

ThanosNet: Attention-Based Trash Classification Using Meta-labels

Alan Sun

University of Maryland, College Park

asun17904@gmail.com

Harry Xiao

Columbia University

hx2310@columbia.edu

Abstract—In modern society, waste recycling and classification has become a necessity for reducing resource consumption and economic loss. For an effective waste classification system to be feasible, such as an intelligent waste sorting trash can, a robust model must be made available. In this study, we propose a novel attention-based trash classification model utilizing deep neural networks and meta-labels. We used time of day, location of the trash can, and distance to landmarks as the meta-labels. We collected ISBNet, a dataset which contains 889 images and their associated meta-labels, distributed over 5 classes (paper, plastic, cans, tetra pak, and landfill). Afterwards, we developed two different systems for the proposed attention-based trash classification, BilinearThanosNet and AdditiveThanosNet, both of which used ResNet50 as a feature extractor. By comparing ThanosNet with state-of-the-art, image-based classification models that use transfer learning on ISBNet, we found that ThanosNet displayed the best performance, with an F_1 score of 0.952. This justifies the use of meta-labels for trash classification and our ThanosNet model architecture.

Keywords—Trash classification, computer vision, deep learning, waste, recycling

I. INTRODUCTION

Society today has an ever-increasing awareness towards the importance of classifying trash properly for the purpose of recycling. Waste that is not properly sorted poses danger to soil, air, and water sources [1], while effective waste management reduces the pressure on landfills and can create beneficial economic and financial effects [2]. Reducing total waste is crucial in conserving and reusing resources, and becoming more sustainable [3], which is enabled through more extensive recycling. However, a key issue that stands in the way of widespread recycling lies in the accurate classification of recyclable and non-recyclable trash, particularly for consumers. At a consumer level, non-recyclable papers and plastics are often mistakenly placed into the recycling bins [4], contaminating and thus disqualifying the entire batch from being recycled. Utilizing machine learning to improve classification is a promising application and could vastly improve these inaccuracies present in consumer classification. The development of an automated trash can that could assist in the sorting of trash is a potentially impactful use case.

Utilizing deep learning to classify trash has been proposed numerous times. Salimi et al. [5], created a trash-bin robot that is capable of detecting trash and classifying it. Similarly, Auto-Trash, a trash can that can automatically sort waste into compost and recyclables made its debut at the 2016 TechCrunch Disrupt Hackathon in New York [6].

In the same year, Yang and Thung [7] released Trashnet, a

dataset that is now used as a benchmark for measuring waste classification performance. Currently, most of the state-of-the-art models on Trashnet use transfer learning to finetune well-known CNN-based models developed for the ImageNet challenge[8], which includes over 14 million images belonging to 20,000 categories. However, these models, which rely solely on image features to classify waste, are not effective for discriminating between objects with similar features but belonging to different classes.

The rising pressure for effective waste classification has already been seen on the local level. In 2019, the Beijing Municipal government implemented mandatory waste management regulations: households and institutions must sort their waste into recyclables, food waste, other, and hazardous material. “Individuals who fail to follow the regulations repeatedly will be fined a maximum of [~30 dollars]” [9]. In accordance with these new regulations, the International School of Beijing (ISB) installed a new trash sorting system that replaced vague landfill and recycle class trash cans with multiple bins: cans, landfill, paper, plastic, and tetra pak. Almost a year later, waste audits reveal (Table I lists the results of the most recent waste audit) that students are not responding to numerous education initiatives employed by the school’s environmental organization. This again stresses the need for a more reliable method of classifying waste at the consumer level.

| Class | Percentage Correct |
|----------|--------------------|
| Plastics | 28.10 |
| Cans | 89.13 |
| Paper | 46.67 |
| Other | 33.09 |

TABLE I

JANUARY 2020 WASTE AUDIT RESULTS

In light of these issues and observations, this study develops a deep convolutional neural network model, ThanosNet, for trash classification which incorporates metadata to improve existing trash sorting systems.

The contributions of this study are as follows:

- 1) Curated ISBNet dataset that includes 889 images belonging to 5 classes. Each picture contains meta labels identifying the location of the trashcan, activity of the trash can with respect to the time of day.
- 2) The development of our network ThanosNet.
- 3) A proof-of-concept that incorporating metadata into the trash classification model can improve precision. Experiments were conducted to demonstrate the performance differences between current state-of-the-art

models and ThanosNet which utilizes meta-labels to make classification decisions.

II. RELATED LITERATURE

Yang and Thung [7] curated the Trashnet dataset in 2016. This dataset contains approximately 2500 images of trash across six classes (cardboard, glass, metal, paper, plastic, and trash). Each class contained approximately 400-500 images that were taken against a monochromatic background. To introduce variance in the dataset, the lighting and pose between images were modified. Data augmentation techniques including random translations, rescaling, shearing, and rotation were applied to further increase the variance of the dataset. The researchers proposed two novel methods for classifying trash: support vector machines and convolutional neural networks. These methods achieved a test accuracy of 63% using a 70/30 random training/testing split.

