processes

MDPI

Article

An Efficient Multi-Label Classification-Based Municipal Waste Image Identification

Rongxing Wu [1], Xingmin Liu [2], Tiantian Zhang [3], Jiawei Xia [4], Jiaqi Li [5], Mingan Zhu [6,*] and Gaoquan Gu [7,*]

[1] School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China; 17787173724@163.com
[2] School of Intelligent and Information Engineering, Shandong University of Traditional Chinese Medicine, Qingdao 266112, China; 15689403550@163.com
[3] School of Computer Science, Northwestern Polytechnical University, Xi'an 710129, China; zhangtiantian@nwpu.edu.com
[4] School of International Education, Beijing University of Chemical Technology, Beijing 100029, China; 2022090129@buct.edu.cn
[5] School of Information and Electrical Engineering, China Agricultural University, Beijing 100193, China; lijiaqi2021@cau.edu.cn
[6] School of Management, Harbin Institute of Technology, Harbin 150001, China
[7] College of Naval Architecture and Ocean Engineering, Naval University of Engineering, Wuhan 430033, China
* Correspondence: 2021112778@stu.hit.edu.cn (M.Z.); davidgu3000@126.com (G.G.)

Abstract: Sustainable and green waste management has become increasingly crucial due to the rising volume of waste driven by urbanization and population growth. Deep learning models based on image recognition offer potential for advanced waste classification and recycling methods. However, traditional image recognition approaches usually rely on single-label images, neglecting the complexity of real-world waste occurrences. Moreover, there is a scarcity of recognition efforts directed at actual municipal waste data, with most studies confined to laboratory settings. Therefore, we introduce an efficient Query2Label (Q2L) framework, powered by the Vision Transformer (ViT-B/16) as its backbone and complemented by an innovative asymmetric loss function, designed to effectively handle the complexity of multi-label waste image classification. Our experiments on the newly developed municipal waste dataset "Garbage In, Garbage Out", which includes 25,000 street-level images, each potentially containing up to four types of waste, showcase the Q2L framework's exceptional ability to identify waste types with an accuracy exceeding 92.36%. Comprehensive ablation experiments, comparing different backbones, loss functions, and models substantiate the efficacy of our approach. Our model achieves superior performance compared to traditional models, with a mean average precision increase of up to 2.39% when utilizing the asymmetric loss function, and switching to ViT-B/16 backbone improves accuracy by 4.75% over ResNet-101.

Keywords: multi-label image classification; waste management; Query2Label; Vision Transformer; asymmetric loss function

check for updates

Citation: Wu, R.; Liu, X.; Zhang, T.; Xia, J.; Li, J.; Zhu, M.; Gu, G. An Efficient Multi-Label Classification-Based Municipal Waste Image Identification. *Processes* **2024**, *12*, 1075. https://doi.org/10.3390/pr12061075

Academic Editor: Aldo Muntoni

Received: 14 April 2024
Revised: 4 May 2024
Accepted: 23 May 2024
Published: 24 May 2024

1. Introduction

With the rapid development of artificial intelligence (AI) technology, its application across various fields such as medical diagnostics [1], autonomous driving [2], language translation [3], and image recognition [4] has become increasingly widespread, significantly advancing and innovating in these areas. AI technology can process and analyze vast amounts of data, extracting valuable information to assist in decision making and prediction, thereby enhancing work efficiency and accuracy. Particularly in the field of image recognition, the integration of deep learning and machine vision technologies has enabled AI to achieve and even surpass human recognition capabilities in aspects such as facial recognition and scene understanding, advancing the field of intelligent image processing.

Among its numerous applications, image-based waste classification [5] has received considerable attention in recent years. As urbanization accelerates and populations grow, the issue of urban waste becomes increasingly severe, making the effective classification and recycling of waste an urgent problem to be addressed. The real-time and efficient classification of municipal waste can enhance recycling efforts and strengthen waste management, while also ensuring the cleanliness of urban environments and public health.Traditional methods of municipal waste classification primarily rely on manual sorting, which is labor-intensive and prone to errors [6]. Techniques such as magnetic separation and eddy current separation are commonly used for segregating metal components, but these do not address the sorting of nonmetallic components [7]. Techniques like air classification and screening have been employed to separate waste based on size and weight, yet these methods lack the precision needed for distinguishing between types of materials that are visually similar but recyclably distinct [8]. Utilizing AI technology for the automatic identification and classification of waste images not only improves the efficiency and accuracy of waste sorting but also reduces labor costs [9]. Moreover, it helps increase the proportion of waste recycling, playing a significant role in environmental protection and resource recycling. For instance, Malik et al. [10] discussed an AI framework that incorporates intelligent recognition and management strategies to improve municipal solid waste image classification. Wang et al. [11] used MobileNetV3 and IoT technology to achieve high-precision identification of garbage, including plastic, paper, and more. Through deep learning models that identify various types of waste, rapid and precise classification of waste can be achieved, guiding the recycling and processing of waste and providing technical support for urban management and environmental protection.

However, in advancing image-based waste classification efforts, we encounter several significant challenges.

Firstly, intelligent recognition of municipal waste is not yet sufficient, with the challenge lying in the diversity and complexity of waste images [12]. Municipal waste encompasses a variety of types of refuse, with the shapes, sizes, colors, and configurations of these items potentially varying widely in images. Moreover, waste often appears against complex backgrounds, increasing the difficulty for intelligent recognition systems to identify and classify it. Improving recognition accuracy is crucial for optimizing resource recycling, reducing landfill volumes, and protecting the environment. Therefore, developing efficient intelligent recognition technologies to address these challenges is particularly important.

Secondly, there is a lack of effective multi-label recognition methods. In practice, an image often contains multiple types of waste, requiring the system to identify all waste types in an image simultaneously. However, traditional image recognition methods mostly focus on single-label recognition and fall short in dealing with complex scenarios that include multiple categories of waste, failing to meet the practical application demands.

