sustainability

MDPI

Article

Recycling Waste Classification Using Vision Transformer on Portable Device

Kai Huang [1,2,†], Huan Lei [1,†], Zeyu Jiao [1,*] and Zhenyu Zhong [1]

[1] Guangdong Key Laboratory of Modern Control Technology, Institute of Intelligent Manufacturing, Guangdong Academy of Sciences, Guangzhou 510070, China; kirehuang@163.com (K.H.); huan.l@giim.ac.cn (H.L.); zy.zhong@giim.ac.cn (Z.Z.)
[2] School of Automation, Guangdong University of Technology, Guangzhou 510006, China
* Correspondence: zy.jiao@giim.ac.cn
† These authors contributed equally to this work.

Abstract: Recycling resources from waste can effectively alleviate the threat of global resource strain. Due to the wide variety of waste, relying on manual classification of waste and recycling recyclable resources would be costly and inefficient. In recent years, automatic recyclable waste classification based on convolutional neural network (CNN) has become the mainstream method of waste recycling. However, due to the receptive field limitation of the CNN, the accuracy of classification has reached a bottleneck, which restricts the implementation of relevant methods and systems. In order to solve the above challenges, in this study, a deep neural network architecture only based on self-attention mechanism, named *Vision Transformer*, is proposed to improve the accuracy of automatic classification. Experimental results on TrashNet dataset show that the proposed method can achieve the highest accuracy of 96.98%, which is better than the existing CNN-based method. By deploying the well-trained model on the server and using a portable device to take pictures of waste in order to upload to the server, automatic waste classification can be expediently realized on the portable device, which broadens the scope of application of automatic waste classification and is of great significance with respect to resource conservation and recycling.

Keywords: waste classification; automatic recycling; deep neural network; self-attention; portable device

Citation: Huang, K.; Lei, H.; Jiao, Z.; Zhong, Z. Recycling Waste Classification Using Vision Transformer on Portable Device. *Sustainability* **2021**, *13*, 11572. https://doi.org/10.3390/su132111572

Academic Editor: Ashutosh Tiwari

Received: 5 September 2021
Accepted: 13 October 2021
Published: 20 October 2021

1. Introduction

In the era of mass production and mass consumption, waste disposal has become an important international issue. The most basic and time-consuming task in waste disposal is waste classification, such as glass, paper, cardboard, plastic and metal, which determines the recycling percentage of recyclable waste and will significantly affect the efficiency of subsequent treatment [1]. Previous waste classification relied on a large amount of manual waste sorting that was not only high cost and inefficient but may also harm the health of operators because of toxic waste [2]. Moreover, with the rapid increase in the amount of urban waste, some inexperienced sorting workers are recruited to supplement the shortage of manpower [3]. The combination of these factors renders waste recycling an increasingly serious and urgent challenge [4]. Therefore, a waste classification method that can be automatically, quickly and easily deployed at waste sorting stations is urgently needed by city managers, governments and non-profit organizations [5].

With the development of robots, artificial intelligence and automation technologies, automatic waste classification, especially visual guidance automatic classification, has become mainstream in waste recycling. Methods based on deep learning [6–8], especially the convolutional neural network (CNN) [9–11], have further improved the accuracy of waste classification and the efficiency of resource recycling. In essence, CNN is a hierarchical data representation model. High-level feature representation depends on lower-level

feature representation, and features with higher-level semantic information are abstractly extracted step-by-step from shallow to deep representation. Each layer of the network learns its own simple representation, but through the connection, it finally forms a powerful feature representation ability. Although CNN has some advantages such as using convolution kernels or filters that constantly extract features, the receptive field should be able to theoretically cover it entirely, etc. However, many studies [12,13] have shown that the actual feeling towards this approach is that it is far less than a solid theory, and it is not ideal for taking full use of features of the context information. Although this can be solved by stacking convolution layers, it will bring about high computational cost and render it difficult for the model to converge during training, which goes against the original intention of using CNN [14]. The accuracy rate of CNN method has been pushed to over 95%, but due to the limitation of global information, it has reached a bottleneck in waste classification. However, considering the amount of waste produced globally every year, even a small improvement will bring about huge economic and environmental benefits, so it is of great significance to break through the bottleneck of CNN and to improve the accuracy of waste classification.

In recent years, the *Transformer* model [15] has made remarkable achievements in machine translation and natural language processing. With the self-attention mechanism, the transformer can process and analyze all words in a text simultaneously, rather than in sequential order. This excellent mechanism enables the transformer model to capture global contextual information to build a distant dependency on the target and extract more powerful features that CNN does not have. Dosovitskiy et al. [16] migrated a transformer to the image classification task and proposed the vision transformer, which only relies on the self-attention mechanism to achieve far more effects than CNN. However, when the waste classification system is deployed in the waste sorting station, a problem that cannot be ignored is the deployability of the system. Most of the existing studies deployed the CNN model or Transformer model on high-performance computing equipment, and it is extremely costly and infeasible to equip each waste sorting station with high-performance computing equipment. In addition, the environment of the waste sorting station is relatively fixed, and the consumption of computing resources can be reduced by reducing the complexity of the background, and the model deployed in the cloud can be accessed through portable devices for classification.

Motivated by the above-mentioned problems, a novel method based on vision transformer is proposed to improve the accuracy of waste classification and can be easily deployed on portable device, which improves the efficiency of resource conservation and recycling. The input image captured by the portable device was uploaded to a cloud server and fed into the vision transformer model that trained on TrashNet. The classification results were real-time presented on a portable device in order to facilitate the recycling of resources. The contributions of this study can be summarized as follows:

- An automatic waste classification method based on vision transformer is proposed to improve the efficiency of resource recycling;
- Experiments show that the proposed method outperforms the existing methods;
- The trained model is deployed on cloud server for real-time and convenient waste classification on portable devices.

2. Materials and Methods

In this section, the overall scheme of the proposed method is first illustrated. The model architecture of vision transformer is then described in detail to explain why it performs better than CNN. Finally, the method for deploying the portable device in the cloud is elucidated.

2.1. Overall Scheme

The overall scheme of the proposed method is illustrated in Figure 1.

