Swift and Smart: A New Paradigm for Real-Time Garbage Segmentation

Jun Wang
University of Washington
Seattle, WA 98195
junw3@cs.washington.edu

Robin Wang
University of Washington
Seattle, WA 98195
qwang6@cs.washington.edu

Abstract

Waste management is a critical and consequential task, as improper waste disposal can induce profound environmental and economic damage, leading to soil, water, and air pollution, increased waste collection and disposal expenses, and health risks for waste-handling personnel. While existing garbage classification algorithms are generally effective, they fail to accurately sort waste that is intermixed or contaminated with other materials. To address this challenge, we introduce a real-time semantic segmentation solution featuring a two-stage, end-to-end pipeline. Our approach combines the capabilities of the state-of-the-art Segment Anything Model (SAM) and Grounding DINO object detector to automate the labeling process for raw waste images. Subsequently, we fine-tune the models to optimize accuracy in classifying garbage objects, facilitating the model to perform sophisticated semantic image segmentation. Our best model combining the MobileNetV3 and the DeepLabv3 architectures has an impressive Mean Intersection over Union (MIoU) score of 0.7865 and over 210 Frames Per Second (FPS). The accuracy and efficiency of our model suggest significant applicability for deployment in complex real-world scenarios.

1. Introduction

Sorting waste into appropriate bins is a minor yet crucial task. Inappropriate trash disposal can lead to severe consequences, including soil, water, and air pollution, increased waste collection and disposal expenses, and health risks for those handling waste. With the advancement of deep learning over the past decade, Artificial Intelligence experts have been refining neural networks to efficiently categorize waste, with the aim of replacing manual sorting. For instance, researchers have evaluated the performance of notable deep learning architectures including DenseNet121 [8], DenseNet169 [8], InceptionResnetV2 [17], MobileNets [7], and Xception [3] using the Trashnet dataset (comprising 2,527 images with six waste categories) [1]. Meanwhile, others have concentrated on optimizing a single model (ResNet50 [5]) and employed methods such as data augmentation to enhance classification accuracy from 91.40% to 95.35% [15]. Despite their high accuracy in classifying different types of garbage, state-of-the-art garbage classification algorithms struggle to accurately identify and sort waste that is mixed or contaminated with other materials. In such cases, garbage segmentation becomes necessary.

We propose an innovative two-stage, end-to-end pipeline aimed at enabling real-time garbage segmentation. The first stage incorporates the data preprocessing phase, wherein we leverage the synergistic capabilities of the Grounding DINO [11] and the Segment Anything Model (SAM) [9]. This integrated approach facilitates automatic labeling of an extensive Kaggle dataset comprising more than 10,000 raw images, featuring 12 distinct categories of household waste (battery, biological, brown glass, cardboard, clothes, green glass, metal, paper, plastic, shoes, trash, and white glass) [16]. The subsequent stage focuses on model training, utilizing fine-tuned variants of the ResNet50 model and the MobileNetV3 model [6]for the classification of the aforementioned annotated images. Additionally, we employ a refined DeepLabv3 model [2] to perform semantic image segmentation tasks.

To evaluate the performance, our refined models accept a diverse range of waste types taken from the test set as raw image inputs and generate bounding boxes and segmentation masks for the objects. Our best model combining the MobileNetV3 and the DeepLabv3 architectures demonstrated a best Mean Intersection over Union (MIoU) score of 0.7865, a best average precision of 0.8147, a best average recall of 0.8112, and an impressive processing speed of 218.93 Frames Per Second (FPS). For comparison, our best model outperforms the baseline model (ResNet50 combined with a Fully Convolutional Network) regarding best average precision, best average recall, and best FPS. Our work, therefore, represents a significant step forward in developing sophisticated, real-time waste sorting solutions with the potential to greatly mitigate the negative impacts of inappropriate waste disposal.

2. Related Work

Due to the success of deep learning at extracting and learning meaningful features from raw data, researchers have developed and evaluated many models to perform garbage classifications. In this section, we summarize several recent works on garbage classification and segmentation.

