applied sciences

MDPI

Article

A Serial Attention Frame for Multi-Label Waste Bottle Classification

Jingyu Xiao [1,†], Jiayu Xu [2,†], Chunwei Tian [3,4,5,*], Peiyi Han [6,*], Lei You [7] and Shichao Zhang [1,*]

1 School of Computer Science, Central South University, Changsha 410083, China; jyxiao@csu.edu.com
2 Shenzhen Research Institute, Guangdong Databeyond Technology Co., Ltd., Shenzhen 518057, China; xujiayu@databeyond.cn
3 School of Software, Northwestern Polytechnical University, Xi'an 710129, China
4 Yangtze River Delta Research Institute of NPU, Taicang 215400, China
5 Research & Development Institute of Northwestern Polytechnical University in Shenzhen, Northwestern Polytechnical University, Shenzhen 518057, China
6 Department of Computer Science, Harbin Institute of Technology, Shenzhen 518055, China
7 School of Bio medical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA; lei.you@uth.tmc.edu
* Correspondence: chunweitian@nwpu.edu.cn (C.T.); hanpeiyi@hit.edu.cn (P.H.); zhangsc@csu.edu.cn (S.Z.)
† These authors contributed equally to this work.

Featured Application: This work is mainly applied to the waste recycling industry, especially to the waste plastic bottle image classification task.

Abstract: The multi-label recognition of damaged waste bottles has important significance in environmental protection. However, most of the previous methods are known for their poor performance, especially in regards to damaged waste bottle classification. In this paper, we propose the use of a serial attention frame (SAF) to overcome the mentioned drawback. The proposed network architecture includes the following three parts: a residual learning block (RB), a mixed attention block (*MAB*), and a self-attention block (SAB). The RB uses ResNet to pretrain the SAF to extract more detailed information. To address the effect of the complex background of waste bottle recognition, a serial attention mechanism containing *MAB* and SAB is presented. *MAB* is used to extract more salient category information via the simultaneous use of spatial attention and channel attention. SAB exploits the obtained features and its parameters to enable the diverse features to improve the classification results of waste bottles. The experimental results demonstrate that our proposed model exhibited good recognition performance in the collected waste bottle datasets, with eight labels of three classifications, i.e., the color, whether the bottle was damage, and whether the wrapper had been removed, as well as public image classification datasets.

Keywords: multi-label image classification; waste bottle; serial attention frame; mixed attention block; self-attention block

Citation: Xiao, J.; Xu, J.; Tian, C.; Han, P.; You, L.; Zhang, S. A Serial Attention Frame for Multi-Label Waste Bottle Classification. *Appl. Sci.* **2022**, *12*, 1742. https://doi.org/10.3390/app12031742

Academic Editor: Manuel Armada

Received: 20 December 2021
Accepted: 31 January 2022
Published: 8 February 2022

1. Introduction

Garbage classification is important in establishing a digital society. Specifically, multi-label waste bottle classification is one of the common tasks in garbage classification. However, due to the various damaging conditions they can be exposed to, the multi-label classification of waste bottles has been incredibly challenging. To address this problem, scholars are devoted to finding solutions in the following three ways: seeking label semantic relationships, objecting proposals, and visual attention methods. In terms of seeking label semantic relationships, some discriminative methods, i.e., recurrent neural network (RNN) [1], Bayesian network [2], and graph convolution network (GCN) [3], have been used to establish semantic relationships. Although these methods have strong learning

abilities, they cannot define adjacent matrices and require higher computational costs. Object proposal methods (i.e., hypotheses-CNN-pooling (HCP) [4] and random crop pooling (RCP) [5]) generate object bounding box proposals, and then they extract useful information from these object proposals for image classification. Although they have a faster execution speed in comparison to traditional sliding windows, obtaining object proposals with these methods may be time consuming. Taking these factors into account, attention methods are presented [6,7]. For instance, multi-class attentional regions (MCAR) [6] and the spatial regularization network (SRN) [7] used the current state to guide the previous state to extract salient features at a lower cost, in order to improve performance. However, most of the previous methods performed poorly, especially with regards to damaged waste bottle classification. Inspired by that, we propose a serial attention frame (SAF) to overcome the aforementioned drawback. This paper includes the following three parts: a residual learning block (RB), a mixed attention block (*MAB*), and a self-attention block (SAB). The RB uses 101-layer ResNet [8] to pretrain SAF to extract more detailed information. To address the effect of the complex background of waste bottle recognition, a serial attention mechanism, containing *MAB* and SAB, is presented. *MAB* is used to extract more salient category information via the simultaneous use of channel attention [9] and spatial attention [9] in multi-label recognition. SAB exploits the obtained features and its parameters to enable the diverse features to obtain scores to improve the classification results of waste bottles. The contributions can be summarized as follows:

(1) A CNN is used to recognize waste bottles.
(2) A serial attention frame, mainly including channel attention, spatial attention, and self-attention mechanisms, is used to extract the salient features of diverse types to improve the multi-label classification of waste bottles.
(3) Image datasets of waste bottles are collected.

The remainder of the paper is organized as follows: Section 2 discusses the related work. Section 3 reveals the architecture of our proposed model. The experiments are shown in Section 4. Section 5 is the conclusion.