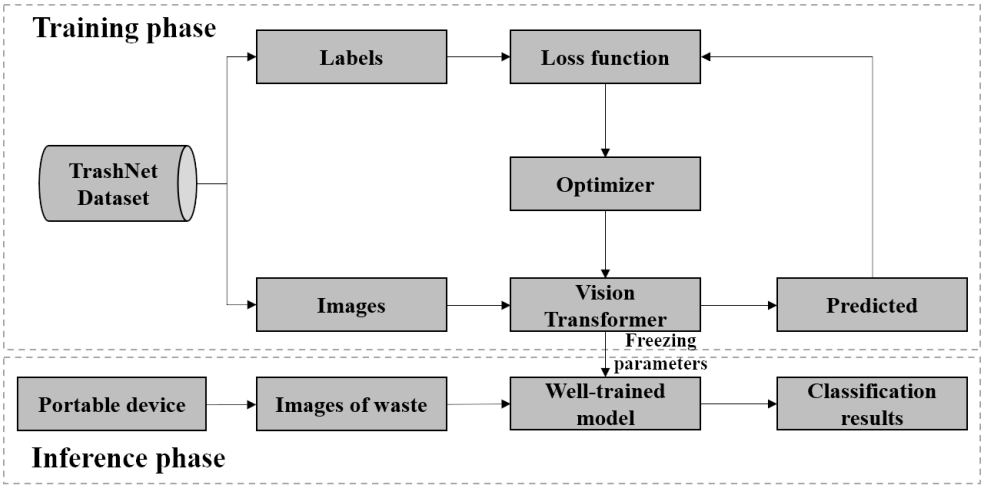

Figure 1. Overall scheme of the proposed method.

The entire method consists of two phases: training and inference. In the training phase, the labeled images in the TrashNet dataset [17] are utilized to train the parameters of the vision transformer model under the guidance of the loss function. After that, by freezing the network parameters, the well-trained model is exploited in the inference phase to process the images captured by the portable device and achieve automatic waste classification. It should be noted that other existing datasets, such as TACO [18], MJU [19] and OpenLitterMap [20], are not suitable for this study. The main reasons include the following: (1) TACO and OpenLitterMap currently induce more waste in natural scenes, which does not match the background of the waste sorting station that this research focuses on. (2) Although the data types of MJU and TrashNet are very similar, the MJU dataset is aimed at detection tasks. In order to improve the deployability of the system, the TrashNet dataset with a relatively fixed background is used in this study to complete an image classification task. We have added this part of the content in the revised manuscript.

2.2. Vision Transformer Model

Natural language is a series of words arranged in a specific order. The task of natural language processing and machine translation is to construct a network model to extract the features of the sequential data so as to complete the tasks of classification or prediction. Analogously, an image can be thought of as an order of pixels in two-dimensional space. Therefore, the structure of the transformer model can be applied to image classification tasks without hindrance, that is, vision transformer. The structure of vision transformer is shown in Figure 2.

Patches. Assume that the image of waste to be classified is $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$, where H and W represent the height and width of the image, respectively, and C indicates the channels of the image. First, the input image is divided into several patches $\mathbf{x}_p \in \mathbb{R}^{N \times \left(P^2 \cdot C\right)}$ of the same size to form a language-like sequential data. By dividing the image here instead of directly using pixels as the input sequence, it can avoid the subsequent pixel-level attention calculation, which takes up a lot of computing resources and storage space.

Linear projection. When the resolution of the images or the size of the patches changes, the number of patches or the number of pixels will change. A trainable linear projection is adopted to map the patches into the D-dimensional space to enable processing of patches of different dimensions and quantities. In this process, D is usually much less than $H \times W$, so the realization of linear projection improves the generalization of the model while reducing the dimension of the input image.

Position embedding. The position information of each pixel in the image is important for the waste classification. However, after being divided into several patches, only the local position information in the patch is contained. It is necessary to attach the corresponding position information when it is input into the model in the form of sequence after linear

projection. Different from the trigonometric function position coding in the transformer [15], this study adopts a trainable random initialization position coding method, which has been proved to be effective in [16]. The position information is added in the form of vectors to the results of patches after linear projection, which can be expressed as follows:

$$\mathbf{z}_0 = \left[\mathbf{x}_{\text{class}}; \mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \cdots; \mathbf{x}_p^N\mathbf{E}\right] + \mathbf{E}_{pos}, \quad \mathbf{E} \in \mathbb{R}^{\left(P^2 \cdot C\right) \times D}, \mathbf{E}_{pos} \in \mathbb{R}^{(N+1) \times D} \tag{1}$$

where $\mathbf{x}_{\text{class}}$ is the one-hot encoding of the class label, $\mathbf{x}_p^i$ represents the ith patch, $\mathbf{E}$ denotes the encoding of the input patches and $\mathbf{E}_{pos}$ indicates the position embedding.

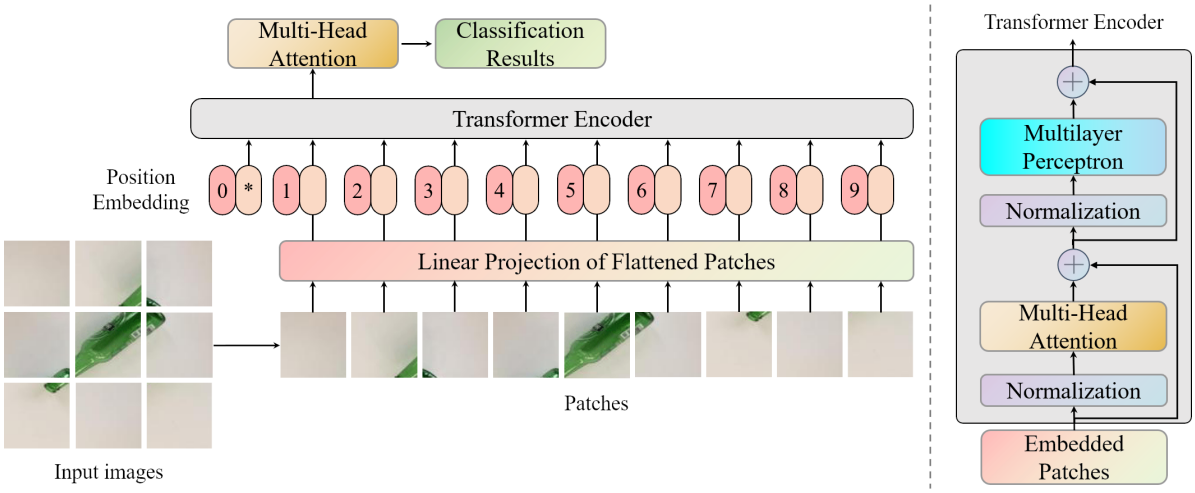

Figure 2. Structure of vision transformer.