The small size of Trashnet motivated Knowles et al. [10] to utilize transfer learning techniques with deep CNN models pre-trained on the ImageNet dataset [8]. Transfer learning for image classification uses a pre-trained model as a feature extractor to extract lower-level features such as edges and lines. Trainable fully-connected layers are then added to classify these features. This enables researchers to train large CNN models with millions of parameters using a small dataset like Trashnet. Knowles et al. utilized the pre-trained weights of the VGG-19 [11] network. In addition to the images in Trashnet, Knowles et al. created a non-waste object class by taking images from the Flowers dataset by the Visual Geometry Group and PASCAL VOC 2012 [12].

Aral et al. [13] further experimented with the efficacy of various transfer learning architectures with established CNN-based models such as DenseNet [14], Inception-ResNet-V2 [15], and Xception [16]. Training, testing, and validation were done in a ratio of 70:17:13, respectively, using Trashnet. Based on their experimental results, DenseNet121 and DenseNet169 performed the best, achieving 95% test accuracy, while Inception-V4 was a close second with 94%. The team also experimented with Adam and Adadelta optimizers, and found that the Adam optimizer resulted in higher test accuracies. Therefore we utilized DenseNet169 with the Adam optimizer as a baseline to demonstrate the effectiveness of our model, ThanosNet, against the current state-of-the-art networks.

Vo et al. [17] continued the trend of transfer learning-based architectures with their DNN-TC model. DNN-TC utilizes ResNext-101 [18] as a feature extractor with the addition of two fully-connected layers following the global average pooling layer. The team also produced their own VN-trash dataset, which consists of images found online and taken in the surrounding environment. It covers the classes of medical waste, organic, and inorganic wastes. This model achieved an accuracy of 94% and 98% when using a 60:20:20 split for training, validation, and testing, respectively, on Trashnet and VN-trash datasets.

A publication most similar to this one is White et al. [19] WasteNet uses DenseNet architecture with fully-connected

layers added on top. A hybrid tuning method was used by first pre-training the classifier layers. Once the performance of these layers began to converge, the remaining layers were unfrozen and a smaller learning rate was applied to calibrate these lower-level feature extractors. The team chose a 50:25:25 split of training, validation, and testing, respectively, using images from the Trashnet dataset. They used a combination of random translation, zooming, shearing, and rotation to augment the images. After training over more than 1000 epochs, WasteNet achieved a F_1 score of 0.970. These experiments showed again the potential of DenseNet [14], and further encouraged us to evaluate our model against a DenseNet feature extractor.

Previous waste classification systems that incorporate CNNs [13], [7], [17], [19], [20], [10] rely on a purely image-based approach. The approach described in this study is designed to be deployed in a modified consumer trash can. This allows our model to use metadata that is associated with the physical trash can including, but not limited to, location, time of day and weight of the trash. In turn, the smart trash can could effectively sort the trash into 5 high level categories: cans, landfill, paper, plastic, and tetra pak.

III. EXPLORATORY DATA ANALYSIS

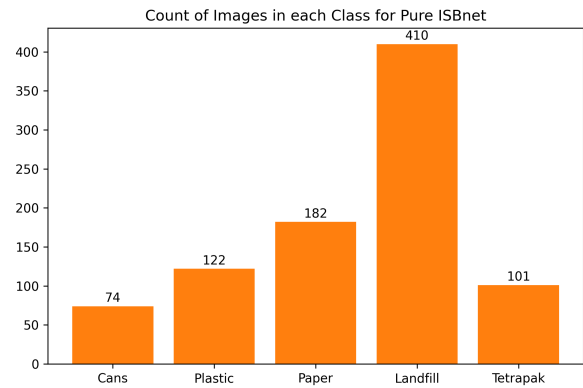


Fig. 1. Class Distribution of ISBNet

ISBNet is hand collected by our group at the International School of Beijing. These images were gathered from trash cans around the school. ISBNet totals 889 images distributed across 5 classes: cans, landfill, paper, plastic, and tetra pak. The distribution of these classes is shown in Figure 1. The data acquisition process involved using a piece of black poster paper as a background; this would create enough contrast for trash belonging to the paper category. These pictures were taken with an iPhone 8 and an iPhone XS. We recorded the trash can in which the piece of trash originated from, and any trash-generating landmarks near this trash can. Section IV details the encoding and formatting of these meta labels. Data augmentation techniques were performed on the images due to the limited size of each class. This included grey-scaling, random rotation, re-scaling, and shearing. Mean subtraction and normalization were also performed on the dataset.