Finally, faced with the task of processing a large volume of municipal waste classification, reducing computational complexity while maintaining high accuracy to achieve real-time processing poses another challenge. With the increasing amount of urban waste, the demand for processing speed also rises. Ensuring the system can rapidly and accurately process large volumes of data, given limited resources, is crucial for the efficient and automated realization of waste classification tasks.

Our work introduces several key contributions to the domain of municipal waste management through image recognition:

- Development of a flexible multi-label image classification framework: We present the Query2Label (Q2L) framework, tailored for the complex task of municipal waste image recognition. This model excels in identifying multiple types of waste within the same image, utilizing self-attention and cross-attention mechanisms to accurately classify waste types, enhancing both accuracy and efficiency.
- Utilization of a novel municipal waste dataset: Our study employs the "Garbage In, Garbage Out" (GIGO) dataset, a newly developed collection of urban waste images. This dataset, with its diversity and real-world scenarios, significantly aids in improving

the model's performance by providing a wide array of waste images for training and testing.

- High accuracy with low computational complexity: Compared to existing models, our approach achieves superior precision in identifying various types of waste while maintaining computational efficiency. This ensures the model's suitability for real-time applications, highlighting its potential for practical deployment in waste management systems.

The structure of our paper is outlined as follows: Section 2 reviews related works in multi-label image classification and intelligent waste identification. In Section 3, we introduce our novel Query2Label framework and the Vision Transformer as the backbone for our intelligent waste recognition model. Section 4 describes the GIGO dataset, our experimental setup, and evaluation metrics. Section 5 discusses the results from our experiments, demonstrating the effectiveness of our model through comparisons and ablation studies. The paper concludes in Section 6 with reflections on our findings and suggestions for future research directions.

2. Related Work

2.1. Multi-Label Image Classification

Multi-label image classification is a key research area in computer vision. Distinguished from traditional single-label classification, this task entails a higher level of complexity, necessitating models to not only identify all pertinent objects within an image but also comprehend the potential interrelations and hierarchical structures among these entities. The surge in deep learning advancements has notably propelled the evolution of multi-label classification methodologies. For instance, convolutional neural networks (CNNs) [13] have been widely leveraged for feature extraction and image representation, whereas the integration of attention mechanisms has further augmented models' acuity for image details and their proficiency in discerning label relevancies [14]. Additionally, graph neural networks (GNNs) have been applied to delineate the intricate relationships between labels, thereby ameliorating classification performance [15].

In practical applications, multi-label image classification finds utility across diverse scenarios, including biodiversity monitoring, social media content analysis, mineral recognition, and medical image diagnosis, to name a few. Within biodiversity monitoring, multi-label classification aids in the automated identification of various animal species depicted in images captured by field cameras, pivotal for ecological research and conservation efforts [16]. In the sphere of social media, multi-label classification of user-uploaded images enables a nuanced comprehension and analysis of user interests and behavioral patterns [17]. In mineral exploration, Qi et al. have effectively harnessed multi-label classification for the swift identification of assorted minerals within mineral images, facilitating preliminary mineral exploration outside laboratory confines [18]. Regarding medical imaging, the adoption of multi-label classification has rendered feasible the diagnosis of multiple pathologies from a single medical image, significantly elevating the efficacy and precision of medical diagnostics [19].

2.2. Intelligent Waste Identification

The relevant literature has provided a comprehensive review of efforts in intelligent waste identification [20,21]. CNNs are the predominant machine learning models utilized for waste identification, accounting for 87% of the models reviewed. Additional models such as support vector machines, hidden Markov models, and classification trees were also employed in a minority of studies. These studies predominantly engaged with datasets based on images (both visible and infrared) and sound, addressing tasks such as single-label classification, bounding box detection, and pixel segmentation. Prominent public datasets include Trashnet and Taco, with classification models based on the Trashnet dataset achieving an average accuracy of 92.9%, and a study employing a ResNext architecture alongside image augmentation techniques reaching a top accuracy of 98% [22–24]. Moreover, some studies have explored data fusion techniques combining visual and acoustic features to enhance classification performance, exemplified by the use of pretrained Visual Geometry Group Network (VGGNet) models and one-dimensional CNNs for waste classification [25,26].

Investigations have identified 13 architecture types using 14 feature extractors or backbones. A common approach involved proposing custom architectures, particularly prevalent among classification models, where 33% of CNNs models opted for custom solutions. ResNet served as the most frequently employed feature extractor, especially within detection models in conjunction with various architectures. Additionally, Darknet and its variations, serving as the default backbone for Yolo architectures, were widely adopted. The feature extractors exhibited more diversity in classification models, with VGGNet and MobileNet being among the most popular [27–29]. These studies highlight the diversity and complexity of applying machine learning techniques for intelligent waste identification, also indicating the range of technologies and approaches considered in developing efficient waste classification systems.

3. Method

3.1. Q2L Framework for Intelligent Multi-Label Waste Image Recognition

In this study, we propose a novel framework, Q2L [30], aimed at addressing the challenges associated with intelligent multi-label recognition of urban waste images. The framework is designed to overcome the limitations of traditional single-label classification methods, especially in complex scenarios involving images with multiple types of waste. Utilizing self-attention and cross-attention mechanisms, Q2L effectively models the intricate relationships among waste types as well as the interactions between waste images and labels, as shown in Figure 1.