Transformer encoder. The transformer's impressive performance in natural language processing and machine translation tasks is mainly due to its self-attention mechanism. The essence of a self-attention mechanism is to establish weight parameters to represent the correlation between each part of the input sequence and the final result so that the model has the ability to reallocate resources according to the importance of the object to be desired. In the task of image classification, vision transformer adopts a similar encoding method to establish the association between various parts of image features and the final classification results so as to obtain the global information of the image. The structure of the transformer encoder is shown on the right side of Figure 2, which is mainly composed of a multi-headed self-attention (MSA) block and multi-layer perceptron (MLP) block. The input is normalized by layer normalization before each block, and the residual connection is exploited after each block in order to reduce the difficulty of model training [21,22].

In MSA, Reference [15] constructed standard qkv self-attention methods, where q represents the feature to be queried, k indicates the feature to be matched and v is the correlation measure value of the feature to be queried and matched. qkv can be obtained by applying linear transformations to z in Equation (1), which can be expressed as follows:

$$\begin{aligned} q &= W^q z \\ k &= W^k z \\ v &= W^v z \end{aligned} \tag{2}$$

where W^q, W^k and W^v are the weights of linear transformations.

The attention weight A can be obtained through the softmax function after the matching degree between q and k is normalized in the D-dimensional space, which can be calculated as follows.

$$A = \text{softmax}\left(\mathbf{q}\mathbf{k}^\top / \sqrt{D_h}\right) \tag{3}$$

Finally, the weight of self-attention SA can be expressed as the result of the weight of attention A and the matching value of the correlation measure value v.

$$\mathrm{SA}(\mathrm{z}) = A\mathrm{v} \tag{4}$$

Varying random initialization mapping weights of the linear transformation W^q, W^k and W^v can map the input vector to different subspaces, which allows the model to understand the input sequence from different perspectives. Therefore, the combination of several attentions at the same time may be better than a single attention. This method of simultaneously calculating multiple self-attentions is named MSA. MSA is an extension of SA and can be obtained by weighing the values of k multiple SAs, which can be expressed as follows:

$$\mathrm{MSA}(\mathrm{z}) = [\mathrm{SA}_1(z); \mathrm{SA}_2(z); \cdots; \mathrm{SA}_k(z)]\mathbf{U}_{msa} \tag{5}$$

where $\mathbf{U}_{msa}$ represents the weight matrix of the weighted summation.

The MLP block consists primarily of a multi-layer perceptron. The encoding process in the entire transformer encoder can be represented as follows:

$$\mathbf{z}'_\ell = \mathrm{MSA}(\mathrm{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \quad \ell = 1 \ldots L \tag{6}$$

$$\mathbf{z}_\ell = \mathrm{MLP}\big(\mathrm{LN}\big(\mathbf{z}'_\ell\big)\big) + \mathbf{z}'_\ell, \quad \ell = 1 \ldots L \tag{7}$$

$$\mathbf{y} = \mathbf{L}\mathrm{N}\Big(\mathbf{z}_L^0\Big) \tag{8}$$

where L stands for Linear function, MSA and MLP are stacked together.

2.3. Deployed on the Cloud

Due the deep learning model requiring computing power and storage space (especially with respect to the dependence on a graphics processing unit (GPU)), it is almost impossible to classify waste using deep learning model on mobile devices. With the development of cloud server and 5G network, uploading the images of waste captured by portable device through mobile communication network and classifying waste with the help of cloud server is possible, which brings the dawn of automatic waste classification. Therefore, in this study, in order to make full use of the well-trained vision transformer model for waste classification, a lightweight web-side server is built and deployed on the cloud in order to facilitate waste classification on portable devices.

Based on an open source lightweight web application framework *Flask* [23,24] written in Python, this study constructs a waste classification system that can be deployed on a cloud server for portable device. Flask has two outstanding advantages: (1) excellent extensibility. Flask does not have a built-in abstraction layer for database processing nor does it form validation support. Instead, Flask supports extensions to add such functionality to your application, which allows it to be easily modified based on new features and added functionality requested by users. (2) Flask is also extremely robust. Flask structure is categorized into two parts: static files and template files. The static files contain files used to process the uploaded data and display them on the website; on the other hand, the template files contain the display template of the classification results. This modular design ensures that the system is well equipped to deal with errors during execution.

2.4. Experimental Details

Dataset. To explore the performance of the vision transformer model, all experiments were implemented on TrashNet dataset, which contains a total of 2527 images of waste. These images include six of the most common waste in daily life: glass, paper, cardboard, plastic, metal and trash. All images were taken using portable devices, including Apple iPhone 7 Plus, Apple iPhone 5S and Apple iPhone SE. The images in the dataset are divided

into train set, validation set and test set according to the official given ratios of 70%, 13% and 17%.

Training and Fine-tuning. Considering the limited number of waste images in the dataset, it is easy to cause under-fitting or it is difficult to converge when directly training the model with randomly initialized parameters. Therefore, in this study, the transfer learning was adopted: The network parameters of the vision transformer model was pre-trained on the ImageNet [25], and the images in the TrashNet dataset were utilized to fine tune the model. During the training, Adam [26] was exploited as the optimizer of the model with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and applied a high weight decay of 0.1. The initial learning rate was set to 10^{-4}, the batch size for the training stage was set to 512 and 64 for evaluation and test stage. A linear learning rate warm up and decay was adopted in the first 500 steps. All the experiments were conducted on an Intel i7-6700 CPU at 4.0 GHz with 16 GB RAM and 8 Nvidia P100 GPU with 16 GB memory. The programming language was Python 3.6, and the integrated development environment was Anaconda 3.

Metrics. When it is actually deployed on portable devices for waste classification, the most important evaluation metrics include accuracy and inference speed. To explore the scalability of the model, the accuracy rate on the TrashNet test set was utilized as the core evaluation metric of the model performance. The average inference time of the model on each image was measured to show the inference speed of waste classification.

Test platform. After the model was constructed, an Android phone was used as a test platform to verify that the model was not affected by the phone system and model. The test platform is Xiaomi 10, the CPU is a Snapdragon 870 chip with a processing frequency of 3.2 GHz, and it has a 2-megapixel high-definition camera.

3. Experimental Results and Discussion

In this chapter, the details of the experiment are first explained in detail in order to illustrate the conditions under which the experiment was performed. The performance of the vision transformer model is then evaluated numerically and compared with existing CNN-based and traditional machine learning methods. Next, the influencing factors of the model performance are discussed seriatim to further demonstrate the advantages of the vision transformer model in waste classification. Finally, the self-attention mechanism of the vision transformer model is visualized in order to further demonstrate the advantages of the vision transformer in waste classification.