2. Related Work

Multi-label classifications can be divided into three patterns, i.e., label semantic relationships, object proposals, and attentive mechanisms. More details of these methods can be found in the following sections.

2.1. Bottle Classifications

In terms of bottle recognition, Muresan et al. extracted the bottle information via segments, identified the interested text area and converted the obtained information to human-readable characters [10]. Fang et al. proposed a binocular inspection algorithm to overcome impurities in ampoules and penicillin bottles [11]. Thiyagarjan et al. firstly collected the logo of bottles via a camera, then classified the obtained logo to recognize bottles [12]. Bottle classifications are meaningful to waste bottle classifications. We created a waste bottle dataset, which is shown in Section 4.1.

2.2. Label Semantic Relationships

An intuitive way of solving multi-label classification problems is to train multiple independent binary classifiers [13]. These methods often refer to finding relationships among labels, e.g., a keyboard and a mouse in a given image, to achieve a classifier [14]. Inspired by that, to capture label semantic relationships, RNN has been presented. Wang et al. [15] proposed an end-to-end frame, as well as CNN-RNN, to learn the dependency of semantic labels through a combination of CNN and RNN. Although this method can identify relationships between different labels, it requires a fixed label sequence in the training process. To address this issue, Chen et al. [16] used an attention and long–short-term memory (LSTM) [17] to achieve an order-free label sequence to obtain a classifier. Although RNNs have obtained remarkable results in image classification, they may rely on heavy manual

parameter tuning to obtain optional parameters. To handle this drawback, GCNs [18] have been developed.

Semantic-specific graph representation learning (SSGRL) [3] directly uses a GCN to enhance the interactions among semantic regions to extract more representative features [19]. Chen et al. [20] used a GCN to build a graph of labels to represent a set of mutually dependent object classifiers. You et al. [21] followed in this manner, but used a cross-modality attention mechanism instead. Additionally, there are some other works regarding multi-label classification. For instance, Guo et al. [2] adopted a conditional dependency network to exploit label co-occurrence information to improve the classification performance. Although these methods can extract representative features via a GCN, they still suffer from the challenge of high computational costs. To tackle this issue, object proposals have been developed.

2.3. Object Proposals

Object proposals use two steps to conduct multi-label classification [22,23]. The first step needs to locate the positions of the objects. The second step needs to recognize different labels. For instance, Yang et al. [24] exploited two types of features, i.e., feature view and label view, to predict the labels of each proposal to improve the recognition results. However, this method referred to bounding box annotations in the training process. To handle this problem, HCP [4] used Edgebox [25] or BING [26] algorithms to generate hypothesis object proposals of an image, then used a CNN with a max-pooling operation to extract salient features and recognize objects. In addition, Wang et al. [5] randomly scaled and cropped an image to extract features, and also used a CNN. Although object proposals are effective for multi-label classification tasks, they may lead to huge human costs in the chosen object proposals or heavy labor for bounding box annotations. Alternatively, attention mechanisms can extract salient features to make a tradeoff between performance and speed, according to the context.

2.4. Attention Mechanisms

Some CNNs tend to increase the depth or width of deep CNNs to improve the performance of image classification. However, that may lead to a higher computational overhead [27]. To overcome this drawback, attention methods are conducted [28]. Due to the low resource consumption, attention methods are extended to multi-label classification tasks, such as recurrent attentional reinforcement learning (RARL) [29] and recurrently discovering attentional regions (RDAR) [1]. Specifically, RARL adopts a recurrent attention reinforcement learning module to recursively learn the attentional regions, and RDAR uses a spatial transformer to efficiently find the locations of interested regions in the multi-label classification tasks. In addition, Gao et al. [6] also tried to discover the attentional regions, but divided the process into three sub-processes using a multi-class attentional region module. Scholars tend to generate an attention map of each label, instead of attentional regions. For example, Zhu et al. [7] used label-specific attention to obtain the spatial and semantic relationships of labels. There are also other attention mechanisms, such as visual attention consistency [30] and the attention pathway [31]. Compared to object proposals and seeking label semantic relationships, attention-based models are good tools to balance performance and computational costs in multi-label classification. Inspired by this, the attention mechanism is used in a CNN for multi-label classification in this paper. Due to the complementarity of different features from different attention mechanisms, we fuse different attention mechanisms to identify more salient features in this paper.

3. The Proposed SAF

The proposed 104-layer SAF contains the following three parts: a residual learning block (RB), a mixed attention block (*MAB*), and a self-attention block (SAB), as shown in Figure 1, where the images collected from the cameras are not very clear, due to the

illumination conditions of the sorting machine. More detailed information is illustrated in Figure 1.

