1] proposed

an application of Faster R-CNN (faster region with convolutional neural networks) [2] based on autonomous navigation to detect the shelf legs and tags, which is useful for automated guided vehicles (AGVs) in warehouses. Max Schwarz et al. [3] replaced the language model of the DenseCap [4] network with linear SVM (support vector machine) to accurately detect objects in warehouses. Han et al. [5] proposed a lightweight model by introducing channel-level sparsification of YOLOv3 (you only look once, version 3) [6]. They fused the channels between shallow feature maps and deep feature maps to obtain more detailed features of stacked parcels.

However, the above research is mainly limited to single-scale object detection in a simulated scene. There are still some problems associated with metal roll detection in warehouses, such as strong metal reflection, image noise, and so on, as shown in Figure 1. Furthermore, metal rolls in the warehouse are densely stacked and can block each other. Most importantly, the image pixel sizes of the front and rear metal rolls are very different; those for the rear metal rolls are less than one-fifteenth of those for the front metal rolls. Lastly, due to the small image pixel size of the rear metal rolls and serious occlusion, a large number of features are lost. Multi-scale metal rolls caused by the spatial change in this complex environment pose a significant challenge for intelligent visual detection.

(**a**) (**b**)

Figure 1. Metal rolls in a non-ferrous-metal processing warehouse: (**a**) a metal roll image taken from the top; (**b**) a metal roll image taken from the side.

Therefore, a new object detection method is proposed to solve the problem of detection of multi-scale metal rolls in a complex environment. The main contributions of this paper are as follows:

(1) A novel attention mechanism, the efficient capture location attention module (ECLAM), is proposed to make the backbone network focus on the extraction of the spatial location features of multi-scale metal rolls.
(2) A new feature fusion module named a location feature fusion module (LFFM), based on ECLAM, is proposed in this paper. In contrast to other traditional feature fusion algorithms, LFFM adopts a nonlinear superposition method to improve the weight of spatial location features in feature maps.
(3) Based on the feature maps that include plenty of spatial location information extracted by ECLAM and LFFM, a new detection method named a location-based multi-scale object detection network (L-MSNet) is proposed to accurately detect multi-scale metal rolls in a complex environment.

The rest of this paper is organized as follows. Section 2 presents some state-of-the-art object detection methods. Section 3 introduces the overall structure of L-MSNet. The proposed ECLAM and LFFM algorithms are elaborated in Section 4. Section 5 presents a series of experiments to examine the performance of the proposed algorithms. The paper ends with a brief conclusion and suggestions for future work.

2. Related Work

At present, some practical algorithms are available to solve multi-scale object detection problems, such as RetinaNet [7], FCOS (fully convolutional one-stage object detection) [8], ATSS (adaptive training sample selection) [9], and PAA (probabilistic anchor assignment) [10].

RetinaNet is a one-stage anchor-based detector that uses a new cross-entropy loss function named focal loss to solve the problem of data imbalance in multi-scale object detection, in order to improve the detection accuracy. In contrast with RetinaNet, FCOS adopts an anchor-free algorithm, which makes it possible for multi-scale objects to be detected by canceling the preset anchor and solves the detection problem posed by extreme differences in object scales. There are three differences between RetinaNet and FCOS: (1) the distinction between negative samples and positive samples; (2) the status of the start of the regression; and (3) the number of anchors for each location. Experimental results [8] show that FCOS outperforms RetinaNet on various datasets.

According to the results, Zhang et al. [9] further explored the above three differences. They found that the key distinction between anchor-free and anchor-based methods is how the positive and negative training samples are defined, and the accuracy of the detection model is closely related to the selection of positive and negative samples during training. They proposed an ATSS algorithm that can dynamically select high-quality training samples to improve the accuracy of multi-scale object detection. ATSS uses an adaptive anchor assignment which computes the standard deviation and mean of IoU values from some close anchors for each ground truth. Nevertheless, although this method can slightly improve the accuracy, it is not suitable for multiple strong, high-quality anchors. Therefore, PAA adaptively selects positive and negative samples through maximum likelihood probabilities to optimize the training process for anchor-based networks, to solve the above problem and improve the detection accuracy for objects in a complex environment.

Although various methods were used in the above literature to improve the detection accuracy of multi-scale objects, all of them ignored some essential features, i.e., the spatial location features. The spatial location features in a complex environment contain the distribution information of the objects. The proposed L-MSNet method can obtain the spatial location features of small and medium-sized metal rolls densely stacked in the distance in the image, which is helpful for accelerating the convergence speed of networks and improving the detection accuracy for multi-scale metal rolls.