The method based on vision transformer is compared with existing CNN-based methods and traditional machine learning methods, respectively. Among them, the RecycleNet proposed by [27] is the current state of the art, which has achieved the highest accuracy of 94.2% on the test set. CNN-based methods, including ResNet50 [28], HOG + CNN (Histogram of Oriented Gradient) [28], Simple CNN [28], DenseNet121 [29], DenseNet169 [29], RecycleNet [27], ResNet + SE (Squeeze-and-Excitation) [30] and Optimized DenseNet121 [31], etc., are representative methods in a certain period of waste classification task. Traditional machine learning methods, such as SVM (Support Vector Machines) + HOG [28] and SIFT (Scale-Invariant Feature Transform) + SVM [17], are the most representative methods before the large-scale application of CNN.

3.1. Performance of Vision Transformer Model

In this study, the proposed vision transformer model is quantitatively evaluated and compared with existing CNN-based and machine learning-based methods to demonstrate its performance in waste classification tasks. Table 1 demonstrates the comparative experimental results. Compared with the previous state of the art, the vision transformer model achieves an improvement of nearly 2% in classification accuracy. As can be observed from the results in Table 1, traditional machine learning-based methods adopt manually selected image features such as HOG or SIFT, which is difficult to effectively represent the features of waste images, and shallow classifiers such as SVM are difficult in terms of effectively realizing classification in feature space. As a result, the classification accuracy

of traditional machine learning-based methods is often low, which limits its application in real classification scenarios. The CNN-based method has significant improvements in classification accuracy compared with the machine learning-based method, which is mainly attributed to the powerful feature extraction ability of CNN.

According to the research of [29,30,32], with the increase in depth (i.e., the number of layers of the CNN) of CNN model, more and more convolutional layers are stacked to improve the representation ability of the model, which can theoretically improve the classification accuracy of the model. However, from [29]'s study, it can be found that from DenseNe121 to DenseNe169, although the depth of CNN increases, the performance of CNN model does not change significantly. To some extent, this reflects the limitations of CNN itself. In other words, the failure of CNN to extract global features mentioned in [16] restricts the further improvement of CNN's performance. Moreover, by comparing the results of ResNet50 in [28] and the results of ResNet50+CBAM (convolutional block attention module) proposed by [32], adding an attention mechanism to the CNN makes the model show a stronger classification effect, which is mainly due to the attention mechanism assuming the function of adaptively assigning weights in the model so that the model can learn the important features more effectively. The attention mechanism will be visualized and discussed in detail later.

Table 1. Results of comparison experiment.

Methods	Accuracy (%)	Total Images	Inference Time per Image (ms)	Epoch
HOG + SVM [28]	23.51	2276 train images, 251 test images [1]	-	-
SIFT + SVM [17]	63	1769 train images, 758 test images [2]	-	-
Simple CNN [28]	79.49	2276 train images, 251 test images	-	40
HOG CNN [28]	81.53	2276 train images, 251 test images	-	40
Resnet50 w/o pre-train [28]	58.70	2276 train images, 251 test images	-	40
Resnet50 [28]	91.40	2276 train images, 251 test images	-	40
DenseNet121 [29]	95	2527 (70% training, 17% testing, 13% valid)	-	10 + 100
DenseNet169 [29]	95	2527 (70% training, 17% testing, 13% valid)	-	7 + 120
Inception V4 [29]	94	2527 (70% training, 17% testing, 13% valid)	-	7 + 120
Inception V4 [29]	89	2527 (70% training, 17% testing, 13% valid)	-	10 + 200
MobileNet [29]	84	2527 (70% training, 17% testing, 13% valid)	-	10 + 200
ResNet18 + SE [30]	87.70	2527 (70% training, 17% testing, 13% valid)	-	100
ResNet34 + SE [30]	88.86	2527 (70% training, 17% testing, 13% valid)	-	100
ResNet50 + SE [30]	91.88	2527 (70% training, 17% testing, 13% valid)	-	100
ResNet18 + CBAM [32]	79.81	2527 (70% training, 17% testing, 13% valid)	-	100
ResNet34 + CBAM [32]	81.44	2527 (70% training, 17% testing, 13% valid)	-	100
ResNet50 + CBAM [32]	82.14	2527 (70% training, 17% testing, 13% valid)	-	100
ResNet18 + RecycleNet [27]	93.04	2527 (70% training, 17% testing, 13% valid)	231 [3]	100
ResNet34 + RecycleNet [27]	93.97	2527 (70% training, 17% testing, 13% valid)	352	100
ResNet50 + RecycleNet [27]	94.20	2527 (70% training, 17% testing, 13% valid)	366	100
Optimized DenseNet121 [31]	94.02	2276 train images, 251 test images	-	40
EfficientNet-B0 [33]	90.02	2527 (70% training, 17% testing, 13% valid)	-	-
EfficientNet-B1 [33]	91.53	2527 (70% training, 17% testing, 13% valid)	-	-
EfficientNet-V2 [34]	94.69	2527 (70% training, 17% testing, 13% valid)	-	-
Vision transformer w/o pre-train	89.06	2527 (70% training, 17% testing, 13% valid)	-	40
Vision transformer	96.98	2527 (70% training, 17% testing, 13% valid)	423	40

[1]: 9/10 of all images in the TrashNet dataset are used as the train set and 1/10 as the test set. [2]: Combine TrashNet's validation set with the test set for the model testing. [3]: This code is available at https://github.com/sangminwoo/RecycleNet (accessed on 5 January 2021).

Moreover, the classification speed of the model is also evaluated to gauge whether it can meet the requirements of the actual deployment. Different from [31], who uses training time as the evaluation metric, since the model is deployed on the cloud in practical application and the model parameters have been frozen, there is no need to retrain the model; thus, only inference time needs to be considered. As observed from the results in the penultimate column of Table 1, the vision transformer model is slightly slower than the

CNN-based methods [27] in the inference stage, with an average of 100 milliseconds (ms) more per image. Although time increased, the inferential speed is close to human response speed [35], which is sufficient to meet the daily waste classification task for deployment on portable devices.

3.2. Self-Attention Mechanism

The self-attention mechanism is the core reason why the vision transformer model performs so well. In order to demonstrate the role of self-attention mechanism in waste classification more intuitively, Figure 3 shows the visualization results of self-attention mechanism of six classes of waste. The brighter areas in the image represent the greater impact on the final classification results, while the darker areas represent the smaller impact on the results. As observed from the visualization results, after training, the self-attention mechanism assigns a higher weight to the areas of waste in the image, and it is similar to the manner human eyes focus on a specific area, which can significantly improve the performance of the model.