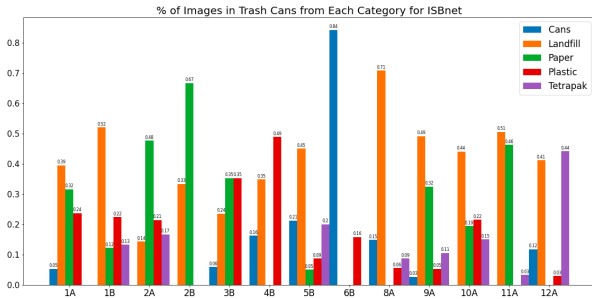


Fig. 2. Distribution of images belonging to each trash class within the individual trashcans. Values are denoted in percentages calculated as the percentage of total photos in the trash can represented by the class.

As illustrated in Figure 2, the variance between each trash category within the individual bins is subtle. There isn’t enough substantive information for a model running purely on metadata to outperform state-of-the-art image-based models. However, the variance in category distributions between trash cans is much more meaningful. Therefore, a model that synergizes metadata and image-based features may exploit the information present in inter-trash can variance to improve classification accuracy.

IV. METADATA

Metadata of all kinds can be collected through sensors in a smart trash can where our model could be implemented, such as location of the trash can and its distance to landmarks, or time of day. The geographical location of a trash can allows us to identify its proximity to certain landmarks. We define a landmark, in the context of trash classification, as an identifiable area that skews the distribution of the type of trash and amount of trash found in a proximal trash can. Landmarks may affect trash generation either through the inherent nature of these landmarks, or through the increased foot traffic experienced by these areas. Examples of landmarks that we identified are cafeterias (eating may produce more contaminated and food-related trash), printers (recyclable paper would be more common next to a printer), or entrances/exits (the large flow of people means more trash is likely to be deposited in the nearby bins). Time of day refers to the time when a piece of trash was thrown into the bin. The time of day may also skew the probability that particular categories of trash are thrown away, such as lunch time, when we would expect more landfill trash.

Incorporating metadata as extra inputs to an image-based neural network decreases the likelihood for *feature confusion*. Items of trash belonging to different categories may have similar features. A network that solely depends on image features is often not able to differentiate between these objects. This is exemplified in Figure 3.

Saliency maps [21] of an image-based trash classifier were generated for images of paper and tissues/napkins, which belong to the landfill class. These are two of the classes with lowest precision for image-based classifiers. The trained classifier shown in Figure 3 incorrectly predicted both landfill pictures, the napkin and tissue, as belonging to the paper

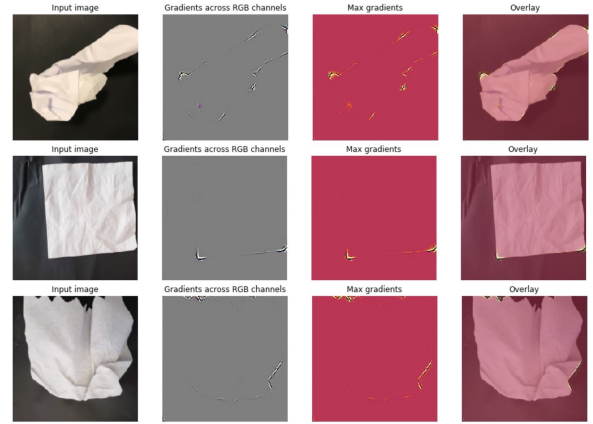


Fig. 3. Saliency maps from a baseline ResNet50 classifier. The images from top to bottom: crumpled piece of paper (paper), a napkin (landfill), crumpled piece of tissue (landfill).

class. The saliency maps illustrated that the image-based model falsely associated rigid edges and crumples in the tissue and napkin with the paper class. Exposing the network to additional time and location information will increase its ability to discern images of similar features, decreasing the likelihood for feature confusion.

A. Location and Distance

Thanks to the ISB administration, we acquired detailed, scaled blueprints, with trash can locations marked, of the two floors where ISBNet was collected. The limited number of trash cans meant trash belonging to different landmarks was grouped together. Figure 4 showcases trash can 8A, an example of this, which received a large amount of trash due to its proximity to a variety of significant landmarks: the cafeteria, library, a lounge, a stairwell, a printer, and bathroom. In turn, a substantial quantity of trash generated from those landmarks was present in 8A, and thus assigned the corresponding trash can metadata. This introduced a larger degree of homogeneity to the metadata than preferred. However, this issue is alleviated by introducing landmark distances, which is introduced later in this subsection.

Using these blueprints, we identified 11 types of trash-generating landmarks around the school. These landmarks are listed below.