3. The Structure of L-MSNet

Aiming at the problem of multi-scale metal roll detection in a complex environment, a multi-scale object detection network called L-MSNet is proposed, based on spatial location features, as shown in Figure 2. L-MSNet is composed of four parts: the backbone network based on ResNet-50 (residual network 50 layers, ResNet-50) [11] and a feature pyramid network based on an attention mechanism (attention-based FPN); the candidate region generator consisting of an RPN (region proposal network) and ROI Align (region of interest align); the FC (full connection layer) classifier, and the FCN (fully convolutional network) image segmentation module [12].

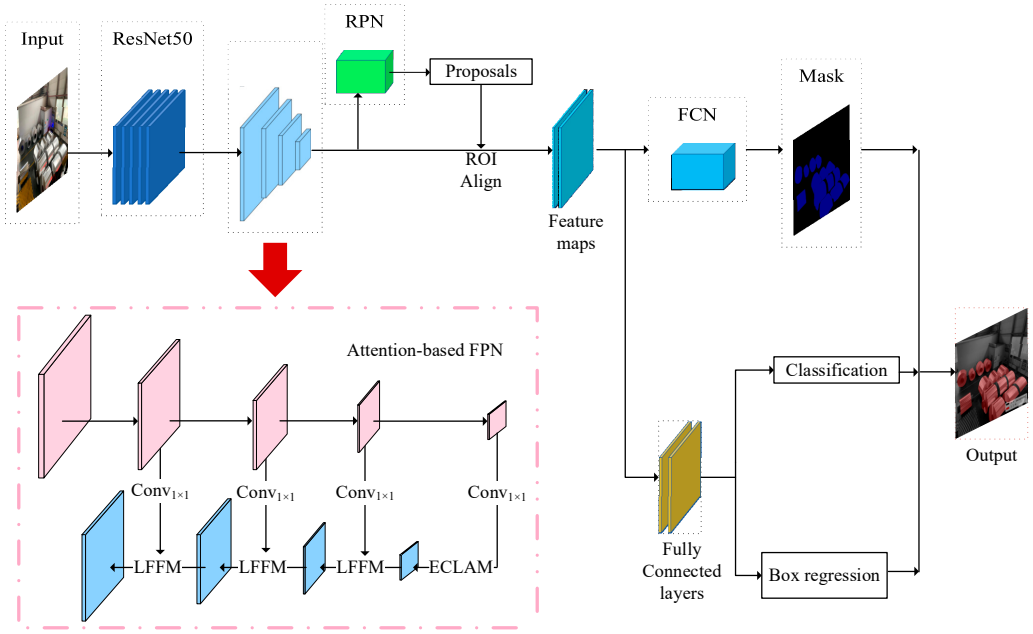

Figure 2. The structure of L-MSNet. The attention-based FPN is used to construct spatial location feature maps based on ECLAM and LFFM, and $Conv_{1\times1}$ is used to change the number of channels.

L-MSNet uses ResNet-50 and the attention-based FPN as the backbone network, to extract the features of metal rolls. ResNet-50 adopts the bypassing connection method, which allows the extracted feature information to spread across multiple hidden layers so that the shallow features and deep features in the network are integrated, to avoid the loss or overfitting of basic features. Then, the extracted basic feature maps are fed into the attention-based FPN. The structure of the attention-based FPN is shown in Figure 2. Firstly, the critical location information of the feature map is extracted through the proposed attention algorithm ECLAM. Then, the multi-scale feature maps are efficiently fused through the proposed LFFM algorithm. Finally, multi-scale feature maps are constructed, which express the key location information more significantly and in more detail.

The RPN aims to generate some regions of interest (ROIs) on the feature maps efficiently and quickly. Firstly, the RPN generates a series of anchor points on the multi-scale feature maps. There are usually three to five regions of interest for each anchor point. By setting multi-scale ROIs, metal rolls can be detected more easily. On this basis, metal rolls can be located by ROI Align, which is based on a bilinear difference algorithm. Finally, the multi-scale feature maps with rich and key location information are input into the FC and FCN to complete detection and segmentation. The loss function of L-MSNet is shown in Equation (1).

$$L = L_{cls} + L_{box} + L_{mask} \tag{1}$$

where L is a multi-task loss that includes three parts, L_{cls} is a classification loss, L_{box} is a bounding-box loss, and L_{mask} is a binary cross-entropy loss which averages all pixels.