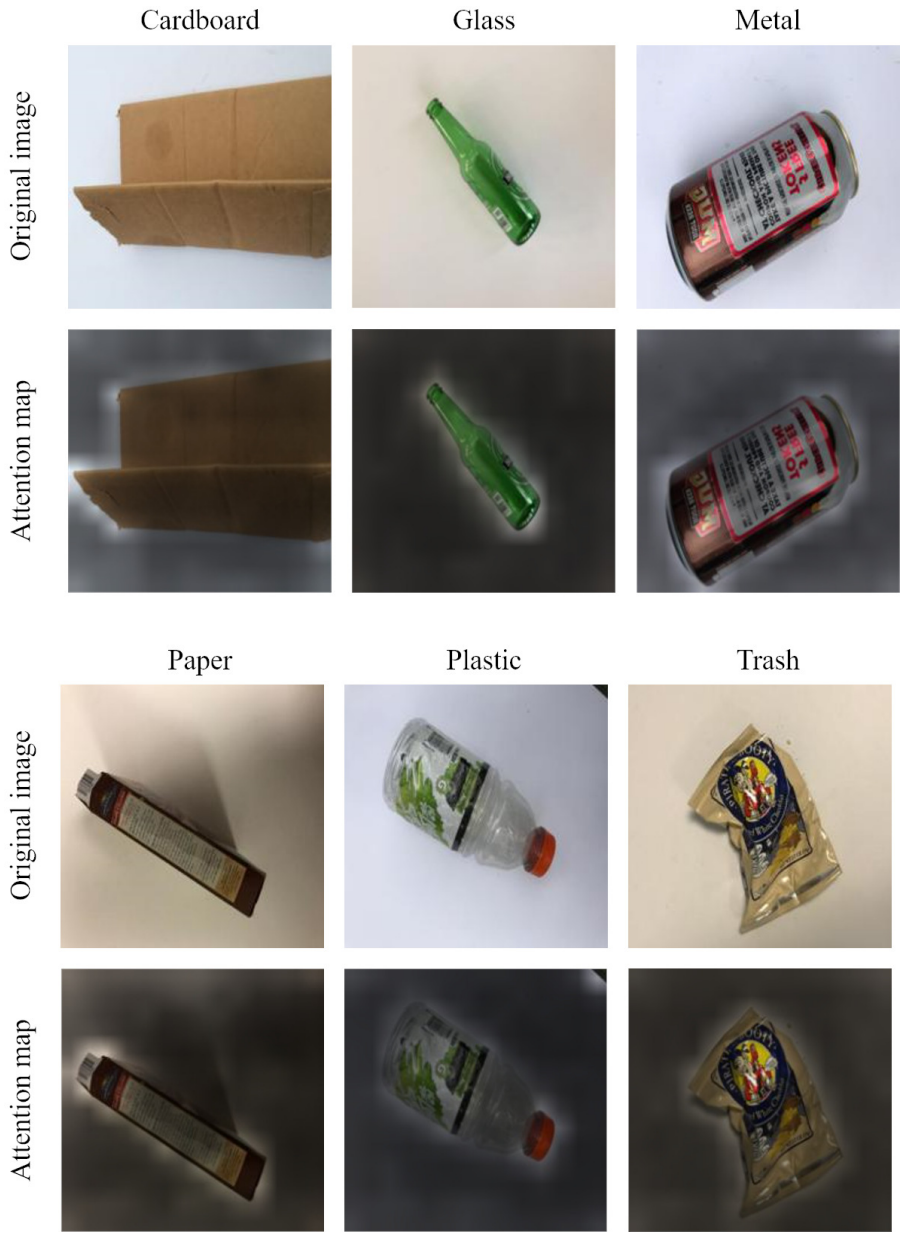

Figure 3. Visualization of self-attention mechanisms.

3.3. Influencing Factors of the Model Performance

Transfer learning. Since deep learning models usually contain tens of millions of parameters, training deep learning models requires huge computational resources or the use of huge datasets. Transfer learning can overcome this problem effectively, reduce the dependence on data quantity and make the model perform better. The last two rows of Table 1 show the performance of the vision transformer model with or without (w/o) transfer learning. It can be observed that pre-training on ImageNet results in a significant improvement of nearly 8% in model performance. This result also shows that the vision transformer model has good generalization, and some common features can be learned from large datasets such as ImageNet, and sufficient features can be extracted from small datasets such as TrashNet to distinguish different classes of waste.

Error analysis of the test set. In addition to considering the performance of the model under different conditions in Table 1 to quantitatively analyze the influencing factors of the model, the error classifications on the test set are also statistically analyzed to explore the reasons for the failure of the model. Table 2 analyzes the classification results of vision transformer on the test set, and some examples of the classification results are visualized in Figure 4 to provide more intuitive understanding of the model performance.

Table 2. Confusion matrix of the classification results.

Predicted \ Ground Truth	Cardboard	Glass	Metal	Paper	Plastic	Trash
Cardboard	69	0	0	0	0	0
Glass	0	78	0	0	2	0
Metal	1	1	67	0	0	0
Paper	0	0	1	104	1	0
Plastic	0	3	0	4	71	0
Trash	0	0	0	0	0	29

Figure 4. Examples of misclassified wastes.

It can be found from the visualization results that incorrect classification is often caused by the inability to perceive the material of the waste or because one kind of waste is covered by another kind of waste. The cardboard in Figure 4a is coated with a layer of reflective paper, which presents a sheen similar to metallic reflection under light, thus resulting in misclassification. The reflective part of the cardboard occupies high weight in the attention map, which confirms that the misclassification of the model may be caused by the misdirection of the reflection. The attention maps in Figure 4b,c shows that the model

can accurately find the areas that need attention. However, it is difficult for people to distinguish plastic or glass bottles (i.e., Figure 4b) and plastic or paper labels (i.e., Figure 4c) even if they cannot perceive the material. This may also be the reason for the incorrect classification of the current vision transformer model. The metal can in Figure 4d has a paper label on its surface, which also leads the model to mistakenly believe that the waste is paper rather than metal.

3.4. Evaluation on Different Datasets

In order to evaluate the generalization of the Vision Transformer model adopted in this study, in addition to using TrashNet dataset, this study also evaluated the GLASSENSE-VISION dataset [36] and Drinking Waste Classification dataset [37], respectively. As shown in Figure 5, the model reached 100% accuracy on the test set of the GLASSENSE-VISION dataset after 80 steps, and it achieved 99.29% accuracy on the test set of the Drinking Waste Classification dataset after 180 steps, as shown in Figure 6.