In a recent study, experts tested popular deep learning models to find the most effective method for categorizing images of garbage from the Trashnet dataset [1]. By employing the Adam and Adadelta optimization methods and fine-tuning the dataset using data augmentation techniques, they discovered that DenseNet121 delivered the best outcome with a test accuracy rate of 95%. This success might be attributed to the unique properties of the DenseNet architecture, which has a track record of excelling in classification tasks. DenseNet is famous for its dense connectivity, in which every layer is linked to every other, thereby boosting information flow and gradient propagation. This architecture not only encourages feature reuse, but also reduces the number of parameters and enhances efficiency. Furthermore, it addresses the issue of the vanishing gradient, making the training of deeper networks more effective. All these elements contribute to DenseNet's capacity to identify intricate patterns and achieve high levels of classification accuracy.

While top-tier neural network models have shown remarkable progress in garbage classification tasks, the realm of garbage segmentation research is relatively unexplored. In a recent study, researchers applied the deep supervision UNet++ model to address this issue, particularly focusing on the classification and segmentation of road garbage [10]. This refined model builds upon its predecessor by incorporating a set of nested, dense skip pathways, improving the network performance by enabling better feature propagation. By training the model on a road garbage segmentation dataset consisting of four categories (stones, leaves, sand, and bottles), the model achieved considerable garbage segmentation improvement with an MIoU of 0.7673. However, despite the high accuracy, the model's real-time performance is limited.

Aside from the lack of focus on researching real-time garbage segmentation, the number of garbage types is another limiting factor of the model's real-world performance. To optimize the recycling process, it is essential to sort waste into groups that undergo similar recycling treatments. Most existing datasets categorize waste into a limited number of classes (typically 2 to 6 at most). The Trashnet dataset, for instance, organizes waste into six categories: paper, glass, trash, metal, plastic, and cardboard [1]. Enhancing the capacity to classify waste into a greater number of classes could substantially improve waste management.

3. Methods

Our approach to real-time garbage segmentation involves an end-to-end pipeline, incorporating automated image annotation, classification, and semantic segmentation. This cohesive pipeline pivots around two stages: data preprocessing and model training. Our choice of semantic segmentation over instance segmentation is governed by the practicality of real-life waste management scenarios, where it is not beneficial to further segregate garbage of identical types, as they would get disposed into the same trash bin. In real-life deployment scenarios where sensors and cameras performance are limited [20], it becomes imperative to optimally balance processing efficiency with the computational expense [18]. Hence, the crux of our methodology is the generation of a high-quality garbage dataset coupled with a focus on an equilibrium between our model's efficiency and its accuracy. This approach empowers our model to not only achieve superior performance metrics but also remain pragmatically feasible for real-time deployment in diverse and complex waste management environments.

3.1. Data Preprocessing

To address the model's bias towards a specific class of garbage, we implemented downsampling on classes that had a significant number of samples (e.g., clothes, shoes). This involved reducing the number of samples in these classes to approximately 700 images each, resulting in a training dataset of approximately 8,000 images. We recognized the dual importance of dataset quality and quantity in training robust garbage segmentation models. Manual annotation, while ensuring accuracy, poses challenges in terms of labor intensity and scalability, hence requiring an automated solution. To address this, we conducted a rigorous investigation, experimenting with different techniques to streamline the annotation process. We found a solution to automate the annotation process and expedite it using a combination of the Grounding DINO and the Segment Anything Model (SAM). This innovative data preprocessing approach not only reduced the workload but also maintained the high quality of our dataset. By automating the labor-intensive aspects of data preparation, we ensured our dataset's balance and integrity, which are critical factors in building effective garbage segmentation models. See Figure 2 for a sample from the annotated dataset.

In the data preprocessing stage, we first applied the Grounding DINO to perform Zero-Shot Object Detection on our dataset encompassing 12 distinct categories of garbage [11]. Grounding DINO excels at detecting objects even when confronted with objects that do not align with the predefined set of classes in the training data. This unique capability enables the model to adapt to novel objects and scenarios, rendering it highly versatile and relevant for subsequent garbage segmentation tasks.

Figure 1. Raw images are first processed by Grounding DINO. Garbages are detected and masked, which are then used as prompts for SAM.