4. Method

4.1. Attention Mechanisms

As a bionic technology, an attention mechanism aims to enhance the ability of the backbone network to extract critical features from the data and suppress other unimportant features. In recent years, attention mechanisms have been widely used in deep learning, leading to many achievements [13–15]. There are three kinds of attention mechanisms commonly used in the field of computer vision: local attention, global attention, and multi-head attention. The first two attention mechanisms are realized on the basis of soft attention and hard attention. Local attention focuses on only one or several small windows in an

image, for the sake of paying attention to capture the nuances. Local attention is conducive to detecting small and medium-sized objects. Global attention focuses on capturing the long-distance dependence in an image and pays attention to the global information. Global attention is conducive to detecting large-scale objects. A multi-head attention mechanism is constructed of several attention mechanisms. Multi-head attention enables models to learn critical features in multiple subspaces, which can enhance the feature extraction ability of backbone networks.

4.2. Efficient Capture Location Attention Module

Figure 3 shows the structure of ECLAM which mainly includes two parts. One carries out a series of operations to factorize the input feature map $\boldsymbol{X} \in R^{C\times H\times W}$ into two one-dimensional feature vectors $\boldsymbol{M}_h \in R^{C\times H\times 1}$ and $\boldsymbol{M}_w \in R^{C\times 1\times W}$ in the horizontal and vertical directions, respectively, where C is the number of channels, and H and W denote the height and width, respectively. The other gives different weights to various features in $\boldsymbol{X}$ through the extracted attention vectors $\boldsymbol{M}_h$ and $\boldsymbol{M}_w$, in order to obtain the final feature map $\boldsymbol{Y} \in R^{C\times H\times W}$, which contains rich and key location information. The overall process of ECLAM is summarized in Equation (2), where $\otimes$ denotes element-wise multiplication.

$$\boldsymbol{Y} = \boldsymbol{X} \otimes \boldsymbol{M}_w \otimes \boldsymbol{M}_h \tag{2}$$

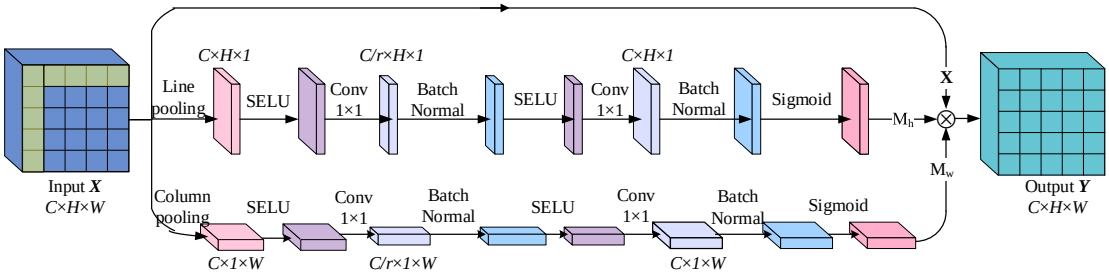

Figure 3. The ECLAM structure. Line pooling is an operation of mean pooling on lines. Accordingly, column pooling is an operation of mean pooling on columns. The size of a feature map is changed four times: after pooling, after convolution 1×1 twice, and after element-wise multiplication.

More specifically, first of all, the mean pooling operation is implemented for each line and column of the input intermediate feature map $\boldsymbol{X} \in R^{C\times H\times W}$. The sizes of the two pooling kernels are $(1, W)$ and $(H, 1)$. As a result, the output of pooling $\boldsymbol{R}(h)_c^h \in R^{C\times H\times 1}$ and $\boldsymbol{R}(w)_c^w \in R^{C\times 1\times W}$ can be formulated as

$$\boldsymbol{R}(h)_c^h = \frac{1}{W}\sum_{i=0}^{W} \boldsymbol{X}_c(h, i) \tag{3}$$

$$\boldsymbol{R}(w)_c^w = \frac{1}{H}\sum_{i=0}^{H} \boldsymbol{X}_c(j, w) \tag{4}$$

where $\boldsymbol{R}(h)_c^h$ denotes the mean pooling result of the c-th channel at the i-th line. Similarly, $\boldsymbol{R}(w)_c^w$ is the mean pooling result of the c-th channel at the j-th column. By encoding each channel of the input feature map $\boldsymbol{X}$ along the vertical and horizontal coordinates, respectively, a 2D feature map $\boldsymbol{X}$ is turned into two 1D vectors. This allows ECLAM to capture long-range dependencies and furnish the location features of metal rolls. What is more, a local channel attention structure represented by L is constructed from batch normalization (BN) [16], 1×1 convolution ($Conv_{1\times 1}$), and a SELU (scaled exponential linear units) activation function [17] and used to aggregate the local channel context, as shown in Equation (5). Therefore, the results $\boldsymbol{M}_h$ and $\boldsymbol{M}_w$ can be expressed as Equations (6) and (7), respectively,

$$L(\boldsymbol{f}) = Conv_{1\times 1}(SELU(BN(Conv_{1\times 1}(SELU(\boldsymbol{f}))))) \tag{5}$$