- Entrance/Exit
- Bathroom
- Lounge
- Stairwell
- Cafeteria
- Theater
- Gym
- Pool
- Printer
- Couch Area
- Library

A trash can’s *proximal landmarks* are defined as the landmarks in the set of all landmarks listed above that are within

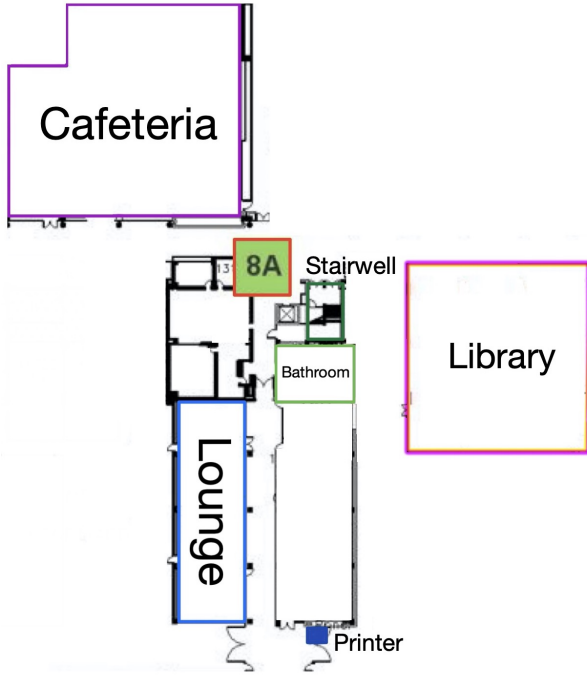


Fig. 4. Trash can "8A" with a cafeteria, stairwell, bathroom, printer, lounge, and library as proximal landmarks.

a 70 meters radius of the trash can. Each trash can was annotated with a one-hot encoding vectors that represents its proximal landmarks.

A second vector, of the same length, described the distances between the bin and each proximal landmark. The distance was defined as the diagonal distance from the center of the bin to the center of the landmark.

We converted these two encoded-vectors into inputs for the network by performing the following process:

Let $\vec{v}_d \in \mathbb{R}^{11}$ and $\vec{v}_{oh} \in \{0, 1\}^{11}$ be the distance between the landmark and the trash can and the one-hot encoded vector of the proximal landmarks, respectively.

$$\vec{v}_c = \vec{v}_d \odot \vec{v}_{oh}$$

The zero entries of the unit vector of \vec{v}_c , \hat{v}_c are replaced with a constant β . We found that the optimal value for β is 10^4 . After, a component-wise negative logarithmic transformation was applied to the unit vector \hat{v}_c . This transformation is reflective of the tendency for consumers to throw away trash in the trash can closest to the trash-generating landmark. The convergence of this logarithmic transformation is shown below:

$$\begin{aligned} \lim_{\hat{v}_{c_i} \rightarrow 0} -\log \hat{v}_c &= \infty \\ \lim_{\hat{v}_{c_i} \rightarrow \beta} -\log \hat{v}_c &= -\infty \end{aligned}$$

The relative influence of a landmark decays exponentially as the distance between the landmark and the trash can increases linearly. Non-proximal landmarks that have an influence of β approach $-\infty$. These negative values, in the

transformed vector, help the network further discriminate between confusing features.

B. Time

The 8am - 6pm school day was segmented into one hour segments. The time when the trash was thrown into the trash can is represented as a unit vector $\hat{v}_t \in \mathbb{R}^{11}$, with the value of each column denoting the likelihood that trash was discarded during this one-hour interval.

We were unable to capture the exact time of day when each individual piece of trash was disposed of. Thus, for each trashcan that we collected from, we identified multiple time periods throughout the day where these trash cans would be most active and exposed to the highest level of foot traffic. Each landmark is categorized into one of three categories: multi-modal, normal, and uniform. Landmarks characterized as multi-modal demonstrate increased foot traffic during multiple, regular, scheduled times of day. For example, the cafeteria is mostly accessed during lunch periods and after school, the lounge is accessed by teachers primarily during lunch, and entrances/exits are accessed by all personnel at the start and end of the school day, 8am and 3pm respectively. Landmarks characterized as normal experience activity clustered around a central mean, adhering to a normal distribution. Landmarks that follow such a distribution include the theater which, on regular days, only experiences traffic after lunch time for assemblies. Landmarks that are described by uniform distributions are accessed throughout the day with similar levels of activity, such as bathrooms and printers.

In order to create the probability distributions for each landmark, the one-hour time intervals between 8am and 6pm were converted into values between 0 and 1. Namely, 8am was considered 0, 1pm (13:00) was considered 0.5, and 6pm (18:00) was considered 1.0. Multi-modal probability distributions were constructed through the addition of normal distributions. Normal distributions were created with means centered on the high activity times; standard deviations were determined as the approximate time before the mean that foot traffic begins a significant uptick. An example is the cafeteria multi-modal distribution. We estimated that the times of highest activity were 12pm (around lunch time, represented as 0.4) and 4pm (after school, represented as 0.8), with a standard deviation of 1 hour (0.1) and 30 minutes (0.05), respectively. This was used to create two normal distributions, which were combined by averaging. The probability distributions for landmarks that have normally-distributed foot traffic were created using the same procedure outlined above. Uniform probability distributions have a value of 1 through all time intervals.