Subsequently, the detected bounding boxes generated by Grounding DINO were utilized to prompt SAM for the purpose of conducting instance segmentation on the input images [9]. Given the extensive diversity and irregular spatial distribution of garbage in the real world, the adoption of instance segmentation through SAM presents a notable advantage by isolating each individual piece of garbage. This isolation ensures a more precise classification and segmentation in subsequent steps.

In order to achieve a seamless integration of SAM and Grounding DINO, and to potentially involve human tuning in the annotation results, the Roboflow API was utilized. This API facilitates the storage and inspection of annotated images, while also enables the incorporation of more fine-grained human adjustments on imperfect annotations. The integration of this API resulted in enhanced efficiency and effectiveness of the data preprocessing pipeline, leading to further improvements in our dataset quality.

Also, various data augmentation techniques were employed to improve the model's ability to learn from data effectively. Resizing, random cropping, and several affine transformations (e.g., horizontal flip, rotation) were applied to introduce variations in the training set. Additionally, resizing and normalizations are applied to ensure all images have the same dimensions and similar pixel value distributions before being fed into the model. For validation and training sets, similar resizing and normalization are performed but without any random transformations to preserve the original data integrity.

3.2. Model Training

In the model training stage, we compared the performances of different backbones (MobileNetV3, ResNet50) and heads (FCN, DeepLabv3) for our real-time garbage segmentation task. We utilized the labeled dataset generated during the preprocessing stage as an input for these models. The role of the backbones in this process was to extract relevant features, which formed the basis for distinguishing different types of waste. Our decision to adopt the ResNet50 architecture stems from its capacity to learn residual mappings and train deep neural networks and its proven effectiveness across various computer vision tasks such as image classification, object detection, and image segmentation made it a solid choice for our classification training task, which directly impacts the performance of subsequent segmentation tasks [5]. Alongside ResNet50, we trained MobileNetV3, an architecture known for its optimization in speed and size. Its efficiency makes it highly suitable for mobile and edge devices, lending versatility to our project.

Figure 2. Visualization of a final usable data sample for our Real-time Garbage Segmentation task. The plastic/biological garbage were segmented and annotated in a row with Grounding DINO and Segment Anything Model, without any human interference.

For the downstream semantic segmentation task, we employed DeepLabv3, a model recognized for its exceptional performance in semantic segmentation. DeepLabv3 incorporates dilated convolutions and atrous spatial pyramid pooling (ASPP) to capture multi-scale context [2]. It expands the receptive field of filters using dilated convolutions, enabling the model to process a larger context without escalating computational complexity. Each unit in the output of this layer is a weighted sum of all units in the corresponding location across all input channels, which can be

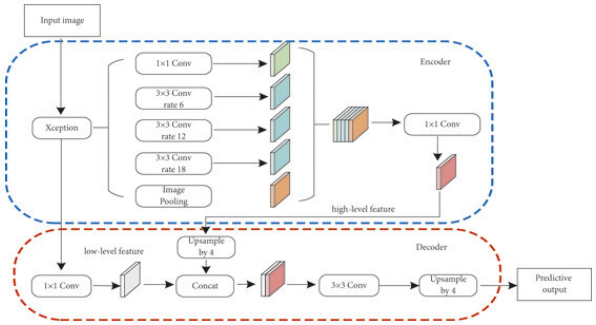

Figure 3. Although Deeplabv3 employs a series of convolutions and upsampling operations, it does not over-complicate the model, since the dilated convolutions allow the network to capture context at various scales, resulting in finer segmentation results, without significantly increasing computational complexity.

thought of as integrating the multi-scale context information. As a result, it provides a finer segmentation result by considering the context of the objects, allowing it to distinguish garbage from the surroundings accurately.

3.3. Baseline Comparison

We have implemented a straightforward baseline method for comparing performances across different backbone and head combinations. Initially, we employed Grounding DINO and SAM for data preprocessing, following the same procedure described earlier. Subsequently, we utilized a pretrained ResNet50 model as a backbone to extract features from the preprocessed dataset, capturing both visual and textual information. Lastly, we used a simple fully connected network as the head, which yields moderate accuracy and efficiency. This combined approach allows for the analysis and comparison to evaluate the similarity or dissimilarity of data samples within the given task, providing a fundamental baseline for our model's assessment. On top of that, we fine-tuned different variations of pretrained ResNet50 [5] and MobileNetV3 [6] to compare both accuracy and efficiency of various state-of-the-art models.