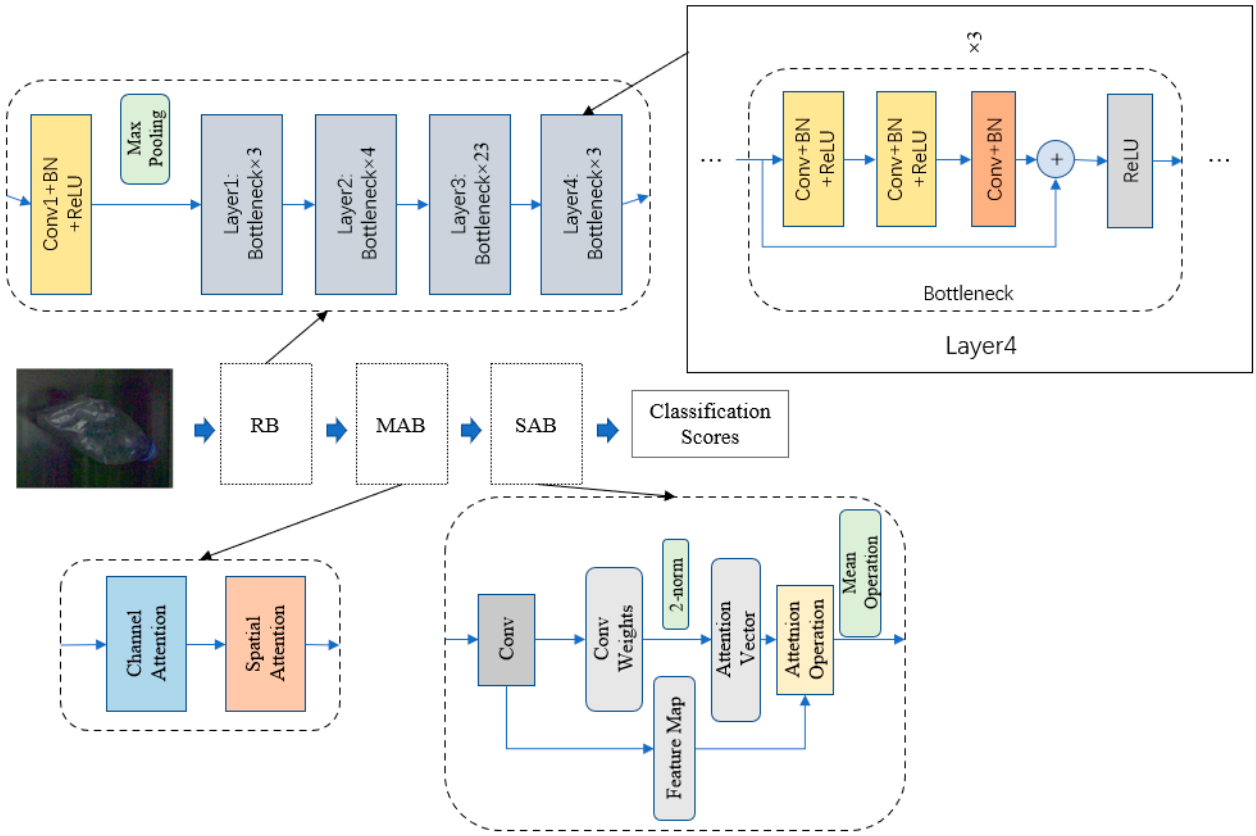


Figure 1. Network architecture of SAF. The residual learning block (RB) is pretrained to learn more context features. The mixed attention block (*MAB*) can extract salient category information via the simultaneous use of channel attention and spatial attention. The self-attention block (SAB) constructs a classifier with attention to improve performance.

RB: The RB uses a 101-layer ResNet [8] to pretrain SAF to extract more detailed information, where its parameters are referred to in [8]. Let us denote a bottle image with a size of 448 × 448 as the input of RB. The RB contains 1-layer Conv1 + BN + ReLU, Layer1, Layer2, Layer3, and Layer4. Specifically, one-layer Conv1 + BN + ReLU denotes the combination of a convolutional layer, batch normalization (BN) [32], and rectified linear unit (ReLU) [33], where its parameters are an input channel number of 3, kernel size of 7 × 7, and output channel number of 64. Moreover, it acts as a one-layer max-pooling layer, which acts as a 9-layer Layer1. Layer1 contains three bottlenecks, and each bottleneck is made up of Conv + BN + ReLU, Conv + BN, and a single ReLU; their parameters are introduced Ref. [8]. The output of Layer1 is the input of Layer2. Layer2 includes four bottlenecks and it connects Layer3. Layer3 is composed of twenty-three bottlenecks and it connects Layer4. Layer4 consists of three bottlenecks and it connects *MAB*. The process can be represented as Equation (1), as follows:

$$O_R = R(I) = L4(L3(L2(L1(MP(CBR(I)))))) \quad (1)$$

Moreover, R expresses a function of ResNet, and is a function of the combination of a convolutional layer, BN, ReLU, and max-pooling layer, respectively. In addition, $L1$, $L2$, $L3$, and $L4$ denote functions of Layer1, Layer2, Layer3, and Layer4, respectively. Further, O_R denotes the output of ResNet as the input of *MAB*.