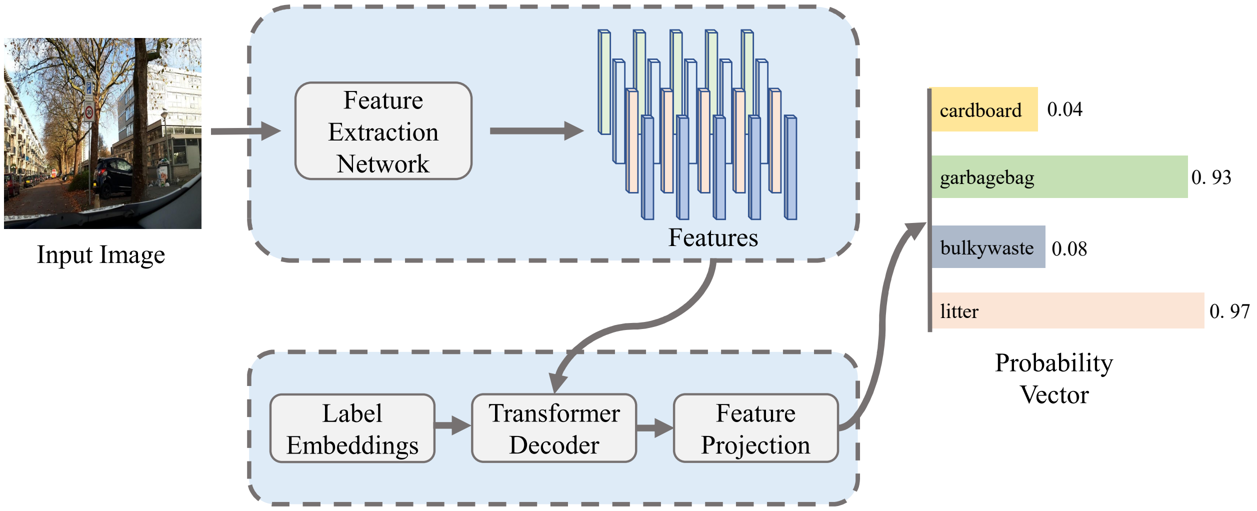


Figure 1. Framework of Query2Label.

Initially, the Q2L framework accepts an input waste image $x \in \mathbb{R}^{H_0 \times W_0 \times 3}$, where $H_0 \times W_0$ represents the height and width of the image, and 3 stands for the RGB channels. A feature extraction network, which can be either a convolutional neural network or a transformer-based network, processes the image to produce a feature map $F \in \mathbb{R}^{H \times W \times d}$, with $H \times W$ and d denoting the height, width, and dimension of the feature map, respectively.

Following feature extraction, the core of the Q2L framework is the Transformer decoder, which models the extracted features and label embeddings. Through the self-attention mechanism, the model captures co-occurrence relationships among waste types, while the cross-attention mechanism aligns visual patterns with corresponding labels.

Specifically, the operations of self-attention (Self-Attn) and cross-attention (Cross-Attn) can be formulated as follows:

$$\text{Self-Attn}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$$\text{Cross-Attn}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K, V represent the query, key, and value matrices, respectively, and d_k is the dimension of the key matrix. These mechanisms allow Q2L to comprehensively process the dependencies and interactions between various waste types, thus enhancing recognition accuracy and efficiency.

3.2. Backbone

To enhance the intelligent multi-label classification of urban waste images, we incorporate the ViT [4] as the backbone within our Q2L framework, distinctively suited for parsing the complexities of waste categorization. ViT's architecture, distinct from conventional convolutional approaches, offers a nuanced understanding of spatial hierarchies and inter-patch relationships critical for identifying various waste components in an image, as shown in Figure 2.

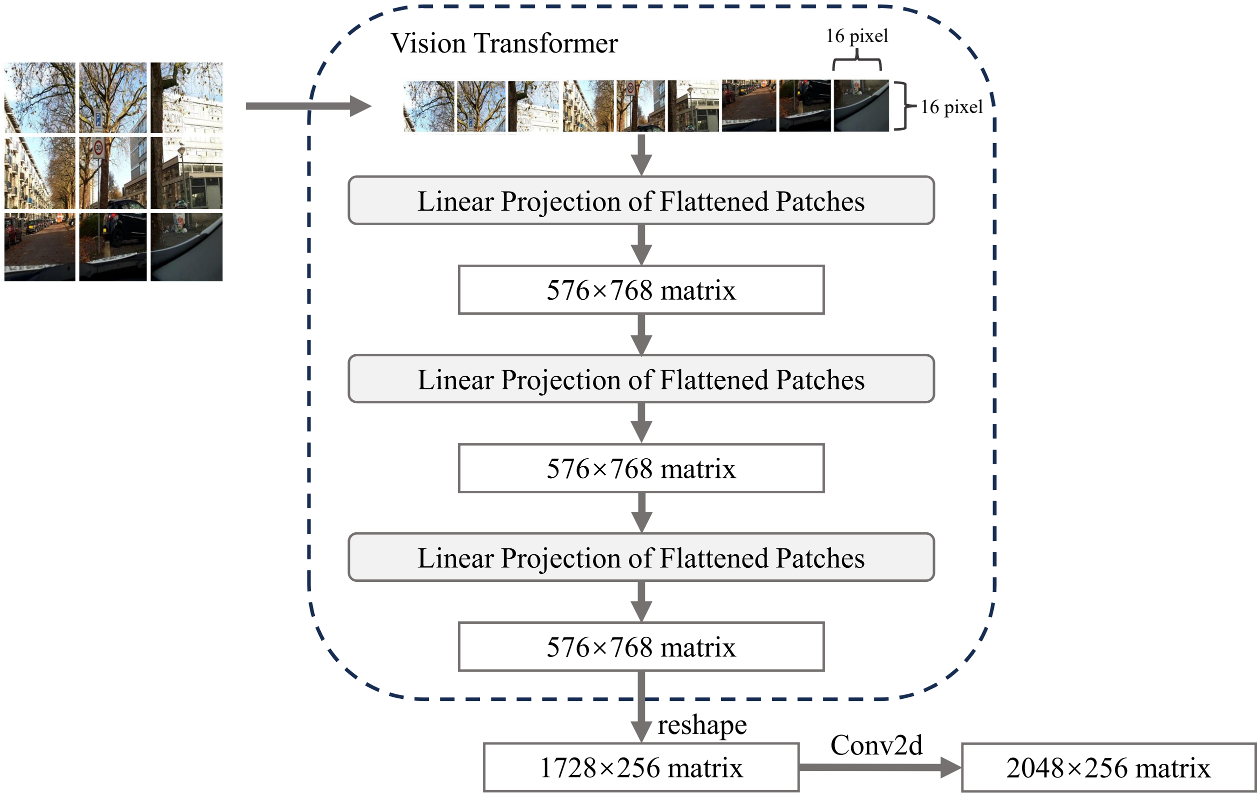


Figure 2. Framework of ViT.