4. Experiments

In this section, we evaluate the accuracy and inference speed of our proposed real-time garbage segmentation solution on our dataset and provide detailed studies about our framework as well as qualitative results.

4.1. Experiment Setup

Our experimental setup utilized Google Colab Pro, which provided us with an Nvidia A100 GPU and 83.5 GB of RAM. The development was carried out on a MacBook Air (Retina, 13-inch, 2018) with a 1.6 GHz Dual-Core Intel Core i5 processor. We used the PyTorch framework for our model implementation.

Dataset The dataset of 8000 images was balanced across the 12 different garbage categories through downsampling (to reach 700 images for each category). Subsequently, it was divided into training, validation, and test sets, following an approximate ratio of 80:10:10, respectively.

Evaluation Metrics We define concrete metrics to ensure a thorough comparison between models, each with varying features such as different loss functions, optimization strategies, and regularization techniques. For classification tasks, we adopt classification accuracy, specifically precision and recall, to measure the performance of our models. For segmentation tasks, we utilize Mean Intersection over Union (MIoU), a metric that gauges the degree of overlap between the predicted and actual segments.

Implementation We fine-tuned all our models in the training phase for 7 epochs (6K iterations). The batch size was set to 8. The backbone is initialized with the ImageNet-pretrained weights with frozen batchnorm layers and other modules are randomly initialized. Data augmentation techniques were employed to improve the model's ability to learn from data effectively. Resizing, random cropping, and several affine transformations (e.g., horizontal flip, rotation) were applied to introduce variations in the training set. Additionally, resizing and normalizations are applied to ensure all images have the same dimensions and similar pixel value distributions before being fed into the model. For validation and training sets, similar resizing and normalization are performed but without any random transformations to preserve the original data integrity. We adopted the AdamW [13] optimizer with an initial learning rate $5e-5$ and weight decay 0.1 for ResNet50 backed models, and learning rate $1e-4$ and weight decay 0.01 for MobileNetV3 backed models. We also introduced a training scheduler with a step size of 200 iterations and a decay rate of 0.1. Additionally, we employed the pixel-wise cross-entropy loss [19] to compute the difference between the predicted and actual classes for each pixel independently, which aligns with the pixel-level accuracy requirement of multi-class semantic segmentation tasks.

4.2. Main Results and Implications

To find out an optimal solution that balances inference speed with the computational expense [18] for real-time garbage segmentation, we compared the performance of our proposed model with the established baseline method on 4 metrics: MIoU, precision, recall, and FPS.

Evaluation on Speed In Table 1, we analyze the impact of different model structures on inference speed. The results indicated that either adopting a light-weighted backbone or a simpler head can lead to significant improvement

in inference speed.

We present 2 key insights: 1) Model complexity is crucial for our task of real-time garbage segmentation; 2) In general, changing the backbone into a lighter one generally can lead to a larger bump than changing the head, as the majority of the inference time is usually taken by the backbone.

Our explanation is that the backbone is responsible for extracting hierarchical features from the input image. These operations, due to their depth and complexity, can consume a significant portion of the computational resources during inference. As such, opting for a more efficient backbone, such as MobileNetV3 [6] instead of ResNet50 [5], can often lead to a noticeable improvement in inference speed. The head of the model, on the other hand, typically performs some task-specific transformations on the features extracted by the backbone. In the case of semantic segmentation, this might include a series of convolutions and upsampling operations to generate a pixel-wise classification of the input image (as shown in Figure 3). While these operations can also be computationally expensive, they usually don't contribute as much to the overall computational cost as the backbone.

Evaluation on Accuracy Table 1 and Figure 4 shows the performance of the different models on the test data. While the combination of ResNet50 and DeepLabv3 achieves the highest accuracy, it falls short in MIoU metric.