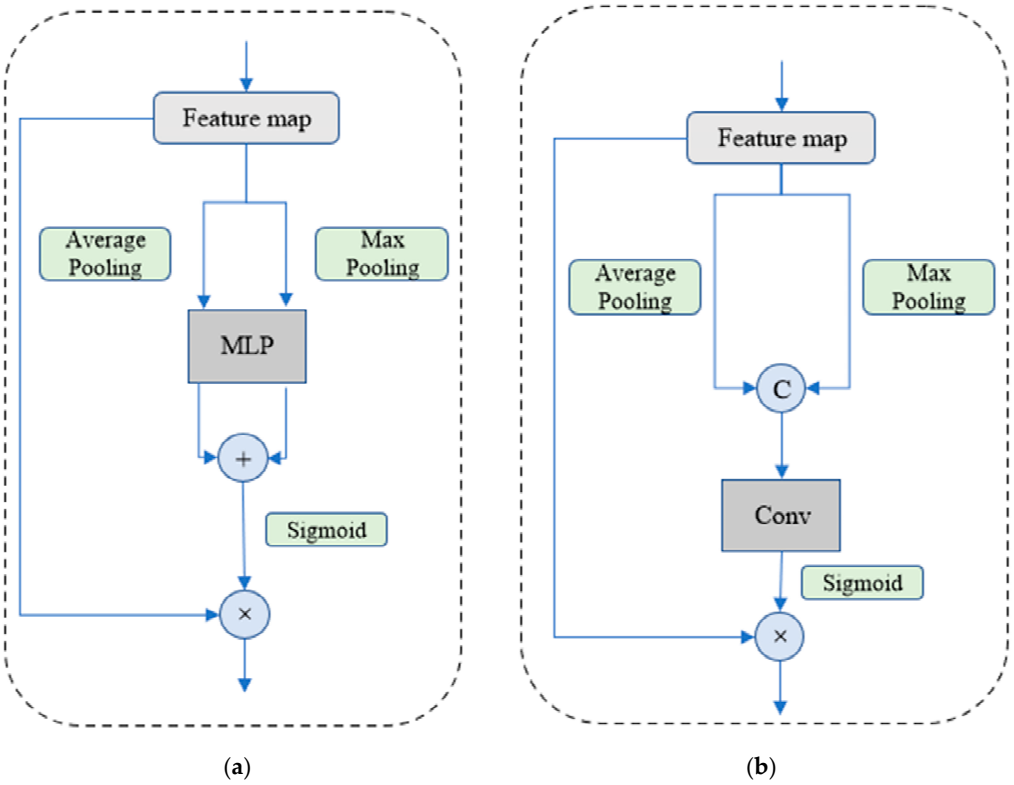
MAB: *MAB* uses one-layer channel attention [9] and one-layer spatial attention [9] to extract more salient category information. Specifically, mixed attention can be formulated as Equation (2), as follows:

$$\begin{aligned} O_{MAB} &= MAB(O_R) \\ &= S(C(O_R)) \end{aligned} \tag{2}$$

where *MAB* is the function of *MAB*, which is composed of the functions (C and S) of channel attention and spatial attention. The channel attention block [9] aims to find out which channels deserve attention, and its structure is shown in Figure 2a. It consists of a max-pooling operation, an average pooling operation, a multi-layer perceptron (MLP), and a sigmoid operation. Specifically, it firstly exerts average-pooling and max-pooling operations to create a map with the size $14 \times 14 \times 2048$, and then a parameter-shared MLP aggregates the obtained feature vectors. More detailed information can be found in Equation (3), as follows:

$$\begin{aligned} O_c &= C(O_R) \times O_R \\ &= f_{sigmoid}(f_{mlp}(f_{avg}(x)) + f_{mlp}(f_{\max}(x))) \times O_R \end{aligned} \tag{3}$$

where $\times$ is the multiplication operation. Let f_{avg} and f_{max} define the average-pooling operation and max-pooling operation, respectively. $f_{sigmoid}$ is used to express the sigmoid function and f_{mlp} is a function of MLP, which is shown in Figure 3. O_c denotes the output of channel attention, which acts as the spatial attention block.



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

Figure 2. Network architecture of channel attention block (**a**) and spatial attention block (**b**). The channel attention block learns which channels of feature maps are important; the spatial attention block finds out "where" feature maps are vital. Together, *MAB* can aggregate more effective information regarding categories.

Figure 3. Our waste bottle images from the collected dataset. There are 3902 images in the datasets. We randomly picked 3224 of them as the training set and the others as the test set. These images are classified according to three categories, i.e., color, whether the bottle is damaged, and whether the wrapper is removed.

The spatial attention block [9] mainly locates the important spatial positions of all the feature channels. It consists of a combination of max pooling and average pooling, as well as a convolutional layer and a sigmoid function. Different from the channel attention block, a convolutional layer is utilized to aggregate feature vectors. The parameters of the convolutional layer are an input channel number of two, kernel size of 3×3, and output channel number of one. The structure of the spatial attention block is shown in Figure 2b, and in Equation (4), as follows:

$$\begin{aligned} O_s &= S(O_c) \times O_c \\ &= f_{sigmoid}(C(Concat(f_{avg}(O_c), f_{\max}(O_c)))) \times O_c \end{aligned} \tag{4}$$

where *Concat* denotes the concatenation operation and C is a convolutional layer with a size of 3×3. Additionally, its output channel is $c1$. O_s expresses the output of spatial attention and it acts the same as SAB.