$$\boldsymbol{M}_h = \sigma(BN(L(\boldsymbol{R}(h)_c^i))) \tag{6}$$

$$\boldsymbol{M}_w = \sigma(BN(L(\boldsymbol{R}(h)_c^j))) \tag{7}$$

where σ represents the sigmoid function, and $\boldsymbol{f}$ is a feature map for processing by L. In addition, $Conv_{1\times1}$ represents a convolution operation with a 1×1 kernel. $\boldsymbol{M}_h$ and $\boldsymbol{M}_w$ contain two $Conv_{1\times1}$ operations, whose sizes are $C/r \times C \times 1 \times 1$ and $C \times C/r \times 1 \times 1$, respectively, and r is the reduction ration for reducing parameters. In this way, channels in the intermediate process can be reduced in order to prevent excessive calculation caused by too many channels. Through the above equations, $\boldsymbol{M}_h$ and $\boldsymbol{M}_w$ can be acquired from the input feature maps $\boldsymbol{X}$ simultaneously. All elements in both $\boldsymbol{M}_h$ and $\boldsymbol{M}_w$ can indicate whether metal rolls are placed in the corresponding lines and columns or not. Accurate location features are of great benefit in helping the backbone network recognize metal rolls better.

4.3. Location Feature Fusion Module

The traditional feature pyramid network (FPN) [18] is commonly used to solve the problem of multi-scale object detection by continuous downsampling and upsampling. The traditional FPN can make feature maps that contain both the shallow and deep information of an image, expanding the receptive field of large-scale feature maps. However, this fusion method cannot enhance the weight of location information in multi-scale feature maps. Furthermore, the location information will be gradually lost with continuous fusions.

In this paper, LFFM is proposed to enhance the weight of location information in multi-scale feature maps. Figure 4 shows the structure of LFFM.

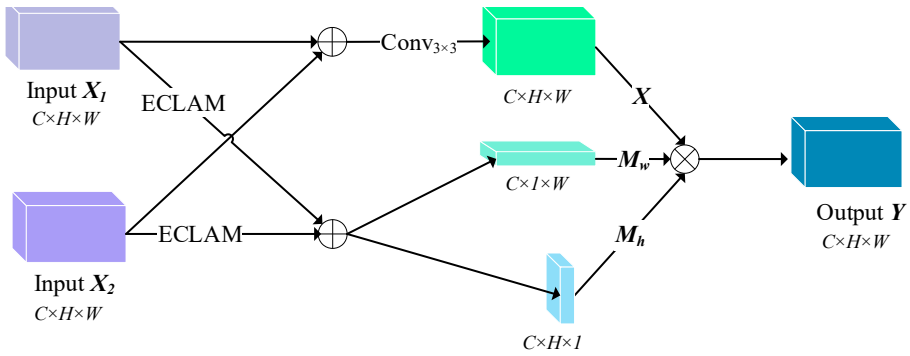

Figure 4. The structure of LFFM and dimensions of each feature map. Similar to $Conv_{1\times1}$, $Conv_{3\times3}$ denotes a convolution operation with a 3×3 kernel in order to remove alias effects.

In contrast to the element addition of the traditional FPN, LFFM uses a nonlinear fusion between two adjacent feature maps to improve the weight of location information. The overall process of LFFM can be summarized by the following equations:

$$\boldsymbol{X} = Conv_{3\times3}(\boldsymbol{X}_1 \oplus \boldsymbol{X}_2) \tag{8}$$

$$\boldsymbol{M}_w = \boldsymbol{M}_{w1} \oplus \boldsymbol{M}_{w2} \tag{9}$$

$$\boldsymbol{M}_h = \boldsymbol{M}_{h1} \oplus \boldsymbol{M}_{h2} \tag{10}$$

where $\boldsymbol{X}_1 \in R^{C\times H\times W}$ and $\boldsymbol{X}_2 \in R^{C\times H\times W}$ refer to a low-level feature map and a high-level feature map, respectively. $\boldsymbol{M}_{h1} \in R^{C\times H\times 1}$ and $\boldsymbol{M}_{h2} \in R^{C\times H\times 1}$ can be calculated by Equation (6) with $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, respectively. Similarly, $\boldsymbol{M}_{w1} \in R^{C\times 1\times W}$ and $\boldsymbol{M}_{w2} \in R^{C\times 1\times W}$ can be calculated by Equation (7). After completing the above calculation, the final output $\boldsymbol{Y} \in R^{C\times H\times W}$ can be obtained using Equation (2). In contrast to the traditional FPN, LFFM not only ensures that each feature map contains high-level semantic information but also makes the location information in multi-scale feature maps more detailed and significant. By introducing LFFM, L-MSNet can effectively and accurately detect metal rolls in complex environments.