The preprocessing stage involves normalizing and resizing input waste images $x \in \mathbb{R}^{H \times W \times C}$ to a uniform dimension of 384 × 384. The image is partitioned into 16 × 16 pixel patches, akin to words in a sentence for natural language processing (NLP) tasks. These patches are linearly projected into a D-dimensional embedding space, creating a sequence of patch embeddings. To retain positional context, necessary for discerning spatial arrangements of waste types, positional embeddings are integrated with the patch embeddings.

The core analytical process employs a transformer encoder that operates on the patch embeddings, augmented with positional information. This encoder, through self-attention mechanisms, enables the model to focus on relevant segments of the waste image for classification. It effectively captures global dependencies across patches, facilitating a comprehensive understanding of the image context. This is vital for accurately identifying and classifying multiple waste items present within a single image.

3.3. Asymmetric Loss Function

In addressing the complex challenge of multi-label waste image classification, it becomes essential to adopt a loss function that can effectively manage the intricacies of this task, including class imbalance and the presence of hard-to-classify instances. Traditional loss functions like binary cross-entropy (BCE) [31] offer a foundational approach by evaluating the prediction accuracy across multiple labels. However, this method may not adequately emphasize the more challenging or less frequent waste categories, leading to suboptimal classification performance.

To navigate these challenges, we introduce an adapted asymmetric loss (ASL) function [32], which is tailored to the unique requirements of waste image classification. The ASL function is designed to mitigate the limitations of conventional loss functions by applying distinct focusing parameters for positive and negative predictions, thereby enhancing the model's sensitivity to rare and difficult-to-detect waste categories.

The ASL function for a given waste image classification task is formulated as follows:

$$L_{asl} = -\sum_{i=1}^{M}[y_i(1-p_i)^{\gamma_+}\log(p_i) + (1-y_i)(p_i-m)^{\gamma_-}\log(1-p_i-m)]$$

where M represents the total number of waste categories, y_i denotes the ground truth label for category i, p_i is the predicted probability for category i, γ_+ and γ_- are the focusing parameters for positive and negative samples, respectively, and m is a margin applied to adjust the model's response to highly confident negative predictions, effectively reducing their influence on the loss calculation.

The introduction of a margin m in the ASL function serves to further refine the focus on challenging negative samples, ensuring that the model does not become complacent with easily classified negatives. By dynamically adjusting the influence of positive and negative samples through γ_+ and γ_-, the ASL function allows for a more nuanced training process. This process encourages the model to prioritize learning from misclassified or rare waste types, which are often overlooked by more conventional approaches.

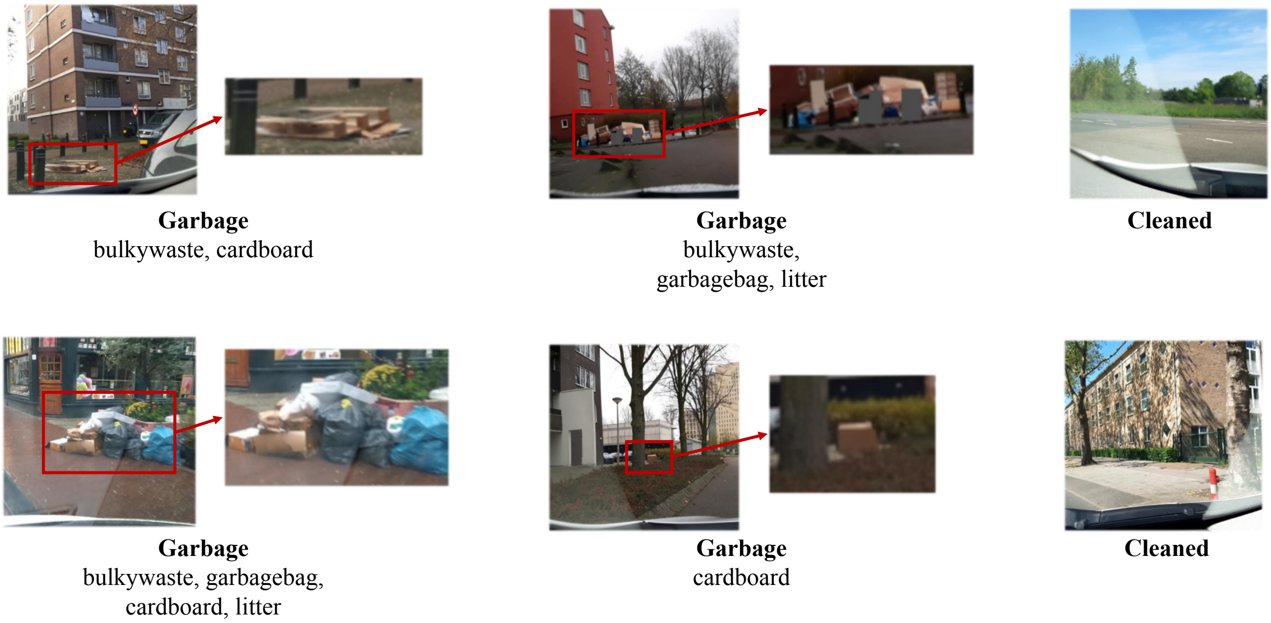
4. Dataset and Experimental Settings

4.1. Dataset—GIGO

In our research, the GIGO dataset [33], developed by Sukel et al. in 2023, underpins our analysis on urban waste classification through machine learning. This dataset, comprising 25,000 street-level images, was compiled to help identify and categorize urban waste. It features images captured by dual cameras mounted on vehicles traversing city roads, presenting a diverse array of urban waste scenes. A key consideration in curating the dataset was the preservation of privacy, with efforts made to obscure identifiable information such as faces and car license plates through YOLO detection and subsequent manual refinement by experts applying masking boxes.

The dataset divides images into categories of either containing waste ("garbage") or not ("cleaned"), further segmenting identified waste into four distinct types: waste bags ("Garbage Bag"), cardboard debris ("Cardboard"), oversized refuse items ("Bulky Waste"), and smaller litter items ("Litter"). Each image marked with garbage presence is accompanied by multi-label annotations, facilitating the recognition of multiple waste types within a single frame. Among the total images, 9351 are categorized as containing garbage, with the remaining 15,647 depicting cleaner urban scenarios.