3.5. Application on Portable Device

Figure 7 shows the effect of the proposed vision transformer-based method deployed on the cloud for waste classification. Figure 7a is the initial interface, which records the amount of various types of waste that has been disposed of. Figure 7b,c show some examples of automatic waste classification, respectively. The portable device used is an Android mobile phone.

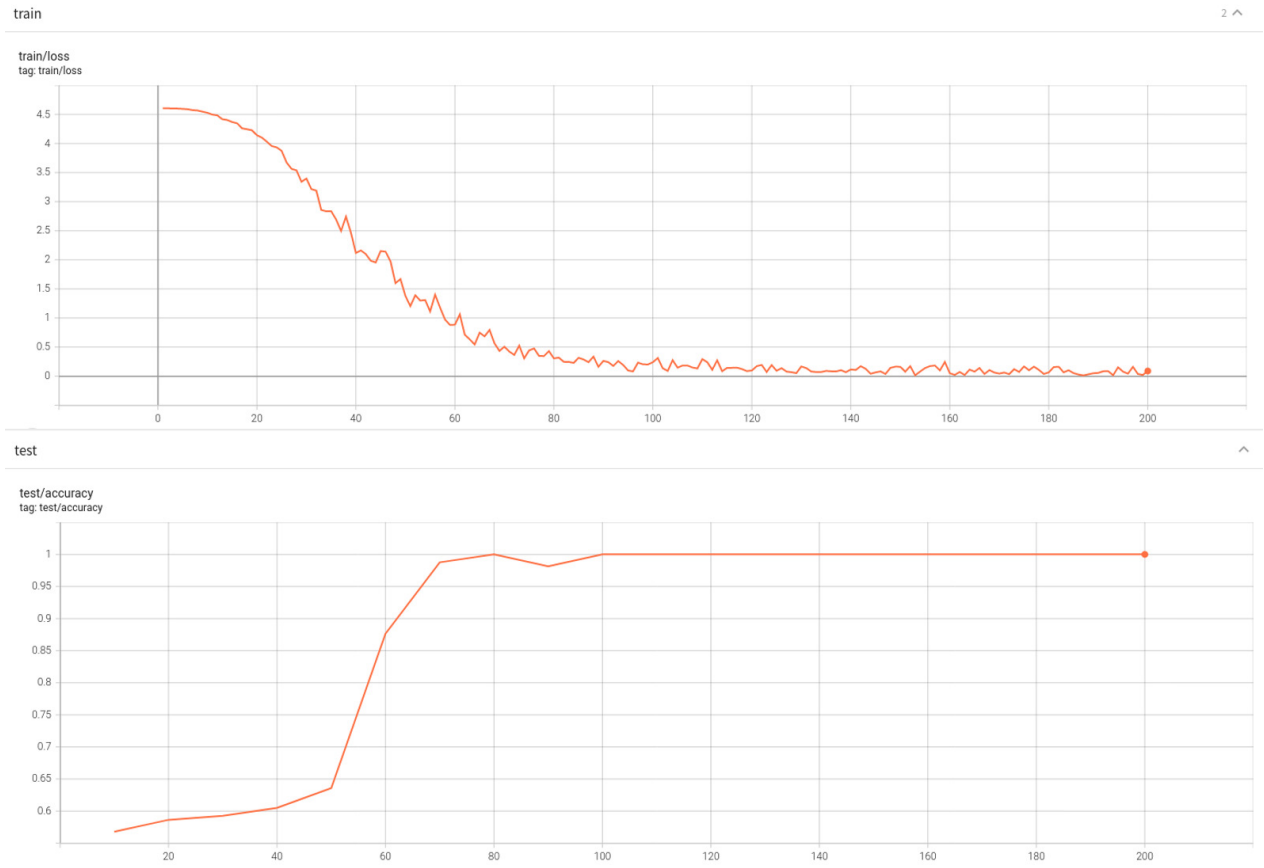

Figure 5. Model performance on GLASSENSE-VISION dataset.