Figure 2 shows the structure for embedding LFFM in the FPN, where $Conv_{1\times1}$ is used to reduce the number of channels. In the process of L-MSNet detecting metal rolls, large-scale metal rolls are detected with small-scale feature maps, in order to accelerate the speed of detection. Accordingly, in order to improve the detection accuracy, a large-scale feature map is used to detect small-scale metal rolls. The selection basis is shown in Equation (11)

$$l = [l_0 + \log_2(\sqrt{wh}/224)] \tag{11}$$

where l is the level of the attention-based FPN, [ ] is the round function, and h and w represent the length and width of the region of the proposal, respectively. The factor 224 is derived from the standard size of ImageNet, and l0 corresponds to the level of the attention-based FPN when h and w are 224 and 224, respectively. The level of the feature map corresponding to each region of the proposal can be accurately calculated using Equation (11), which can help to avoid excessive calculation during detection. In contrast to the traditional FPN, the novel attention-based FPN adds ECLAM and LFFM, which ensures that the multi-scale feature maps contain complete key location information. Subsequent experiments prove that L-MSNet based on ECLAM and LFFM can better detect multi-scale metal rolls in complex environments.

5. Experiments and Analysis

5.1. Experimental Settings

5.1.1. Dataset and Implementation Details

The datasets used in the following experiments were all collected from actual large non-ferrous-metal warehouses. We adopted the well-known COCO dataset [19] format to construct a metal roll dataset, and the details are shown in Table 1.

Table 1. Dataset analysis results.

Attribute Name	Attribute Value
Image type	RGB
Number of images	368
Image size	1408 × 1088 and 1920 × 1088
The distribution ratio of image sizes	About 3:1
Metal roll numbers in images	About 2000
The maximum metal roll	About 586,000 pixels
The minimum metal roll	About 500 pixels

All the metal rolls were labeled accurately at the pixel level by LabelMe [20] labeling software. These images can be roughly divided into two viewing angles: the side and the top, as shown in Figure 1. It can be seen that metal rolls are closely stacked together in the image, blocking each other, and some long-distance metal rolls occupy fewer than 32 × 32 pixels. In addition, the images are affected by equipment, light, the site, and other factors and contain varying degrees of metal reflection, image noise, and pedestrian occlusion, which are common phenomena in production. The use of non-idealized data can improve the robustness of L-MSNet and its application in production. In the process of training and testing, images are randomly divided into two groups in a ratio of 7:3. In addition, a series of operations such as noise, rotation, and occlusion are performed on the training images to increase the number of training images, as shown in Figure 5.

(**a**) (**b**) (**c**)

Figure 5. Results of image processing: (**a**) a result for pepper noise; (**b**) a result for rotation; (**c**) a result for occlusion.

The software and hardware names and version numbers used in the experiments are shown in Table 2. In addition, for parameter setting, an Adam optimizer [21] was used in this study. The initial learning rate was set to 0.003, the batch size was set to 10, and the number of iterations was 10,000. Furthermore, the number of $Conv_{1\times1}$ channels C and the reduction rate r of ECLAM were set to 256 and 32.

Table 2. Experimental environment configuration.

Attribute Name	Attribute Value
Operating system version	Ubuntu18.04.3 LTS
Graphics	NVIDIA GeForce RTX 2080Ti × 4
Processor	Intel(R) Xeon(R) Silver 4114 CPU @2.20 GHz × 2
RAM	256 G
CUDA	11.1
Data processing	Python 3.7.10, OpenCV 4.5.2
Deep learning framework	Pytorch 1.8.0, Detectron 2.0.4

5.1.2. Evaluation Indexes

In order to verify the effectiveness of L-MSNet, especially for the detection effect for metal rolls with different scales, four evaluation indexes were used: the average precision (AP), average precision for small objects (AP_s), average precision for medium objects (AP_m), and average precision for large objects (AP_l). AP denotes the average precision when the intersection over union (IoU) increases from 0.5 to 0.95 in steps of 0.05. AP_s refers to objects occupying an area in the image of less than 32 × 32 pixels. Similarly, AP_m indicates objects occupying an area between 1024 and 9216 pixels, and AP_l represents objects whose area is larger than 9216 pixels. The precision is calculated using Equation (12)

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

where TP refers to the total number of positive samples which are predicted as positive and FP is the count of false negative samples which are predicted as positive. The precision is the fraction of true positive samples in the group of samples declared positive by the classifier.

5.2. Experimental Results and Analysis

5.2.1. Ablation Experiments

L-MSNet without ECLAM and PFFM is regarded as the benchmark model. In order to verify the influence of ECLAM and LFFB, a series of ablation experiments was carried out. By embedding the ECLAM algorithm in the x-axis and y-axis in L-MSNet, the effectiveness of ECLAM in the x-axis and y-axis directions, respectively, was verified. The experimental results are shown in Table 3.