In our work, the dataset GIGO utilized is broadly described in the figure and tables below. Figure 3 presents a selection of samples from the dataset. Tables 1 and 2 display the distribution of classes within the dataset.



Garbage
bulkywaste, cardboard

Garbage
bulkywaste,
garbagebag, litter

Cleaned

Garbage
bulkywaste, garbagebag,
cardboard, litter

Garbage
cardboard

Cleaned

Figure 3. Examples of the GIGO dataset, with zoomed window of objects to be identified.

Table 1. Statistics of classes.

Class Name	Number of Images
Not Garbage	15,647
Garbage	9351
Garbage Bag	1957
Cardboard	4391
Bulky Waste	5055
Litter	4863

Table 2. Statistics of garbage images.

Number of Garbage	Number of Images
0	15,647
1	4676
2	2802
3	1496
4	378

4.2. Experimental Settings

Our experimental settings are guided by the following parameters, as shown in Table 3.

Table 3. Experimental settings for urban waste image classification.

Parameter	Value
Batch size	24
Optimizer	RMSprop with Momentum [34]
Initial learning rate	1×10^{-2}
Decay	0.95
Decay steps	10,000
Momentum	0.9
Final learning rate	1×10^{-5}

4.3. Evaluation Metrics

To assess the models' performance in multi-label classification, we employ a comprehensive set of metrics: precision, recall, F1 score, mean average precision (mAP), and floating point operations per second (FLOPs). These metrics offer insights into both the effectiveness and efficiency of our models.

Precision and recall gauge the accuracy of positive predictions and the model's ability to identify all relevant instances, respectively, defined as follows:

$$Precision(P) = \frac{TP}{TP + FP}$$

$$Recall(R) = \frac{TP}{TP + FN}$$

where TP denotes true positives, FP false positives, and FN false negatives.

The $F1$ score balances precision and recall, calculated as the harmonic mean of the two:

$$F1 = 2 \cdot \frac{P \cdot R}{P + R}$$

Mean average precision (mAP) assesses the model across all labels by averaging the precision–recall curve's area under the curve for each label. The AP for a single label is computed as follows:

$$AP = \sum_{k=0}^{n-1} [R(k) - R(k+1)] \cdot P(k)$$

Finally, mAP averages the AP values across all labels, while FLOPs measure the model's computational complexity by counting the required floating-point operations for a single instance prediction.

These metrics collectively inform our understanding of the models' capabilities and limitations, providing a comparative basis for our work, as they are widely used to measure performance in different classification tasks [14,35,36] and guide us in refining approaches to multi-label classification in urban waste management.

5. Experimental Evaluations

5.1. Comparison of Different Backbone Networks

This study conducts a thorough analysis of backbone network selection for the task of intelligent waste multi-label classification, as shown in Table 4. Given the pivotal role of backbone networks in image feature extraction, a comparison of several popular pretrained models, including ResNet-101 [37], MobileNetV3 [38], and ViT-B/16, was undertaken, focusing on key metrics such as the number of parameters, FLOPs, and mAP. Particular attention was paid to the performance of the ViT-B/16 as a backbone network in the intelligent waste multi-label classification task. The ViT-B/16, with its unique Transformer architecture, offers a novel approach to processing image data. By segmenting images into sequenced blocks and applying a self-attention mechanism, it captures global dependencies within the image, crucial for complex multi-label classification tasks.

Table 4. Comparison of different backbone networks.

Backbone Network	Number of Parameters	FLOPs	mAP (%)
ResNet-101	44.5 M	45.8 G	87.61
MobileNetV3	5.4 M	12.3 G	82.54
ViT-B/16	86.0 M	35.8 G	92.36

Based on the experimental outcomes, ViT-B/16 was selected as the backbone network for the intelligent waste multi-label classification task in this study. This choice is predicated on the superior performance demonstrated by ViT-B/16, along with its potential for handling complex pattern recognition in images. The adoption of ViT-B/16 not only offers an efficient solution for the task of intelligent waste classification but also highlights the extensive applicability and powerful potential of Transformer architectures in visual tasks.

5.2. Comparison of Different Loss Functions

In this study, we explored the intelligent waste multi-label classification model, emphasizing the impact of different loss functions on model performance, as shown in Table 5. The selection of suitable loss functions for experimentation and comparison was crucial to achieve desirable training outcomes. A comprehensive evaluation of various loss functions revealed their significant influence on model performance, especially in multi-label classification tasks. We experimented with various combinations of parameters for asymmetric loss to identify the optimal parameter combination that enhances the model's mAP. The model achieved optimal performance, with an mAP of 92.36%, when the parameters were set to $\gamma_- = 4$ and $\gamma_+ = 0$. This indicates the model's high accuracy and generalization capability in addressing complex multi-label classification challenges.

Table 5. Comparison of different loss functions.

Loss Function	Parameter Setting	mAP (%)
Binary Cross-Entropy	—	89.97
Focal Loss [39]	$\alpha = 0.25, \gamma = 2$	91.08
Asymmetric Loss	$\gamma_- = 4, \gamma_+ = 0$	92.36

5.3. Confusion Matrix

Figure 4 showcases the performance of different models in the garbage image recognition task, focusing on identifying Bulky Waste, Garbage Bag, Cardboard, and Litter. Our model, integrating the Q2L architecture with ViT/B-16 and ASL, outperforms YOLOv7 [40] and ResNet-101 [41] across all categories, with precision scores of 92% for Bulky Waste, 89% for Garbage Bag, 94% for Cardboard, and 88% for Litter. The literature acknowledges the reliability of YOLOv7 and ResNet-101 for various classification tasks, but our model demonstrates significant advancements. Our model surpasses others by more effectively handling the scale diversity of garbage items and accounting for the interrelationships between different waste categories. By leveraging the Q2L framework, which efficiently maps queries to their corresponding labels, combined with the ASL's innovative approach to handling label imbalance, our model achieves an exceptional balance of accuracy and adaptability.