We propose 2 explanations for such discrepancy: 1) High classification ability does not guarantee a comparably good segmentation ability. For instance, ResNet50 is a deeper and more complex model than MobileNetV3, which enables it to learn more complex features. This can potentially lead to higher classification accuracy but not necessarily better semantic segmentation performance, since semantic segmentation depends greatly on pixel-level information, not just object-specific features. It needs to identify boundaries and the extent of objects, which requires understanding the context of surrounding pixels as well; 2) Overfitting: models that learn more complex features often have a higher capacity (more parameters), which increases the risk of overfitting. In this case, the model backed by ResNet50 would be too attuned to the training data for semantic segmentation and did not generalize well to unseen data. It might have picked up on very specific patterns in the training data that do not hold in the validation or test set. By contrast, the combination of MobileNetV3 and DeepLabv3 achieved outstanding results in all 3 accracy metrics: MIoU, precision, and recall. MobileNetV3 is designed to be computationally efficient and is particularly suitable for mobile devices. Despite its reduced complexity, when combined with the powerful segmentation capabilities of DeepLabv3, it performs well on this segmentation task, achieving a desirable trade-off between performance and computational cost.

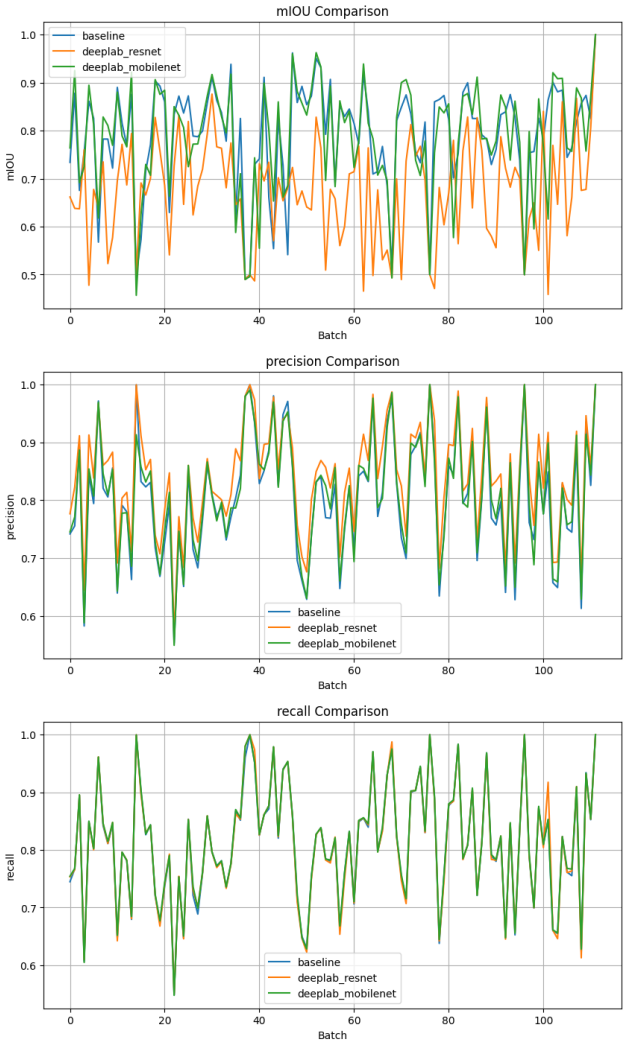

Figure 4. Performance comparison of various models using MIoU, precision, and recall metrics on the test dataset.

Final Implication Based on our experiment results and analysis, we can see the combination of MobileNetV3 and DeepLabv3 reveal the strongest deployment applicability, with the highest FPS, solid MIoU, precision, and recall performance.

5. Discussion and Future Work

5.1. Limitations

The current approach has several limitations. Firstly, despite the considerable time-saving benefits of the automatic data labeling pipeline, the quality of the segmentation and annotation results can be further improved through manual annotation. For example, users can use Roboflow's annotation editor to modify the object's labeling and segmentation boundary. Aside from improving the quality of the

Table 1. Best performance of different models on the test dataset

	ResNet50 + DeepLabv3	MobileNetV3 + DeepLabv3	ResNet50 + FCN (head)
Best MIoU	0.7911	0.7865	0.8011
Best Average Precision	0.8350	0.8147	0.8129
Best Average Recall	0.8490	0.8112	0.8039
Best FPS	118.48	218.93	170.85

dataset, the size of the dataset can be expanded for more effective model training. Our dataset currently comprises roughly 700 images for each class, but we hypothesize that the model could perform better by doubling the size of the dataset. One possible strategy for expanding the training dataset is by incorporating various unused data augmentation techniques in this research, such as adding Gaussian noise.