Each bin was assigned a convolutional probability distribution, which is the average of all waste-generating probability distributions from the bin's proximal landmarks. Consider the example below for the bin b_m which has n proximal landmarks. Its corresponding probability density function (PDF) is defined as $f_m(x)$. b_m 's probability density function

with respect to time can be described using the following:

$$f_m(x) = \sum_i^n \frac{1}{n} f_i(x)$$

where f_i is the PDF for landmark i . Finally, the time vector was determined by evaluating the area under the trash can PDF for each time interval. This is recorded into the corresponding column of the time vector \vec{v}_t . This process is shown below for time intervals $[a, b]$ and $f_m(x)$:

$$\vec{v}_t[i] = \int_a^b f_m(x) dx$$

where $\vec{v}_t[i]$ is the value that denotes the area of the time interval $[a, b]$. This vector is then normalized into a unit vector, \hat{v}_t . Thus, the final length of the time vector is 10.

V. EXPERIMENTS

This section is organized into four subsections. The first section, *Weighted Loss Function*, explains and justifies the loss function we employed. In the second section *Baseline Experiments*, transfer learning methods are used with pre-trained ImageNet models to establish a baseline score for comparison against ThanosNet. Later, *Metadata Experiments* goes into detail regarding the architecture and performance of our ThanosNet model. The last section, *Results*, analyzes the performance ThanosNet relative to the baseline models established in the *Baseline Experiments* subsection.

For all of our experiments, we utilized stratified cross-validation training with a fold count of 5, each fold being trained over 50 epochs. The performance of each model was evaluated through average maximum validation macro F_1 in each fold, and the corresponding average validation loss in the same epoch. As the dataset was highly imbalanced, we used macro F_1 as the validation metric to gauge the precision and recall of our model, which is defined by the following,

$$F_1 = 2 \cdot \frac{pr \cdot re}{pr + re}$$

where pr and re represent precision and recall, respectively.

A. Weighted Loss Function

For this multi-class classification model, let X denote the input image and metadata. The scalar $y_{label} \in \{0, 1, 2, 3, 4\}$ denotes the label representing the class that input X belongs to: cans, landfill, paper, plastic, and tetra pak respectively. $\vec{y}_{predict} \in \mathbb{R}^5$ represent the models prediction of the input X . Component i of the prediction vector $\vec{y}_{predict}$ is indexed by $\vec{y}_{predict}[i]$. The standard cross-entropy loss function is defined as the following:

$$L(\vec{y}_{predict}, y_{label}) = -\vec{y}_{predict}[y_{label}] + \log\left(\sum_j e^{\vec{y}_{predict}[j]}\right)$$

The trained model defined by the standard cross-entropy loss function above shows poor performance as it demonstrates a poor recall rate. The reason stems from the fact that ISBNNet is imbalanced. Therefore, a weighted loss function

was used in place of the standard cross-entropy loss, where the weights of a class are inversely proportional its class size. The weight of class i is defined by

$$\omega_i = \frac{\sum_j^n \|j\|}{\|i\|}$$

where n denotes the set of all classes: cans, landfill, paper, plastic, and tetra pak. $\|j\|$ is the size of class j , and $\|i\|$ is the size of class i .

Thus, the weighted cross-entropy loss function is defined by the following:

$$L(\vec{y}_{predict}, y_{label}) = \omega_{y_{label}} \left(\log\left(\sum_j e^{\vec{y}_{predict}[j]}\right) - \vec{y}_{predict}[y_{label}] \right)$$

In our experiments, we discovered that using such a weighted loss function resulted in better performance.

B. Baseline Experiments

We experimented with VGG16, ResNet50 [22], and DenseNet169 as feature extractors for our image-based, baseline models. These are networks that have performed well in prior literature [19], [13], [17], [20] and thus are a valuable benchmark for comparison. The pre-trained ImageNet model had its respective classification layers removed and replaced with three fully connected layers with ReLU activation functions in between. During training, regularization techniques including batch normalization, dropout, and l_2 regularization were applied. We trained with a batch size of 32. The Adam optimizer was used with a learning rate of 10^{-5} and weight decay of 10^{-10} .

All image inputs were first resized to 256×256 , then center-cropped to 224×224 to match the ImageNet feature extractor input parameters. These images were then normalized based on the mean and standard deviation in the ImageNet training set: $\mu_r = 0.485, \mu_g = 0.456, \mu_b = 0.406$ and $\sigma_r^2 = 0.229, \sigma_g^2 = 0.224, \sigma_b^2 = 0.225$. A random-resized crop was experimented with during training instead of the center crop. This resulted in an extremely volatile training F_1 score, preventing the model from converging. This suggests that object localization plays an important factor in classification as there was a large variance in trash size. The random resized crop could be cropping out important features.

C. Metadata Experiments

To incorporate metadata into the network, two attachments to the existing baseline networks were proposed: a bilinear [23] attachment, and an additive attachment. The architecture of these two variants of ThanosNet, AdditiveThanosNet and BilinearThanosNet, are shown in Figure 5 and Figure 6, respectively.