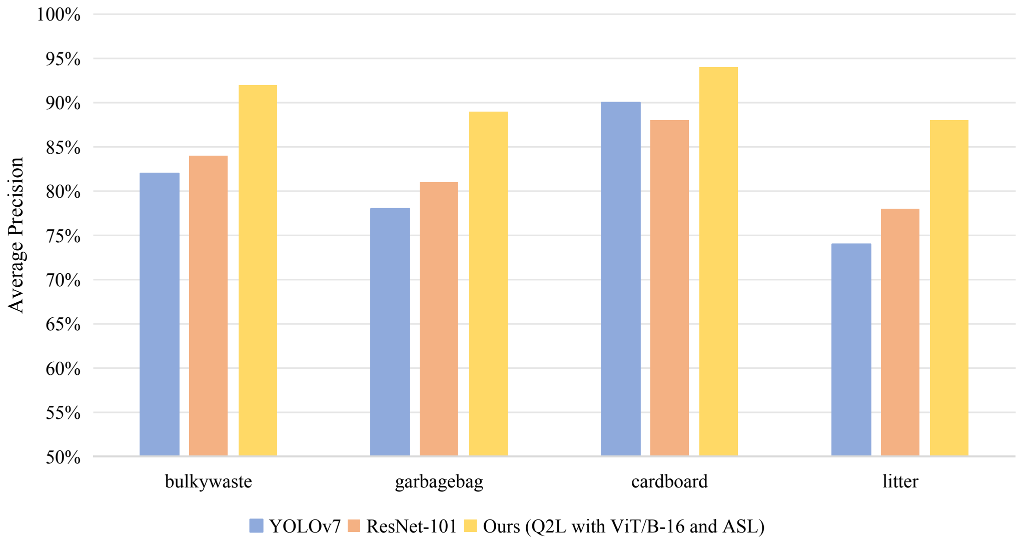


Figure 4. Comparison of different models.

Below, we present the confusion matrix scenarios among different categories, as shown in Figure 5. Simultaneously, the combined confusion matrix, regarding the accurate identification of the presence of garbage, is also displayed, as illustrated in Figure 6. A confusion matrix is a tool used to visualize the performance of a classification model by laying out the predicted labels against the true labels. Each cell in the matrix contains a percentage that represents the proportion of predictions for that particular label combination. The diagonal values of the matrix signify the model's accuracy in correctly predicting each category. In the matrix for different categories, a value of 0.92 in the diagonal position for Bulky Waste indicates that 92% of the actual Bulky Waste instances were correctly identified by our model. Conversely, the off-diagonal values indicate the rates of misclassification. For example, a value of 0.02 at the intersection of the Bulky Waste row and the Garbage Bag column reflects that 2% of the items truly labeled as Bulky Waste were mistakenly predicted as Garbage Bag. The results from our confusion matrices demonstrate high model accuracy and reliability in waste classification, with high true positive and true negative rates that show the model's effectiveness in practical municipal waste management applications.

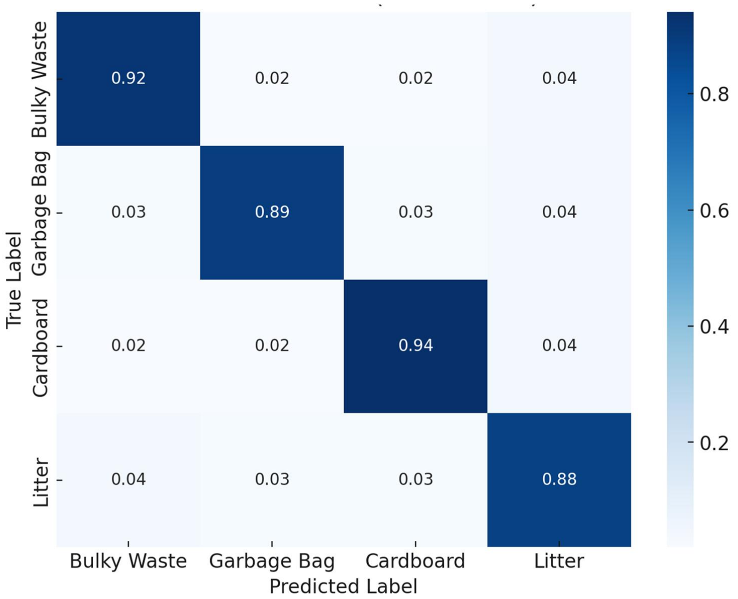


Figure 5. Confusion matrix among different categories.

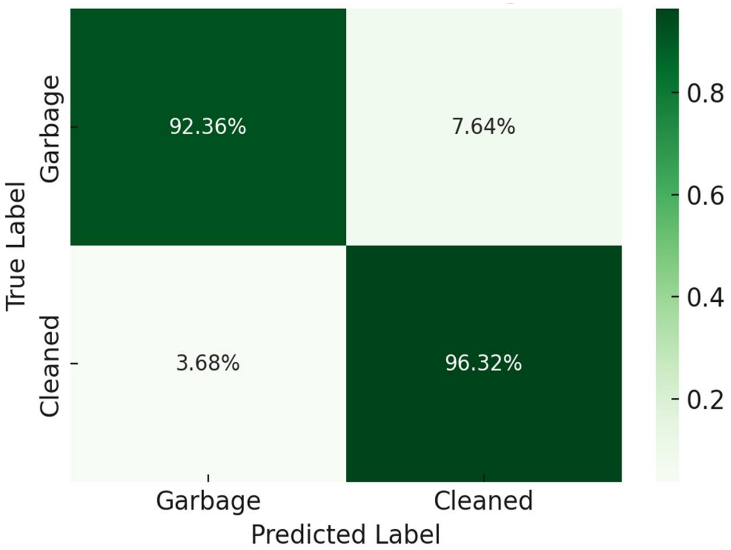


Figure 6. Combined confusion matrix.