SAB: SAB exploits the obtained features and its parameters to enable the diverse features to improve the classification results of waste bottles. SAB uses the following self-attention mechanism [34] to implement its function: Firstly, SAB utilizes convolution with a size of 1×1 to refine the obtained features of *MAB*; its input channel is 2048 and the output channel is $c1$ ($c1 = 8$). The obtained features are represented by f and its dimension is $c1 \times 14 \times 14$. Secondly, we obtained the weights of a convolutional layer of 1×1 as W, and its dimension is $c1 \times 14 \times 14$. Thirdly, we use 2-norm to deal with the weights of each channel, and regard them as new weights of the attention mechanism, as shown in Equation (5), as follows:

$$W^j = \sqrt[2]{\left(w_1^j\right)^2 + \left(w_2^j\right)^2 + \left(w_3^j\right)^2 + \ldots + \left(w_d^j\right)^2} \tag{5}$$

where j ($j = 1, 2, 3, 4, 5, 6, 7, 8$) is the order number of a channel and d ($d = 1, 2, 3 \ldots \ldots, 195, 196$) denotes the order number corresponding to the pixel point. Fourthly, we use an attention operation to learn these obtained features. That is, we use the obtained features

to divide the obtained weights of each channel into new features. This process can be formulated as Equation (6), as follows:

$$f_j = f_j / W^j \tag{6}$$

where f_j denotes the obtained features of the jth channel in f. Finally, we calculated the mean operation to deal with all the feature points of each channel. Its implementations can be shown in Equation (7), as follows:

$$S^j = \frac{1}{196} \sum_{i=1}^{196} x_i^j \tag{7}$$

where x_i^j denotes the i-th feature point in fi. S^j is the score of the jth category, where $j \in \{1,2,3,4,5,6,7,8\}$. If S^j is more than 0.5, it is regarded as the jth category.

4. Experiment

4.1. Dataset

Our bottle image dataset containing 3902 waste bottle images is captured by Guangdong Databeyond Technology Co., Ltd., Shenzhen, China. Figure 4 shows a few examples of our collected dataset. We randomly choose 3224 waste bottle images as a training dataset and other images as a test dataset to obtain a classifier of waste bottles. Specifically, the test dataset is divided into eight categories, such as white, green, blue, wrapper removed, wrapper not removed, damaged and not damaged, according to color, whether it is damaged and whether the wrapper is untagged. Additionally, to further test the classification performance of the proposed SAF, we choose public VOC2007 [35] and WIDER Attribute [36] to evaluate the classification results. VOC2007 has been widely used to evaluate multi-label recognition models, as well as for object detection in image segmentation models. It contains 9963 images from a total of 20 labels, which are divided into two subsets, train-val set and test set. We adopt the train-val set to pretrain SAF and the test set for evaluation. WIDER Attribute is a large human attribute dataset, which has 57,524 images and 14 human attribute binary labels. It has a train-val set and test set. In this paper, we follow [7]'s method to deal with the unspecified labels.

4.2. Loss Function

Inspired by MCAR [6] and class-specific residual attention (CSRA) [34], we choose binary cross-entropy (BCELoss) [37] as the loss function to optimize parameters in this paper.

4.3. Experimental Settings

We conduct our experiments on an NVIDIA Titan RTX GPU, an Intel(R) Core(TM) i7 CPU, and a RAM of 32 GB. The CUDA version and cnDNN version are 11.4 and 9, respectively. We implement our model with PyTorch of 1.8.0 and python of 3.7.6. The initial learning rate is set as 0.01 and it varies by 0.1 for every four epochs. The batch size is 16 and we iteratively train the model for 30 epochs in total. In addition, we utilize ResNet101 pretrained on the ImageNet set as the backbone. The Adam [38] algorithm is utilized to optimize model parameters, with a momentum of 0.9 and weight decay of 0.0001.