Each of these two variants seek to replicate an intuition behind small-scale trash classification. Metadata, such as the location and time distribution of the trash can, provides information that inherently skews the prediction of a trash item. Therefore, the first variant, AdditiveThanosNet, replicates

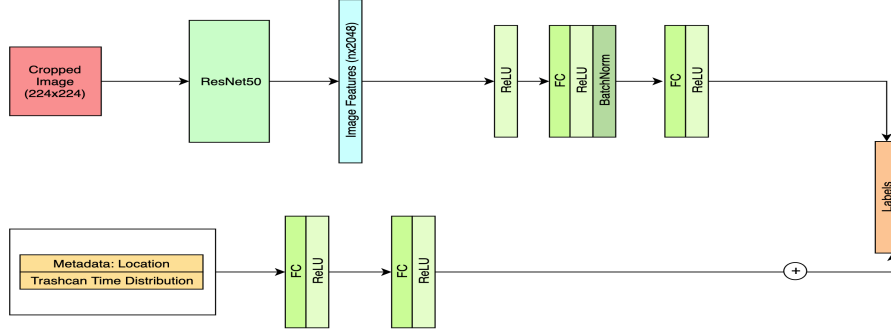


Fig. 5. ThanosNet with additive metadata attachment

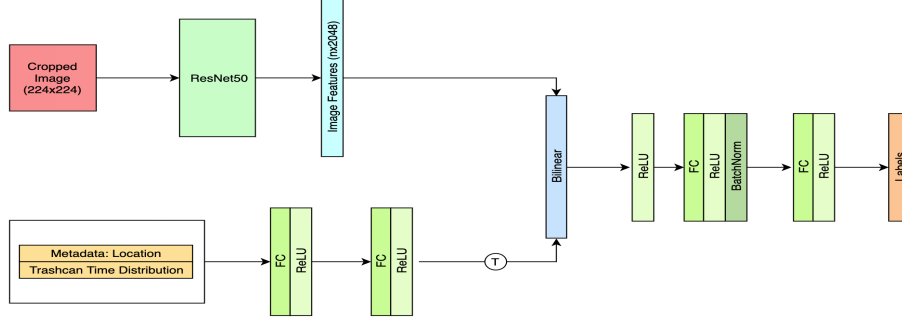


Fig. 6. Thanosnet with bilinear metadata attachment

this inherent bias by using the additive attachment as a bias parameter for the prediction layer.

The second, ThanosNet with a bilinear attachment, captures the intuition that for a given image, the importance of its extracted features is correlated to the context provided by location and time metadata. Furthermore, this bilinear model uses the meta-labels to create an attention mask, forgetting relatively ambiguous, irrelevant features while retaining features that contain more pertinent information. In this bilinear model, the outer product between the transpose of the metadata features and the image features was performed to create a bilinear matrix. Principal component analysis was then applied on this bilinear matrix. The bilinear layer is defined by the following:

$$y = x_1^T A x_2 + b$$

where x_1 represents features extracted from the metadata, x_2 represents image-based features, A is a parameter matrix, and b is the bias of this layer.

For both networks, the metadata input was created by concatenating the location meta-label and time meta-label.

D. Results

As seen in Table II, ResNet50 achieved the best results out of the three baseline ImageNet models with a 0.9240 average 5-fold validation macro F_1 score, while DenseNet169 and VGG16 achieved 0.8750 and 0.9198, respectively. Therefore, we used ResNet50 as the feature extractor for ThanosNet.

However, it should be noted that the baseline VGG16 and DenseNet169 networks were able to converge at a faster rate.

Of the two ThanosNet variations, the bilinear variant performed the best, achieving a F_1 of 0.952 compared to the 0.907 from the additive model. We observe that BilinearThanosNet had the best performance in terms of loss, macro F_1 score, and accuracy, while AdditiveThanosNet was comparable with ResNet50 and DenseNet169. The performance difference between these two variants of ThanosNet is attributed to the subtle variance between the class distributions of each trash can. This hinders the additive bias to significantly impact the predictions of the network positively. We believe that the improvements of BilinearThanosNet come from having an attention-like mask at the level of the feature extractor. The increased distance between the feature layer and prediction layer gives the network more opportunity to correct errors from the attention mask and reduce the effect of noise from the data.

As shown in Figures 7 and 8, BilinearThanosNet

| Model | Loss | Macro F_1 | Accuracy |
|-------------------|-------|--------------|----------|
| VGG16 | 0.406 | 0.875 | 0.875 |
| ResNet50 | 0.273 | 0.924 | 0.920 |
| DenseNet169 | 0.287 | 0.920 | 0.918 |
| AdditiveThanosNet | 0.290 | 0.907 | 0.906 |
| BilinearThanosNet | 0.161 | 0.952 | 0.947 |

TABLE II

BASILINE MODELS AND THANOSNET MODELS RESULTS

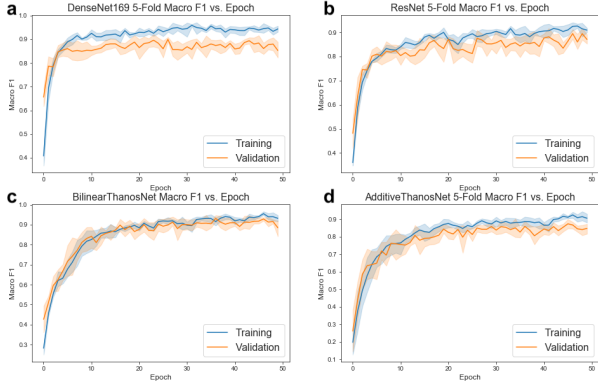


Fig. 7. The macro F_1 in the training (blue) and validation (orange) process for a) DenseNet169, b) ResNet50, c) BilinearThanosNet, and d) AdditiveThanosNet.