5.4. Ablation Experiment

In this research, a series of ablation experiments were conducted to verify the effectiveness of proposed model optimization strategies. These experiments aimed to analyze the impact of different configurations on the performance of the intelligent waste multi-label classification model. The configurations included the baseline model, backbone modification, loss function modification, and combined modification.

1. Baseline model: Utilized the initial backbone network and loss function settings, serving as the performance comparison benchmark.
2. Backbone modification: Altered only the backbone network to ViT-B/16, assessing its impact on model performance.
3. Loss function modification: Maintained the backbone network while changing the loss function to asymmetric loss, exploring the performance improvement due to this modification.
4. Combined modification: Simultaneously changed the backbone network to ViT-B/16 and the loss function to asymmetric loss, examining the model's performance under the combined effect of these optimizations.

The ablation experiments' results, as shown in Table 6, demonstrate that each optimization strategy effectively enhances the model's performance in the intelligent waste multi-label classification task. Specifically, the model achieves its highest performance, with an mAP of approximately 95%, when both the backbone network and the loss function are modified. This underscores the effectiveness of the proposed optimization strategies and their potential application in intelligent waste classification.

Table 6. Summary of ablation experiment results.

Configuration	**mAP (%)**
Baseline model	88.62
Backbone modification (ViT-B/16)	89.97
Loss function modification (asymmetric loss)	90.20
Combined modification	92.36

6. Conclusions and Future Work

This study explores how AI is effectively being used in image recognition for managing municipal waste. With cities growing rapidly, efficient waste sorting is more important than ever. The study suggests that AI can improve sorting accuracy and efficiency, which will help boost recycling efforts. The paper introduces a novel framework named Query2Label, combined with ViT/B-16 as the backbone and an asymmetric loss function, to tackle the inherent complexities of multi-label waste image classification. Through meticulous experimentation on the "Garbage In, Garbage Out" dataset, it demonstrates the framework's superiority in recognizing diverse waste types against varying backdrops, achieving remarkable precision and recall metrics over conventional methods like YOLOv7 and ResNet-101.

Despite its advancements, the study identifies room for improvement in areas such as handling the vast diversity within municipal waste categories and further reducing computational demands to enable real-time processing. The current model, while efficient, might struggle with highly cluttered scenes or rare waste items not adequately represented in the training dataset.

For future work, we envisage several key areas of development to further enhance the capabilities of our image recognition framework for municipal waste management:

- Dataset expansion and diversification: To enhance the model's generalization capabilities across a broader spectrum of waste types and scenarios, it is imperative to expand and diversify the training dataset. This expansion could include a variety of waste materials and configurations, as well as a more extensive range of environmental conditions. Additionally, incorporating data from multiple cities can mitigate the influence of specific urban aesthetics and municipal characteristics, which will further enhance the model's adaptability and performance across diverse urban settings.
- Integration of multiple sensory inputs: Incorporating data from additional modalities, such as infrared imaging, depth sensing, and perhaps even acoustic sensors, could significantly enhance the model's ability to distinguish between different types of waste in visually complex scenes. This multi-modal approach might reveal characteristics of materials that are not apparent in visual-spectrum photographs alone.
- Development of lightweight models: Investigating and developing more efficient model architectures that maintain high accuracy while being computationally less demanding is essential. This could facilitate the deployment of advanced waste classification systems on mobile or embedded devices, enabling real-time processing and decision making at the point of waste collection or sorting.

In summary, the future enhancements of our municipal waste management image recognition framework focus on expanding the training dataset, integrating multiple sensory inputs, and developing lightweight models. These initiatives aim to improve waste type generalization, enhance recognition in complex scenes, and enable real-time decision making, setting a foundation for more accurate and efficient waste classification systems.

Author Contributions: Conceptualization, R.W. and T.Z.; methodology, J.X. and X.L.; software, J.X.; validation, R.W., T.Z. and J.L.; formal analysis, T.Z.; investigation, J.L.; resources, X.L.; data curation, J.L.; writing—original draft preparation, R.W.; writing—review and editing, T.Z., G.G. and J.X.; visualization, T.Z.; supervision, M.Z. and G.G.; project administration, M.Z. and G.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data will be made available on request.

Acknowledgments: The authors would like to thank the anonymous reviewers for their helpful and constructive comments which greatly contributed to the improvement of the final version of the paper.