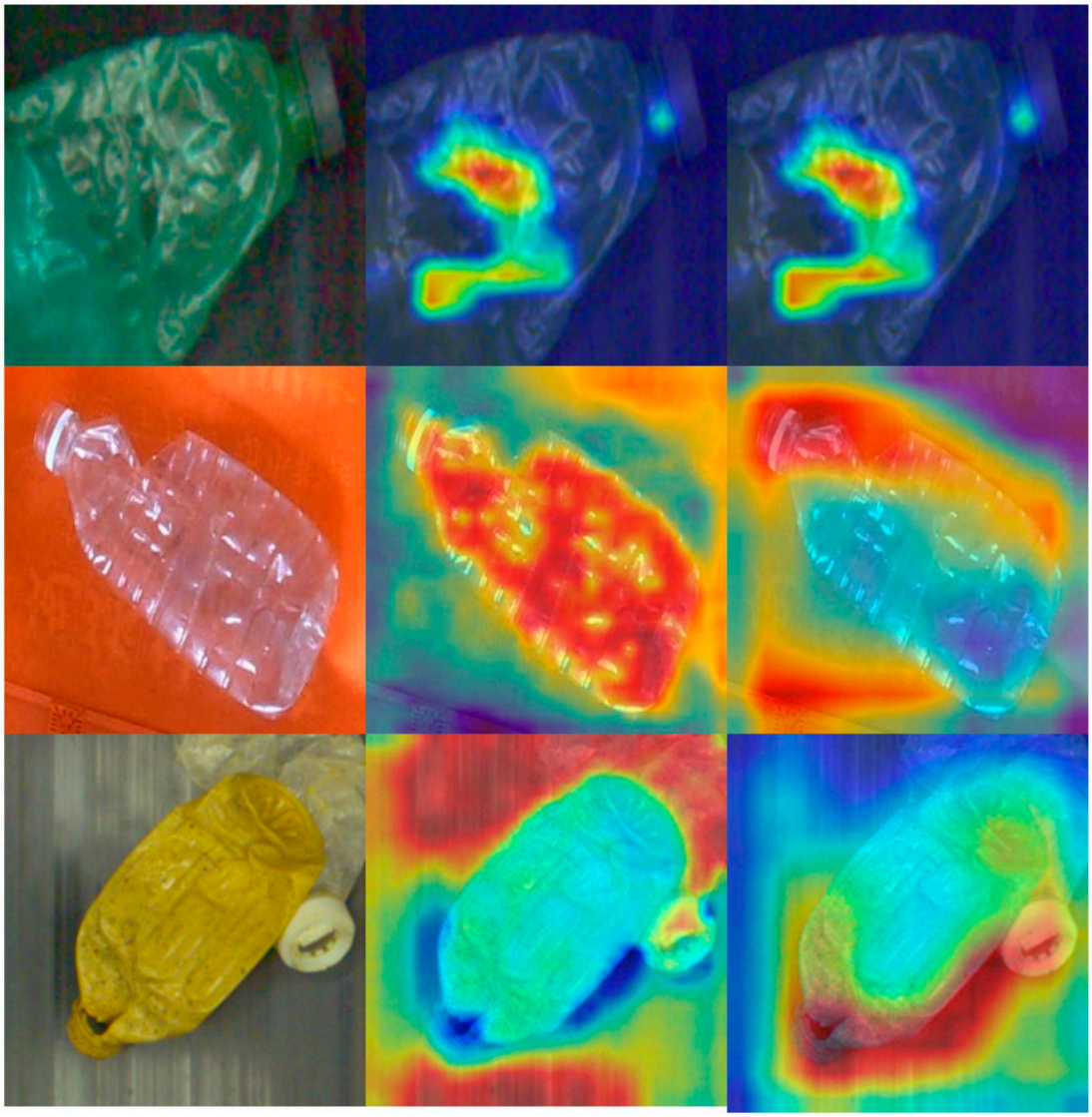


Figure 4. One bottle image (on the left) and its grad cam map of a waste bottle with the damaged classification on the last layer of RB (in the middle) and on the last layer of *MAB* (on the right). Our model can precisely locate the "damaged area" and the *MAB* makes it more focused (we mark relatively unimportant pixels for classification in blue and relatively important ones in red).

4.4. Ablation Experiments

We firstly verify the effectiveness of each module. As described in Figure 4, RB and *MAB* have a positive effect on SAF. As shown in Table 1, RB + SAB + classifier outperforms RB + classifier on mean average precision (mAP) [39], overall F1 measure (CF1) [40] and per-category F1 measure (OF1) [40], where RB + SAB + classifier denotes the combination of RB, SAB, and classifier. This also shows the effectiveness of SAB. SAF obtains better results than RB + SAB + classifier on three indexes, i.e., mAP, CF1, and OF1 in Table 1, which tests the effectiveness of SAB. These materials prove that the proposed techniques have good performance in waste bottle image classification.

Table 1. Results of key techniques in mAP, CF1 and OF1 (%) on waste bottle dataset.

Methods	mAP	CF1	OF1
RB + Classifier	97.39	93.41	96.40
RB + SAB + Classifier	97.46	93.63	96.33
SAF	97.79	94.50	97.05

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

Eleven methods, i.e., RDAR [1], semantic-specific graph representation learning (SSGRL) [3], HCP [4], SRN [7], graph convolutional networks for multi-label classification (ML-GCN) [20], feature view and label view (Fev + Lv) [24], RARL [29], CSRA [34], deep hierarchical contexts (DHC) [36], visual attention (VA) [41], visual attention aggregation (VAA) [42], ASL [43], Query2labels [44], CNN-RNN [15], VGG + SVM [45], and VeSPA [46], are used as comparison methods on mAP, OF1, and CF1 to test the performance in waste bottle image classification. As shown in Table 2, we can observe that our model exceeds the combination of RB and CSRA, where its parameters are the same as SAF. Though SAF has more parameters, the FLOPs and running time just slightly surpass CSRA. Considering its performance improvement, this is totally negligible. As illustrated in Table 3, our SAF outperforms the state-of-the-art methods, i.e., ML-GCN and SSGRL in mAP on VOC2007. From Table 4, we can observe that the proposed SAF outperforms DHC, VA, SRN, and VAA in mAP, CF1, and OF1 on WIDER Attribute. According to these illustrations, it is known that the proposed method is very competitive on waste bottle image classification and public image classification datasets.

Table 2. Comparison results of two methods on metrics, parameters, FLOPs, and running time on waste bottle dataset.

Methods	Metrics (%)			Parameters (M)	Running Time (ms per Image)
	mAP	CF1	OF1		
ASL [43]	81.03	72.10	71.41	53.576	2.30
Query2labels [44]	87.09	82.57	89.23	193.507	7.07
Resnet101 + CSRA [34]	97.25	93.55	96.48	42.516	8.42
SAF	97.79	94.50	97.05	52.129	8.62

Table 3. Classification results of different methods on VOC2007.

Methods	CNN-RNN [15]	VGG + SVM [45]	FeV + LV [24]	HCP [4]	RDAR [1]	RARL [29]	SSGRL [3]	ML-GCN [20]	SAF
mAP (%)	84.0	89.7	90.6	90.9	91.9	92.0	93.4	94.0	94.1

Table 4. Classification results of different methods on WIDER Attribute.