Table 3. Results of the benchmark model obtained by adding ECLAM in different directions.

Base	*X*	*Y*	*AP* (%)	AP_s (%)	AP_m (%)	AP_l (%)
√	—	—	59.2	36.7	56.2	78.6
√	√	—	60.5	38.6	58.1	79.2
√	—	√	60.3	38.3	58.2	79.2
√	√	√	**62.3**	**39.2**	**61.1**	**80.4**

It can be seen from Table 3 that adding unidirectional attention can improve the detection performance of the benchmark model, and the effects of the two directions on the benchmark model are basically same. Embedding unidirectional ECLAM can improve the AP of the benchmark model by about 2.1% to 2.3%. Accordingly, by adding the ECLAM algorithm in the x-axis and y-axis directions, the four evaluation indexes of the benchmark model can be greatly improved. In particular, AP_s and AP_m are improved by 2.5% and 4.9%, respectively, compared with the benchmark model. The experimental results show that ECLAM can improve the detection accuracy of multi-scale metal rolls by extracting the location features in two directions.

Based on the above experiments, the LFFM algorithm was added to L-MSNet. The detection results of LFFM and the traditional FPN were compared to verify the superiority of LFFM for the fusion of location features. The experimental results are shown in Table 4. After embedding LFFM, the detection effect of L-MSNet on metal rolls reached the optimal level of 63.8%. Compared with the traditional FPN, the four evaluation indexes AP, AP_s, AP_m, and AP_l increased by 1.5%, 2.3%, 2.4%, and 0.8%, respectively. The results show that LFFM can efficiently enhance the weight of location information in multi-scale feature maps, further improving the accuracy of L-MSNet.

Table 4. Results for the benchmark model with the addition of LFFM.

Base	ECLAM	FPN	LFFM	*AP* (%)	AP_s (%)	AP_m (%)	AP_l (%)
√	—	√	—	59.2	36.7	56.2	78.6
√	√	√	—	62.3	39.2	61.1	80.4
√	√	√	√	**63.8**	**41.5**	**63.5**	**81.2**

5.2.2. Comparative Experiments

Mask R-CNN [22] is a general network framework that has been developed based on Faster R-CNN. It is used to solve the problem of multi-scale object detection and has resulted in considerable achievements in many fields [23,24]. In order to verify the detection performance of the algorithm fairly and intuitively, the backbone network of Mask R-CNN was set as ResNet-50 + FPN. The detection results for both are shown in Figure 6. In order to clearly distinguish the differences between them, the differences in the image are marked and enlarged within red boxes. According to the comparison results, the first column contains some images of large-scale metal rolls. L-MSNet network correctly detected all metal rolls in the image, while Mask R-CNN missed some of them. The second and third columns clearly show the advantages of L-MSNet in detecting small and medium-sized metal rolls. Here, L-MSNet successfully detected more metal rolls. Even small and medium-sized metal rolls with serious occlusal problems could be successfully detected with higher confidence scores. The experimental results show that L-MSNet can solve the problem of detecting multi-scale metal rolls better than Mask R-CNN.