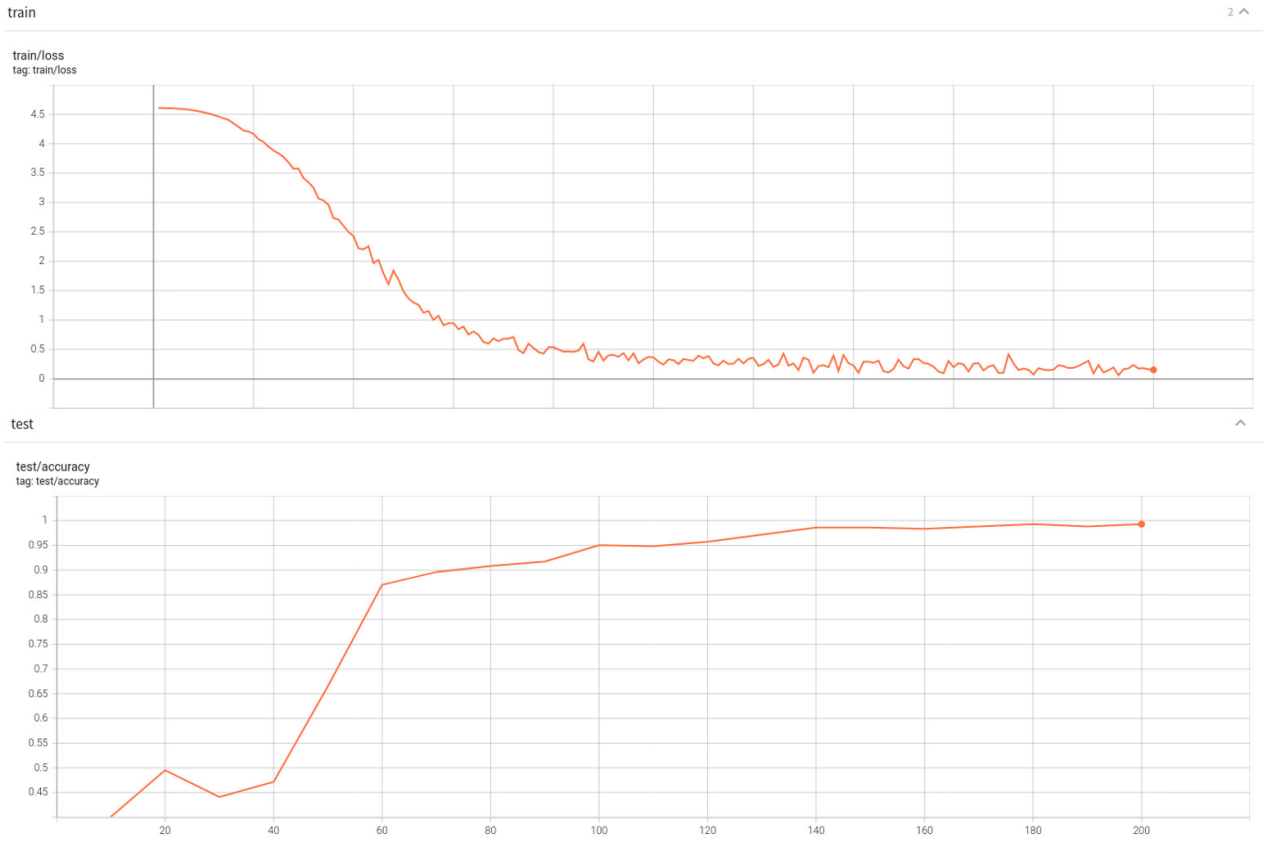

Figure 6. Model performance on Drinking Waste Classification dataset.

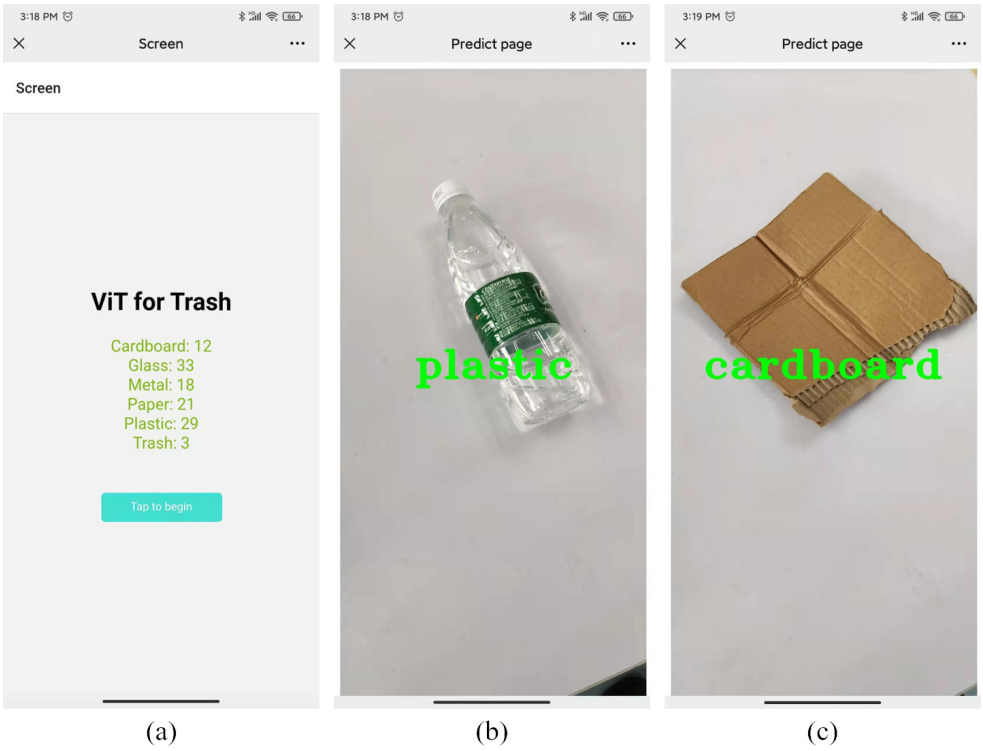

Figure 7. Examples of vision transformer-based waste classification on a portable device. (**a**) is the main interface of the mobile terminal, (**b**) is the recognition result of plastic bottles, (**c**) is the recognition result of cardboard.

4. Conclusions

A waste classification method based on vision transformer is proposed and deployed on the cloud for portable device in this study, which can reduce labor costs and improve waste disposal efficiency and is of great significance for resource conservation and recycling. Since CNN has insufficient ability to pay attention to global information and reaches a bottleneck with respect to accuracy in waste classification, the vision transformer model built in this study can effectively overcome this disadvantage by using the self-attention mechanism to adaptively allocate weight between each part of waste image and the final classification result. By pre-training the model on ImageNet and using images in the TrashNet dataset to fine-tune network parameters, the vision transformer achieves an accuracy rate of 96.98%. Furthermore, by visualizing the weight of attention of different parts in the image, the core parts affecting the classification results are analyzed intuitively, which also explains to some extent why vision transformer outperforms the existing CNN-based method. Furthermore, the analysis of various influencing factors on the model performance shows that transfer learning is very effective and even essential in the construction of waste classification model, and misclassification is usually caused by the reflection or the inability to perceive the material or the waste itself is covered by another kind of waste. Finally, the well-trained model has been deployed on the cloud and been applied to portable devices in order to achieve more convenient waste classification. Further studies should be conducted to find out the clearer causes of misclassification and carry out multiple target detection of wastes. In addition, in order to promote more efficient resource conservation and recycling, further optimization of the model is needed in order to accelerate the inference speed of the vision transformer-based method. Similarly, this work focuses more on image classification tasks. For waste in natural scenes, it is necessary to consider object detection or segmentation methods in order to obtain more accurate waste locations.

Author Contributions: Conceptualization, Z.J.; methodology, K.H. and Z.J.; validation, H.L. and Z.Z.; investigation, H.L., Z.J. and Z.Z.; writing—original draft preparation, K.H.; writing—review and editing, H.L. and Z.Z.; funding acquisition, Z.J. and Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the financial support from GDAS' Project of Science and Technology Development (Grant No. 2021GDASYL-20210103090) and GDAS' Project of Science and Technology Development (Grant No. 2019GDASYL-0502007, 2020GDASYL-20200302015).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: All authors extend their sincerest thanks to the reviewers. We thank Yingjie Cai from the Department of Electrical Engineering, Chinese University of Hong Kong, for her guidance.