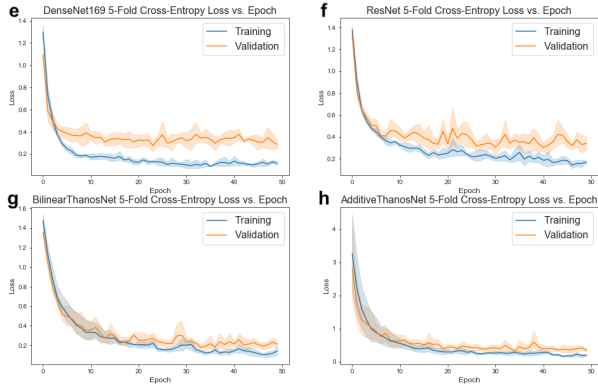


Fig. 8. The loss in the training (blue) and validation (orange) process for e) DenseNet169, f) ResNet50, g) BilinearThanosNet, and h) AdditiveThanosNet.

shows good stability, unlike ResNet50, the second highest-performing model, which struggled with volatility. We believe that this is induced by the inherent noise of the dataset and augmentations that were applied to the training set. For object-dense categories including landfill and paper, where there is a low sample to features ratio, the image-based networks struggle to generalize these features sufficiently. However, since BilinearThanosNet utilizes metadata it is not nearly as dependent on these image features. Therefore, the usage of metadata not only improves performance, but also stabilizes training by mitigating noisy, object-dense classes.

The confusion matrix of the experimental models are displayed in Figure 9. From this, we can see that BilinearThanosNet outperforms the best image-based model, ResNet50, in almost all trash categories. In particular, BilinearThanosNet demonstrates a significant improvement in paper classification over ResNet50 with 0.983 precision and 0.978 recall rate. Since the paper class is prone to feature confusion and has a low sample to features ratio, metadata played an important role in providing additional contextual information that clarifies whether the paper is recyclable or not. Again, this reduces BilinearThanosNet's dependence on potentially noisy image features within the paper category and improves classification performance.

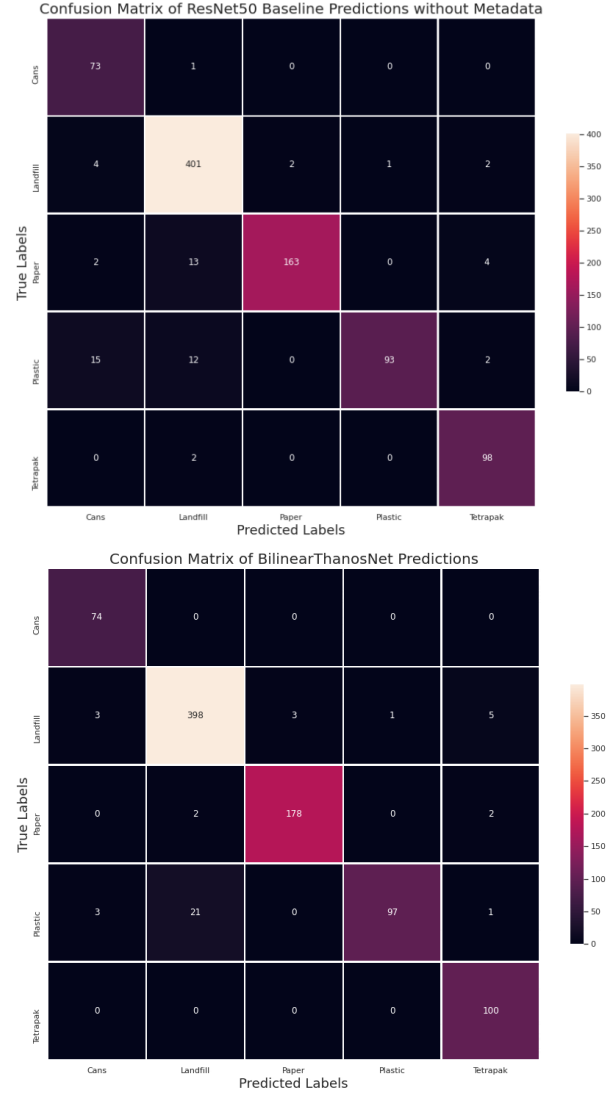


Fig. 9. Confusion matrix of ResNet50 model without incorporating metadata and BilinearThanosNet respectively, generated by combining validation predictions from each model in each fold.