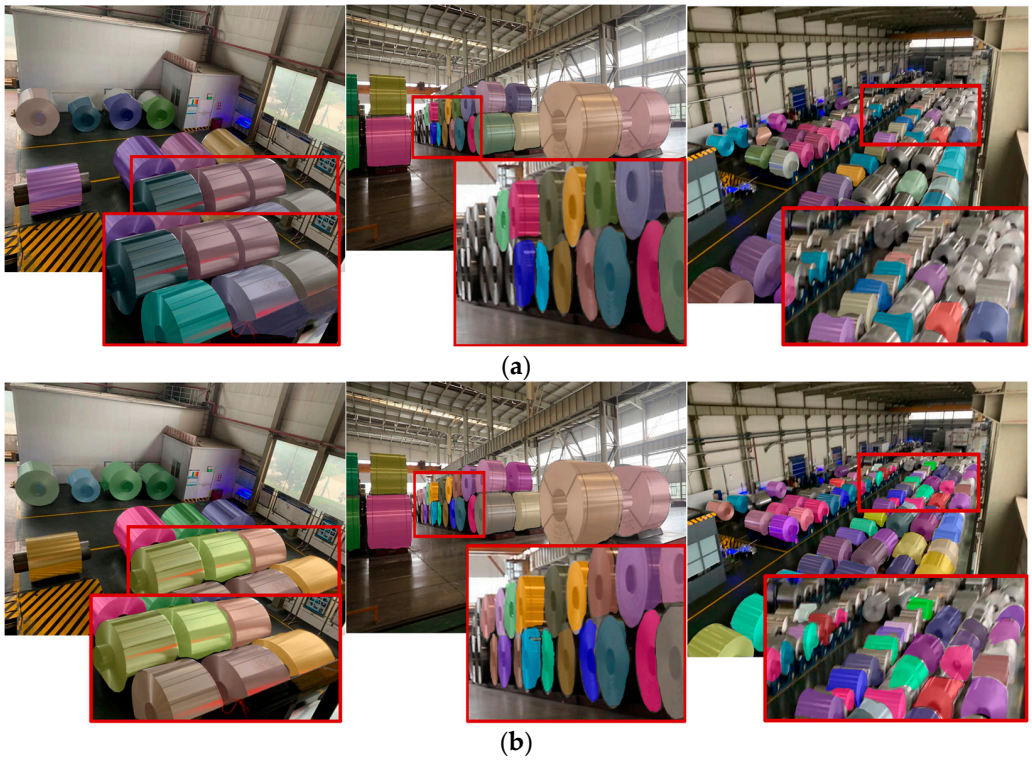

Figure 6. Comparison of the detection results between benchmark model and L-MSNet. (**a**) Detection result for Mask R-CNN; (**b**) detection result for L-MSNet.

In order to further verify the effectiveness of L-MSNet, L-MSNet was compared with mainstream target detection modules such as RetinaNet [7], FCOS [8], ATSS [9], PAA [10], BorderDet [25], and Yolov5m [26]. The backbone network for each module was ResNet-50 + FPN. The results are shown in Table 5.

Table 5. Comparison of the detection results between L-MSNet and other algorithms.

Modules	*AP* (%)	AP_s (%)	AP_m (%)	AP_l (%)	Param.
RetinaNet	58.6	34.1	57.8	77.9	38.0 M
FCOS	58.5	36.5	55.4	77.4	32.3 M
ATSS	60.7	36.2	58.7	81.1	32.2 M
PAA	61.0	37.5	58.1	**81.4**	32.3 M
BorderDet	61.6	37.2	60.2	**81.4**	33.6 M
Yolov5m	61.3	37.4	60.1	80.7	35.7 M
L-MSNet	**63.8**	**41.5**	**63.5**	81.2	**39.8 M**

According to Table 5, The average precision of L-MSNet was the best, especially for the detection of small and medium-sized metal rolls. Compared with other modules, the average precision of L-MSNet was about 2~5% higher than the others. Furthermore, the AP_s and AP_m values of L-MSNet were about 3~7% higher than those of other modules. For detecting large metal rolls, although the AP_l for PAA and BorderDet was 0.2% higher than for L-MSNet, the evaluation indexes AP_s and AP_m for L-MSNet were much higher, i.e., more than 4% and 5% higher, respectively. The average precision of L-MSNet was also significantly better than those of PAA and BorderDet. Moreover, although L-MSNet has slightly more parameters than the other networks, the precision of L-MSNet was significantly higher than the others. It is quite acceptable for a non-ferrous-metal manufacturing workshop to build a more accurate warehouse system by introducing some additional parameters.

The experimental results show that the detection effect of L-MSNet on multi-scale metal rolls is clearly better than that of other modules, which fully proves the effectiveness of this method. In order to solve the problem of multi-scale metal roll detection in complex environments, the attention algorithm ECLAM makes the ResNet-50 backbone network focus on extracting location features from the RGB images and constructing feature maps containing rich key location information. Furthermore, LFFM enhances the weight of location information in multi-scale feature maps by nonlinearly fusing two adjacent feature maps. By taking advantage of ECLAM and LFFM, L-MSNet can accurately detect multi-scale metal rolls in a complex environment.

6. Conclusions and Future Work

Aiming at the problem of detection of multi-scale metal rolls in a complex environment, a novel detection method was proposed to overcome the difficulties. The proposed method used the attention algorithm ECLAM to integrate the spatial position information in the horizontal and vertical directions, which can ensure that the backbone focuses on extracting the spatial position features of multi-scale metal rolls. On this basis, a new nonlinear location feature fusion algorithm LFFM was proposed, to fuse the spatial location information between two adjacent feature maps. LFFM ensures that location features which are extracted by ECLAM are expressed more significantly and in more detail. Taking advantage of ECLAM and LFFM, a new detection network L-MSNet was proposed, to detect multi-scale metal rolls accurately in a complex environment. The experimental results showed that L-MSNet had high detection accuracy for metal rolls in a complex environment. Compared with other state-of-the-art detection networks such as RetinaNet, FCOS, and ATSS, the average precision of L-MSNet was about 2~5% higher than the others, and the AP_s and AP_m values of L-MSNet were about 3~6% higher than for the other modules. The detection of large-scale metal rolls was also improved.