Methods	DHC [36]	VeSPA [46]	VA [41]	RARL [29]	SRN [7]	VAA [42]	SAF
mAP (%)	81.3	82.4	82.9	82.9	86.2	86.4	87.0
CF1 (%)	-	-	-	-	75.9	-	77.2
OF1 (%)	-	-	-	-	81.3	-	82.0

5. Conclusions

To obtain a robust classifier for waste bottles, we propose a serial attention frame (SAF). The SAF relies on a residual learning block, a mixed attention block, and a self-attention

block. The residual learning block uses ResNet to pretrain SAF to extract more detailed information, according to the transfer learning idea. The mixed attention block exploits spatial attention and channel attention to extract more salient category information. In addition, to identify diverse features, a self-attention block utilizes its obtained features and its parameters to improve the classification performance of waste bottles. In the future, we intend to extend SAF to deal with other waste image classifications. Additionally, we will enhance the illumination of the sorting machine to improve the quality of the collected waste images, in order to improve the performance of waste bottle classifications.

Author Contributions: Conceptualization, J.X. (Jingyu Xiao) and J.X. (Jiayu Xu); methodology, S.Z. and C.T.; software, J.X. (Jingyu Xiao); validation, J.X. (Jingyu Xiao); formal analysis, J.X. (Jingyu Xiao); investigation, J.X. (Jingyu Xiao) and L.Y.; resources, P.H.; data curation, J.X. (Jiayu Xu); writing—original draft preparation, J.X. (Jingyu Xiao); writing—review and editing, C.T.; visualization, J.X. (Jingyu Xiao); supervision, S.Z. and C.T.; project administration, S.Z. and C.T.; funding acquisition, C.T. and S.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the Shenzhen Science and Technology Program under Grant 2021A1515110079, in part by the Fundamental Research Funds for the Central Universities under Grant D5000210966, in part by Basic Research Plan in Taicang under Grant TC2021JC23, and in part by the Key Project of NSFC under Grant 61836016. Information regarding the funder and the funding number should be provided. And The APC was funded by the Key Project of NSFC under Grant 61836016.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Data was obtained from Guangdong Databeyond Technology Co., Ltd. and are available at https://www.databeyond.cn/ (accessed on 19 December 2021) with the permission of Guangdong Databeyond Technology Co., Ltd.

Acknowledgments: This work was supported in part by the Shenzhen Science and Technology Program under Grant 2021A1515110079, in part by the Fundamental Research Funds for the Central Universities under Grant D5000210966, in part by Basic Research Plan in Taicang under Grant TC2021JC23, and in part by the Key Project of NSFC under Grant 61836016.

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

References

1. Wang, Z.; Chen, T.; Li, G.; Xu, R.; Lin, L. Multi-label image recognition by recurrently discovering attentional regions. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 464–472.
2. Guo, Y.; Gu, S. Multi-label classification using conditional dependency networks. In Proceedings of the International Joint Conference on Artificial Intelligence, Barcelona, Spain, 16–22 July 2011.
3. Chen, T.; Xu, M.; Hui, X.; Wu, H.; Lin, L. Learning semantic-specific graph representation for multi-label image recognition. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 522–531.
4. Wei, Y.; Xia, W.; Lin, M.; Huang, J.; Ni, B.; Dong, J.; Zhao, Y.; Yan, S. HCP: A flexible CNN framework for multi-label image classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 1901–1907. [CrossRef] [PubMed]
5. Wang, M.; Luo, C.; Hong, R.; Tang, J.; Feng, J. Beyond object proposals: Random crop pooling for multi-label image recognition. *IEEE Trans. Image Process.* **2016**, *25*, 5678–5688. [CrossRef] [PubMed]
6. Gao, B.; Zhou, H. Learning to Discover Multi-Class Attentional Regions for Multi-Label Image Recognition. *IEEE Trans. Image Process.* **2021**, *30*, 5920–5932. [CrossRef]
7. Zhu, F.; Li, H.; Ouyang, W.; Yu, N.; Wang, X. Learning spatial regularization with image level supervisions for multi-label image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 22–25 July 2017; pp. 5513–5522.
8. 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 (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
9. Woo, S.; Park, J.; Lee, J.; 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.
10. Muresan, M.P.; Szabo, P.A.; Nedevschi, S. Dot Matrix OCR for Bottle Validity Inspection. In Proceedings of the IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP), Tokyo, Japan, 15–17 May 2019.