VI. CONCLUSION

In this study, we contributed a novel method for trash classification through our ThanosNet model, and compiled the 889 images and associated metadata that forms the ISBNet dataset. We demonstrated the effectiveness of ThanosNet and the associated concepts through comparisons against image-based, state-of-the-art models for trash classification. The results show ThanosNet outperforming the other models with an F_1 score of 0.952.

ThanosNet has important implications and applications in the field of trash classification. In particular, the creation of a smart trash can for consumer use is promising, as it may be a solution to the difficulties consumers encounter when trying to classify recyclable trash. Successful implementation and deployment of such technology could result in visible improvements in recycling and trash management within

various communities. Moreover, the idea of an attention-based classification system can be utilized by trash classification applications at the macroscopic level. Recycling plants could geo-tag incoming trash with landmarks labels, such as residential communities, industrial parks, shopping centers, and hospitals. These meta labels would further improve proposed image-based classification methods.

For future experimentation and research, we will continue to expand ISBNet to include more images and relevant fields of metadata. A larger, more comprehensive ISBNet dataset that is balanced will likely improve the performance of models trained and tested on it. Such a dataset would also be beneficial for creating accurate deployment models, as it would more closely model real-world, trash-generating probability distributions. Furthermore, we believe that for time metadata, landmarks that possess multi-modal distributions cannot be effectively represented with time intervals that are one hour long. Therefore, collecting the "real" time of when the garbage was disposed, possibly as a scalar, would expel the need for estimating complex probability distributions.

ACKNOWLEDGEMENT

We thank George Lin Wu, Minkyu Colin Jung, and Andy Kim for labelling and photographing all of the images that constructed ISBNet. We thank Hannah Graham, Hyoree Kim, and Alex Zheng for bringing this issue to light and inspiring the development of this project. This work was supported by the International School of Beijing's branch of the Net Impact organization.

REFERENCES

- [1] P. Jain, A. Jain, R. Singhai, and S. Jain, "Effect of Biodegradation and Non-Degradable Substances in Environment," *International Journal of Life Sciences*, pp. 58–64, 2017.
- [2] M. Nizar, E. Munir, Irvan, and F. Amir, "Examining the Economic Benefits of Urban Waste Recycle Based on Zero Waste Concepts," *Advances in Social Science, Education and Humanities Research*, pp. 300–309, 2018.
- [3] M. Abkenari, A. Rezaei, and N. Pournayeb, "Recycling Construction Waste Materials to Reduce the Environmental Pollutants," *International Journal of Architectural, Civil and Construction Sciences*, pp. 1138–1142, 2015.
- [4] Y. Luo, I. Zelenika, and J. Zhao, "Providing immediate feedback improves recycling and composting accuracy," *Journal of Environmental Management*, pp. 445–454, 2019.
- [5] I. Salimi, B. S. Bayu Dewantara, and I. K. Wibowo, "Visual-based trash detection and classification system for smart trash bin robot," in *2018 International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC)*, pp. 378–383, 2018.
- [6] J. Donovan, "Auto-Trash sorts garbage automatically at the TechCrunch Disrupt Hackathon," 2016.
- [7] M. Yang and G. Thung, "Classification of Trash for Recyclability Status," pp. 1–6, 2016.
- [8] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.
- [9] M. Ming, "China Focus: Beijing to implement citywide mandatory household garbage sorting," 2019.
- [10] J. Knowles, S. Kennedy, and T. Kennedy, "OscarNet: using transfer learning to classify disposable waste," 2016.
- [11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2015.
- [12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results." <http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html>.
- [13] R. A. Aral, S. R. Keskin, M. Kaya, and M. Hacıömeroğlu, "Classification of TrashNet Dataset Based on Deep Learning Models," *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, pp. 2058–2062, 2019.
- [14] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2016.
- [15] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, inception-resnet and the impact of residual connections on learning," *CoRR*, vol. abs/1602.07261, 2016.
- [16] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *CoRR*, vol. abs/1610.02357, 2016.
- [17] A. H. Vo, L. Hoang Son, M. T. Vo, and T. Le, "A Novel Framework for Trash Classification Using Deep Transfer Learning," *IEEE Access*, vol. 7, pp. 178631–178639, 2019.
- [18] S. Xie, R. B. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," *CoRR*, vol. abs/1611.05431, 2016.
- [19] G. White, C. Cabrera, A. Palade, F. Li, and S. Clarke, "WasteNet: Waste Classification at the Edge for Smart Bins," 2020.
- [20] Z. Yang and D. Li, "WasNet: A Neural Network-Based Garbage Collection Management System," *IEEE Access*, vol. 8, pp. 103984–103993, 2020.
- [21] A. Z. Karen Simonyan, Andrea Vedaldi, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," 2013.
- [22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.
- [23] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 12 2015.