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

References

1. Al-Antari, M.A. Artificial intelligence for medical diagnostics—Existing and future aI technology! *Diagnostics* **2023**, *13*, 688. [CrossRef] [PubMed]
2. Grigorescu, S.; Trasnea, B.; Cocias, T.; Macesanu, G. A survey of deep learning techniques for autonomous driving. *J. Field Robot.* **2020**, *37*, 362–386. [CrossRef]
3. Kolhar, M.; Alameen, A. Artificial Intelligence Based Language Translation Platform. *Intell. Autom. Soft Comput.* **2021**, *28*. [CrossRef]
4. 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 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
5. Ruiz, V.; Sánchez, Á.; Vélez, J.F.; Raducanu, B. Automatic image-based waste classification. In Proceedings of the From Bioinspired Systems and Biomedical Applications to Machine Learning: 8th International Work-Conference on the Interplay Between Natural and Artificial Computation, IWINAC 2019, Almería, Spain, 3–7 June 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 422–431.
6. Dada, M.A.; Obaigbena, A.; Majemite, M.T.; Oliha, J.S.; Biu, P.W. Innovative approaches to waste resource management: implications for environmental sustainability and policy. *Eng. Sci. Technol. J.* **2024**, *5*, 115–127. [CrossRef]
7. Smith, Y.R.; Nagel, J.R.; Rajamani, R.K. Eddy current separation for recovery of non-ferrous metallic particles: A comprehensive review. *Miner. Eng.* **2019**, *133*, 149–159. [CrossRef]
8. Zurbrugg, C. Urban solid waste management in low-income countries of Asia how to cope with the garbage crisis. *Present. Sci. Comm. Probl. Environ. (SCOPE) Urban Solid Waste Manag. Rev. Sess. Durban S. Afr.* **2002**, *6*, 1–13.
9. Choi, J.; Lim, B.; Yoo, Y. Advancing Plastic Waste Classification and Recycling Efficiency: Integrating Image Sensors and Deep Learning Algorithms. *Appl. Sci.* **2023**, *13*, 10224. [CrossRef]
10. Malik, M.; Sharma, S.; Uddin, M.; Chen, C.L.; Wu, C.M.; Soni, P.; Chaudhary, S. Waste classification for sustainable development using image recognition with deep learning neural network models. *Sustainability* **2022**, *14*, 7222. [CrossRef]
11. Wang, C.; Qin, J.; Qu, C.; Ran, X.; Liu, C.; Chen, B. A smart municipal waste management system based on deep-learning and Internet of Things. *Waste Manag.* **2021**, *135*, 20–29. [CrossRef]
12. Das, S.; Lee, S.H.; Kumar, P.; Kim, K.H.; Lee, S.S.; Bhattacharya, S.S. Solid waste management: Scope and the challenge of sustainability. *J. Clean. Prod.* **2019**, *228*, 658–678. [CrossRef]
13. Li, Z.; Liu, F.; Yang, W.; Peng, S.; Zhou, J. A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 6999–7019. [CrossRef] [PubMed]
14. 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 , Las Vegas, NV, USA, 27–30 June 2016; pp. 2285–2294.
15. Chen, Z.M.; Wei, X.S.; Wang, P.; Guo, Y. Multi-label image recognition with graph convolutional networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5177–5186.
16. Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; Belongie, S. The inaturalist species classification and detection dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8769–8778.
17. 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]
18. Gao, Q.; Long, T.; Zhou, Z. Mineral identification based on natural feature-oriented image processing and multi-label image classification. *Expert Syst. Appl.* **2024**, *238*, 122111. [CrossRef]
19. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R.M. Chestx-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2097–2106.
20. Arbeláez-Estrada, J.C.; Vallejo, P.; Aguilar, J.; Tabares-Betancur, M.S.; Ríos-Zapata, D.; Ruiz-Arenas, S.; Rendón-Vélez, E. A Systematic Literature Review of Waste Identification in Automatic Separation Systems. *Recycling* **2023**, *8*, 86. [CrossRef]
21. Sinthiya, N.J.; Chowdhury, T.A.; Haque, A.B. Artificial intelligence based Smart Waste Management—A systematic review. In *Computational Intelligence Techniques for Green Smart Cities*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 67–92.
22. 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.
23. Proença, P.F.; Simoes, P. Taco: Trash annotations in context for litter detection. *arXiv* **2020**, arXiv:2003.06975.
24. Singh, S.; Gautam, J.; Rawat, S.; Gupta, V.; Kumar, G.; Verma, L.P. Evaluation of transfer learning based deep learning architectures for waste classification. In Proceedings of the 2021 4th International Symposium on Advanced Electrical and Communication Technologies (ISAECT), Alkhobar, Saudi Arabia, 6–8 December 2021; pp. 1–7.
25. Funch, O.I.; Marhaug, R.; Kohtala, S.; Steinert, M. Detecting glass and metal in consumer trash bags during waste collection using convolutional neural networks. *Waste Manag.* **2021**, *119*, 30–38. [CrossRef]
26. Lu, G.; Wang, Y.; Xu, H.; Yang, H.; Zou, J. Deep multimodal learning for municipal solid waste sorting. *Sci. China Technol. Sci.* **2022**, *65*, 324–335. [CrossRef]

27. Chen, Y.; Sun, J.; Bi, S.; Meng, C.; Guo, F. Multi-objective solid waste classification and identification model based on transfer learning method. *J. Mater. Cycles Waste Manag.* **2021**, *23*, 2179–2191. [CrossRef]
28. Feng, B.; Ren, K.; Tao, Q.; Gao, X. A robust waste detection method based on cascade adversarial spatial dropout detection network. In Proceedings of the Optoelectronic Imaging and Multimedia Technology VII, Online, 11–16 October 2020; SPIE: Bellingham, WC, USA, 2020; Volume 11550, pp. 179–188.
29. Cai, H.; Cao, X.; Huang, L.; Zou, L.; Yang, S. Research on Computer Vision-Based Waste Sorting System. In Proceedings of the 2020 5th International Conference on Control, Robotics and Cybernetics (CRC), Wuhan, China, 16–18 October 2020; pp. 117–122.
30. Liu, S.; Zhang, L.; Yang, X.; Su, H.; Zhu, J. Query2label: A simple transformer way to multi-label classification. *arXiv* **2021**, arXiv:2107.10834.
31. Ruby, U.; Yendapalli, V. Binary cross entropy with deep learning technique for image classification. *Int. J. Adv. Trends Comput. Sci. Eng.* **2020**, *9*, 5393–5397.
32. Ridnik, T.; Ben-Baruch, E.; Zamir, N.; Noy, A.; Friedman, I.; Protter, M.; Zelnik-Manor, L. Asymmetric loss for multi-label classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 82–91.
33. Sukel, M.; Rudinac, S.; Worring, M. GIGO, Garbage In, Garbage Out: An Urban Garbage Classification Dataset. In Proceedings of the International Conference on Multimedia Modeling, Bergen, Norway, 9–12 January 2023; Springer: Berlin/Heidelberg, Germany, 2023; pp. 527–538.
34. Zou, F.; Shen, L.; Jie, Z.; Zhang, W.; Liu, W. A sufficient condition for convergences of adam and rmsprop. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 11127–11135.
35. Li, Y.; Chen, Y.; Dai, X.; Chen, D.; Liu, M.; Yuan, L.; Liu, Z.; Zhang, L.; Vasconcelos, N. Micronet: Improving image recognition with extremely low flops. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 468–477.
36. Tripathi, M. Analysis of convolutional neural network based image classification techniques. *J. Innov. Image Process. (JIIP)* **2021**, *3*, 100–117. [CrossRef]
37. Wu, Z.; Shen, C.; Van Den Hengel, A. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognit.* **2019**, *90*, 119–133. [CrossRef]
38. Koonce, B.; Koonce, B. MobileNetV3. In *Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 125–144.
39. Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; Yang, J. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21002–21012.
40. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
41. 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.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.