At present, metal roll images are independent of each other, and the features do not take into account the synergistic effect. In our future research work, we will build a multi-camera cooperative work system to detect metal rolls from multiple angles, in order to achieve high-precision dynamic detection of metal rolls. What is more, we will segment the metal rolls in the images more finely, in order to locate the metal rolls precisely in the non-ferrous-metal manufacturing workshop, which is also important for building an intelligent warehouse management system.

Author Contributions: Formal analysis, project administration, and funding acquisition, D.X.; methodology and writing—original draft preparation, H.L.; visualization and validation, R.W. and Y.H.; writing—review and editing, Y.W.; investigation and data curation, Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by The National Key Research and Development Program of Chinese Intelligent Robot under grant number 2018YFB1309000, The National Natural Science Foundation of China under grant number 61973320, The joint fund of Liaoning Province State Key Laboratory of Robotics under grant number 2021KF2218, The Youth Program of National Natural Science Foundation of China under grant number 61903138, and The Postgraduate Scientific Research Innovation Project of Hunan Province under grant number QL20210048, and supported by the Fundamental Research Funds for the Central Universities of Central South University.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Sun, Y.; Su, T.; Tu, Z. Faster R-CNN based autonomous navigation for vehicles in warehouse. In Proceedings of the IEEE International Conference on Advanced Intelligent Mechatronics, Munich, Germany, 3–7 July 2017; pp. 1639–1644.
2. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
3. Schwarz, M.; Milan, A.; Lenz, C.; Munoz, A.; Periyasamy, A.S.; Schreiber, M.; Schüller, S.; Behnke, S. Nimbro picking: Versatile part handling for warehouse automation. In Proceedings of the IEEE International Conference on Robotics and Automation, Singapore, 29 May–3 June 2017; pp. 3032–3039.
4. Johnson, J.; Karpathy, A.; Fei-Fei, L. Densecap: Fully convolutional localization networks for dense captioning. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4565–4574.
5. Han, S.; Liu, X.P.; Han, X.; Wang, G.; Wu, S.B. Visual Sorting of Express Parcels Based on Multi-Task Deep Learning. *Sensors* **2020**, *20*, 6785. [CrossRef] [PubMed]
6. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
7. Lin, T.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *42*, 318–327. [CrossRef] [PubMed]
8. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 9627–9636.
9. Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; Li, S.Z. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9759–9768.
10. Kim, K.; Lee, H.S. Probabilistic Anchor Assignment with IoU Prediction for Object Detection. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 355–371.
11. 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.
12. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 640–651. [CrossRef]
13. Niu, Z.Y.; Zhong, G.Q.; Yu, H. A review on the attention mechanism of deep learning. *Neurocomputing* **2021**, *452*, 48–62. [CrossRef]
14. Hou, Q.B.; Zhou, D.Q.; Feng, J.S. Coordinate Attention for Efficient Mobile Network Design. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* **2021**, 13708–13717. [CrossRef]
15. Liang, H.; Zhou, H.; Zhang, Q.; Wu, T. Object Detection Algorithm Based on Context Information and Self-Attention Mechanism. *Symmetry* **2022**, *14*, 904. [CrossRef]
16. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, Miami, FL, USA, 7–9 July 2015; pp. 448–456.
17. Klambauer, G.; Unterthiner, T.; Mayr, A.; Hochreiter, S. Self-normalizing neural networks. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 972–981.
18. Lin, T.Y.; Dollar, P.; Girshick, R.; He, K.M.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Hawaii Convention Center, Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
19. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
20. Russell, B.C.; Torralba, A.; Murphy, K.P.; Freeman, W.T. LabelMe: A database and web-based tool for image annotation. *Int. J. Comput. Vis.* **2008**, *77*, 157–173. [CrossRef]
21. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
22. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 386–397. [CrossRef] [PubMed]
23. Zuo, L.; He, P.; Zhang, C.; Zhang, Z. A robust approach to reading recognition of pointer meters based on improved mask-RCNN. *Neurocomputing* **2020**, *388*, 90–101. [CrossRef]
24. Cheng, T.; Wang, X.; Huang, L.; Liu, W. Boundary-Preserving Mask R-CNN. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 660–676.
25. Qiu, H.; Ma, Y.; Li, Z.; Liu, S.; Sun, J. Borderdet: Border feature for dense object detection. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 549–564.
26. Yan, B.; Fan, P.; Lei, X.; Liu, Z.; Yang, F. A real-time apple targets detection method for picking robot based on improved YOLOv5. *Remote Sens.* **2021**, *13*, 1619. [CrossRef]