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

References

1. Wang, Z.; Peng, B.; Huang, Y.; Sun, G. Classification for plastic bottles recycling based on image recognition. *Waste Manag.* **2019**, *88*, 170–181. [CrossRef] [PubMed]
2. Seike, T.; Isobe, T.; Harada, Y.; Kim, Y.; Shimura, M. Analysis of the efficacy and feasibility of recycling PVC sashes in Japan. *Resour. Conserv. Recycl.* **2018**, *131*, 41–53. [CrossRef]
3. Borowski, P.F. Environmental pollution as a threats to the ecology and development in Guinea Conakry. *Ochr. Środowiska Zasobów Nat.* **2017**, *28*, 27–32. [CrossRef]
4. Zelazinski, T.; Ekielski, A.; Tulska, E.; Vladut, V.; Durczak, K. Wood dust application for improvment of selected properties of thermoplastic starch. Inmateh. *Agric. Eng.* **2019**, *58*, 37–44.
5. Żelaziński, T. Properties of biocomposites from rapeseed meal, fruit pomace and microcrystalline cellulose made by press pressing: Mechanical and physicochemical characteristics. *Materials* **2021**, *14*, 890. [CrossRef]

6. Vo, A.H.; Vo, M.T.; Le, T. A novel framework for trash classification using deep transfer learning. *IEEE Access* **2019**, *7*, 178631–178639. [CrossRef]
7. AR, A.R.; Hasan, S.; Mahmood, B. Automatic waste detection by deep learning and disposal system design. *J. Environ. Eng. Sci.* **2019**, *15*, 38–44.
8. Chu, Y.; Huang, C.; Xie, X.; Tan, B.; Kamal, S.; Xiong, X. Multilayer hybrid deep-learning method for waste classification and recycling. *Comput. Intell. Neurosci.* **2018**, *2018*, 5060857. [CrossRef]
9. Tiyajamorn, P.; Lorprasertkul, P.; Assabumrungrat, R.; Poomarin, W.; Chancharoen, R. Automatic Trash Classification using Convolutional Neural Network Machine Learning. In Proceedings of the 2019 IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM), Bangkok, Thailand, 18–20 November 2019; pp. 71–76.
10. Yu, Y. A Computer Vision Based Detection System for Trash Bins Identification during Trash Classification. *J. Phys. Conf. Ser.* **2020**, *1617*, 012015.
11. Ruiz, V.; Sánchez, Á.; Vélez, J.F.; Raducanu, B. Automatic image-based waste classification. In Proceedings of the International Work-Conference on the Interplay between Natural and Artificial Computation, Almeria, Spain, 3–7 June 2019; pp. 422–431.
12. Chao, Y.W.; Vijayanarasimhan, S.; Seybold, B.; Ross, D.A.; Deng, J.; Sukthankar, R. Rethinking the faster r-cnn architecture for temporal action localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1130–1139.
13. Liu, L.; Wu, F.X.; Wang, Y.P.; Wang, J. Multi-Receptive-Field CNN for Semantic Segmentation of Medical Images. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 3215–3225. [CrossRef]
14. Liang, J.; Zhang, T.; Feng, G. Channel Compression: Rethinking Information Redundancy Among Channels in CNN Architecture. *IEEE Access* **2020**, *8*, 147265–147274. [CrossRef]
15. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
16. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 $\times$ 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
17. Yang, M.; Thung, G. Classification of trash for recyclability status. *CS229 Proj. Rep.* **2016**, *2016*, 1–6.
18. Proença, P.F.; Simões, P. TACO: Trash Annotations in Context for Litter Detection. *arXiv* **2020**, arXiv:2003.06975.
19. Wang, T.; Cai, Y.; Liang, L.; Ye, D. A Multi-Level Approach to Waste Object Segmentation. *Sensors* **2020**, *20*, 3816. [CrossRef]
20. Lynch, S. OpenLitterMap. com–open data on plastic pollution with blockchain rewards (littercoin). *Open Geospat. Data Softw. Stand.* **2018**, *3*, 1–10. [CrossRef]
21. Wang, Q.; Li, B.; Xiao, T.; Zhu, J.; Li, C.; Wong, D.F.; Chao, L.S. Learning deep transformer models for machine translation. *arXiv* **2019**, arXiv:1906.01787.
22. Baevski, A.; Auli, M. Adaptive input representations for neural language modeling. *arXiv* **2018**, arXiv:1809.10853.
23. Aslam, F.A.; Mohammed, H.N.; Lokhande, P. Efficient way of web development using python and flask. *Int. J. Adv. Res. Comput. Sci.* **2015**, *6*, 54–57.
24. Mufid, M.R.; Basofi, A.; Al Rasyid, M.U.H.; Rochimansyah, I.F. Design an mvc model using python for flask framework development. In Proceedings of the 2019 International Electronics Symposium (IES), Surabaya, Indonesia, 27–28 September 2019; pp. 214–219.
25. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE conference on computer vision and pattern recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
26. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
27. Bircanoğlu, C.; Atay, M.; Beşer, F.; Genç, Ö.; Kızrak, M.A. RecycleNet: Intelligent waste sorting using deep neural networks. In Proceedings of the 2018 Innovations in Intelligent Systems and Applications (INISTA), Thessaloniki, Greece, 3–5 July 2018; pp. 1–7.
28. Meng, S.; Chu, W.T. A Study of Garbage Classification with Convolutional Neural Networks. In Proceedings of the 2020 Indo–Taiwan 2nd International Conference on Computing, Analytics and Networks (Indo-Taiwan ICAN), Rajpura, India, 7–15 February 2020; pp. 152–157.
29. Aral, R.A.; Keskin, Ş.R.; Kaya, M.; Hacıömeroğlu, M. Classification of trashnet dataset based on deep learning models. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 2058–2062.
30. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
31. Mao, W.L.; Chen, W.C.; Wang, C.T.; Lin, Y.H. Recycling waste classification using optimized convolutional neural network. *Resour. Conserv. Recycl.* **2021**, *164*, 105132. [CrossRef]
32. 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.
33. Huynh, M.H.; Pham-Hoai, P.T.; Tran, A.K.; Nguyen, T.D. Automated Waste Sorting Using Convolutional Neural Network. In Proceedings of the 2020 7th NAFOSTED Conference on Information and Computer Science (NICS), Ho Chi Minh City, Vietnam, 26–27 November 2020; pp. 102–107.
34. Tan, M.; Le, Q.V. Efficientnetv2: Smaller models and faster training. *arXiv* **2021**, arXiv:2104.00298.

35. Thorpe, S.; Fize, D.; Marlot, C. Speed of processing in the human visual system. *Nature* **1996**, *381*, 520–522. [CrossRef] [PubMed]
36. Sosa-Garcia, J.; Odone, F. "Hands on" visual recognition for visually impaired users. *ACM Trans. Access. Comput. (TACCESS)* **2017**, *3*, 1–30. [CrossRef]
37. Kaggle. Drinking Waste Classification Dataset. 2020. Available online: https://www.kaggle.com/arkadiyhacks/drinking-waste-classification (accessed on 12 July 2020).