5.2. Future Work

Future works in the field can focus on further improving the current model's performance on semantic segmentation tasks by exploring different variations of vision transformers, such as MobileViT [14], Transformer iN Transformer (TNT) [4], and Swin Transformer [12]. We believe the self-attention mechanism would enable the model to understand the contextual relationships between pixels or regions and make more informed segmentation decisions. Moreover, by attending to all tokens (image patches) in the input instead of relying on local receptive fields like in CNNs, the model could better understand the spatial relationships between objects and background. Besides, classifying garbage into 12 classes is unlikely to be practical in real life due to space constraints and increased costs for acquiring and maintaining more trash cans. Therefore, we encourage researchers to determine a suitable number of classes based on user studies and real-world analysis.

5.3. Conclusion

In this research, we validate the viability of real-time garbage segmentation using models that judiciously balance computational efficiency and accuracy. We propose the combination of MobileNetV3 and DeepLabv3 as our solution for real-time garbage segmentation tasks. Our findings underscored the importance of maintaining a moderate model complexity in real-time garbage segmentation tasks: a lighter backbone like MobileNetV3 is crucial and can lead to a more significant increase in inference speed than simplifying the head.

Additionally, our contribution extends to the novel data processing paradigm we developed, significantly improving segmentation and annotation efficiency and effectiveness. We hope our dataset processing paradigm set a foundation for future research in other real-time semantic segmentation tasks.

References

[1] Rahmi A. Aral, Şeref R. Keskin, Mahmut Kaya, and Murat Hacıömeroğlu. Classification of trashnet dataset based on deep learning models. In *2018 IEEE International Conference on Big Data (Big Data)*, page 2058–2062, 2018. 1, 2

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. In *arXiv preprint arXiv:1706.05587*, 2017. 1, 3

[3] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *arXiv preprint arXiv:1610.02357*, 2016. 1

[4] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. In *arXiv preprint arXiv:2103.00112*, 2021. 6

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 3, 4, 5

[6] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. In *arXiv preprint arXiv:1905.02244*, 2019. 1, 4, 5

[7] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. In *arXiv preprint arXiv:1704.04861*, 2017. 1

[8] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *arXiv preprint arXiv:1608.06993*, 2016. 1

[9] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *arXiv preprint arXiv:2304.02643*, 2023. 1, 3

[10] Jiacai Liao, Libo Cao, Wei Li, Xiexing Feng, Jianhua Li, and Feng Yuan. Road garbage segmentation and cleanliness assessment based on semantic segmentation network for cleaning vehicles. In *IEEE Trans. Veh. Technol.*, 2021. 2

[11] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *arXiv preprint arXiv:2303.05499*, 2023. 1, 2

[12] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021. 6

[13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. 4

[14] Sachin Mehta and Mohammad Rastegari. Mobilevit: Lightweight, general-purpose, and mobile-friendly vision transformer. In *arXiv preprint arXiv:2110.02178*, 2022. 6

[15] Shanshan Meng and W. Chu. A study of garbage classification with convolutional neural networks. In *2020 Indo – Taiwan 2nd International Conference on Computing, Analytics and Networks (Indo-Taiwan ICAN)*, pages 152–157, 2020. 1

[16] Mostafa Mohamed. Garbage classification (12 classes), 2021. Accessed: 2023-04-24. 1

[17] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *arXiv preprint arXiv:1602.07261*, 2016. 1

[18] Yanping Tan, Wei Guo, Ke Yang, Jingjing Huang, and Zhengping Yang. Design of intelligent garbage classification system based on internet of things technology. *Journal of Physics: Conference Series*, 2187, 2022. 2, 4

[19] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7283–7293, 2021. 4

[20] Xianjun Yi, Yinyi Liang, and Hongchi Peng. Garbage classification system based on artificial intelligence and internet of things. *2022 International Conference on Artificial Intelligence and Computer Information Technology (AICIT)*, pages 1–5, 2022. 2