11. Fang, J.; Wang, Y.; Wu, C. Binocular automatic particle inspection machine for bottled medical liquid examination. In Proceedings of the 2013 Chinese Automation Congress, Changsha, China, 7–8 November 2013; pp. 397–402.
12. Thiyagarajan, K.; Meenakshi, R.; Suganya, P. Vision based bottle classification and automatic bottle filling system. In Proceedings of the International Conference on Advances in Human Machine Interaction, Bangalore, India, 3–5 March 2016.
13. Kiranyaz, S.; Ince, T.; Gabbouj, M. *Image Classification and Retrieval by Collective Network of Binary Classifiers*; Springer: Berlin/Heidelberg, Germany, 2014.
14. Xiao, J.; Tang, S. Joint Learning of Binary Classifiers and Pairwise Label Correlations for Multi-label Image Classification. In Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), Shenzheng, China, 6–8 August 2020.
15. Wang, J.; Yang, Y.; Mao, J.; Huang, Z.; Huang, C.; Xu, W. CNN-RNN: A unified framework for multi-label image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2285–2294.
16. Chen, S.F.; Chen, Y.C.; Yeh, C.K.; Wang, Y.C.F. Order-free RNN with visual attention for multi-label classification. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI), New Orleans, LA, USA, 2–7 February 2018; pp. 6714–6721.
17. Jin, C.; Weihua, L.I.; Chen, J.I. Bi-directional Long Short-term Memory Neural Networks for Chinese Word Segmentation. *J. Chin. Inf.* **2018**, *32*, 29–37.
18. Kip, F.T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In Proceedings of the International Conference on Learning Representations, Toulun, France, 24–26 April 2017.
19. Hochreiter, S.; Schmidhuber, J. Long short-termmemory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
20. Chen, Z.; Wei, X.; Wang, P.; Guo, Y. Multi-Label Image Recognition with Graph Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 5172–5181.
21. You, R.; Guo, Z.; Cui, L.; Long, X.; Bao, Y.; Wen, S. Cross-modality attention with semantic graph embedding for multi-label classification. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 12709–12716.
22. Wu, J.; Yu, Y.; Huang, C. Deep multiple instance learning for image classification and auto-annotation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 8–10 June 2015.
23. Liu, Y.; Li, S.J.; Cheng, M.M. RefinedBox: Refining for Fewer and High-quality Object Proposals. *Neurocomputing* **2020**, *406*, 106–116. [CrossRef]
24. Yang, H.; Zhou, J.T.; Zhang, Y.; Gao, B.; Wu, J.; Cai, J. Exploit bounding box annotations for multi-label object recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 280–288.
25. Zitnick, C.L.; Dollar, P. Edge boxes: Locating object proposals from edges. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; pp. 391–405.
26. Cheng, M.M.; Zhang, Z.; Lin, W.Y.; Torr, P. BING: Binarized normed gradients for objectness estimation at 300fps. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 24–27 June 2014; pp. 3286–3293.
27. Tian, C.; Xu, Y.; Li, Z.; Zuo, W.; Liu, H. Attention-guided cnn for image denoising. *Neural Netw.* **2020**, *124*, 117–129. [CrossRef] [PubMed]
28. Paoletti, M.E.; Moreno-Álvarez, S.; Haut, J.M. Multiple Attention-Guided Capsule Networks for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2014**. [CrossRef]
29. Chen, T.; Wang, Z.; Li, G.; Lin, L. Recurrent attentional reinforcement learning for multi-label image recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Zurich, Switzerland, 6–12 September 2018; pp. 6730–6737.
30. Guo, H.; Zheng, K.; Fan, X.; Yu, H.; Wang, S. Visual attention consistency under image transforms for multi-label image classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 729–739.
31. Luo, Y.; Jiang, M.; Zhao, Q. Visual Attention in Multi-Label Image Classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 16–20 June 2019; pp. 820–827.
32. 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 June 2015; Volume 37, pp. 448–456.
33. Nair, V.; Hinton, G. Rectified Linear Units Improve Restricted Boltzmann Machines. In Proceedings of the International Conference on Machine Learning, Haifa, Israel, 21–24 June 2010; Volume 27, pp. 807–814.
34. Ke, Z.; Wu, J. Residual Attention: A Simple but Effective Method for Multi-Label Recognition. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Virtual Event, 11–17 October 2021; pp. 184–193.
35. Everingham, M.; Gool, L.V.; Williams, C.K.I.; Winn, J.; Zisserman, A. The PASCAL Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
36. Li, Y.; Huang, C.; Loy, C.C.; Tang, X. Human attribute recognition by deep hierarchical contexts. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016; Volume 9910, pp. 684–700.
37. Hsieh, C.Y.; Lin, Y.A.; Lin, H.T. A Deep Model with Local Surrogate Loss for General Cost-sensitive Multi-label Learning. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 3239–3246.

38. Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
39. Henderson, P.; Ferrari, V. End-to-end training of object class detectors for mean average precision. In Proceedings of the Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016; pp. 198–213.
40. Wang, Y.; He, D.; Li, F.; Long, X.; Zhou, Z.; Ma, J.; Wen, S. Multi-label classification with label graph superimposing. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12265–12272.
41. Guo, H.; Fan, X.; Wang, S. Human attribute recognition by refining attention heat map. *Pattern Recognit. Lett.* **2017**, *94*, 38–45. [CrossRef]
42. Sarafianos, N.; Xu, X.; Kakadiaris, I.A. Deep imbalanced attribute classification using visual attention aggregation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; Volume 11215, pp. 680–697.
43. Ben-Baruch, E.; Ridnik, T.; Zamir, N.; Noy, A.; Zelnik-Manor, L. Asymmetric Loss for Multi-Label Classification. Asymmetric loss for multi-label classification. *arXiv* **2020**, arXiv:2009.14119.
44. Liu, S.; Zhang, L.; Yang, X.; Su, H.; Zhu, J. Query2label: A simple transformer way to multi-label classification. *arXiv* **2021**, arXiv:2107.10834.
45. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
46. Sarfraz, M.S.; Schumann, A.; Wang, Y.; Stiefelhagen, R. Deep view-sensitive pedestrian attribute inference in an end-to-end model. In Proceedings of the British Machine Vision Conference, London, UK, 4–7 September 2017.