

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 bin, 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 bin, 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 bin 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 bin 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 existing vague landfill and recycle trash bins 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 trash bin, activity of the trash bin 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 bin 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 AND MOTIVATION

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]. Based on their experimental results, DenseNet121 and DenseNet169 performed the best, while Inception-V4 was a close second.

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.

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 state-of-the-art results on TrashNet.

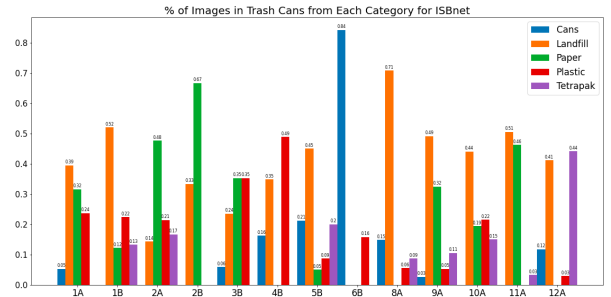


Fig. 1. Distribution of images belonging to each trash class within the individual trash bins. Values are denoted in percentages calculated as the percentage of total photos in the trash bin represented by the class.

Previous waste classification systems that incorporate CNNs [13], [7], [17], [19], [20], [10] rely on a purely image-based approach. However, in a real environment, classifiers that sort waste into high-level categories, such as plastic, cans, paper, and landfill, are subject to many diverse features from the variety of objects present. As a result, pure image-based classifiers are vulnerable to low generalizability and feature confusion, explained in a later section. Moreover, purely image-based approaches assume that distributions between classes and objects of that class are uniform across all locations and time. However, intuitively, these factors have a large impact on said distributions. For example, during meal times, we would expect an influx of trash belonging to the landfill class (food scraps, wrappers, plastic utensils). Alternatively, for a trash bin next to a printer, we would expect a large amount of recyclable paper being disposed of.

In a response to these limitations, our model utilizes meta-data that is associated with the physical trash bin including, but not limited to, location, time of day and weight of the trash. Metadata provides context for an image and information that reflects the likely distributions of trash within a trash bin at given points in time. As illustrated in Figure 1, there is a distinct difference in trash distribution between the various trash bin locations. Therefore, a model that synergizes metadata and image-based features may exploit the information present in inter-trash bin variance to enhance classification capabilities.

We envision the approach described in this study to be deployed in a modified consumer trash bin. In turn, the smart trash bin could effectively sort trash into 5 high-level categories: cans, landfill, paper, plastic, and tetra pak. By using metadata, our model is not as dependent on image features as current state-of-the-art waste classification models, improving precision for object-dense classes and overall accuracy post-implementation.

III. EXPLORATORY DATA ANALYSIS

ISBNet is hand collected by our group at the International School of Beijing. The trash in these images was gathered from trash bins 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

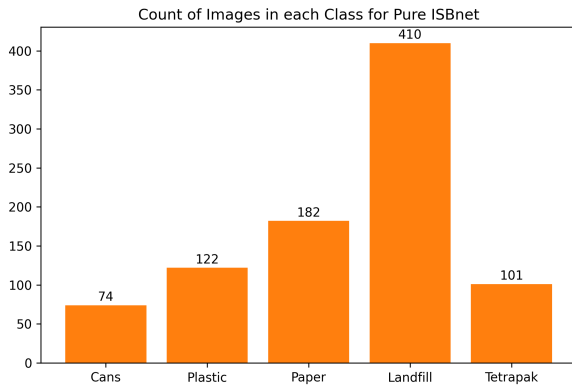


Fig. 2. Class Distribution of ISBNet

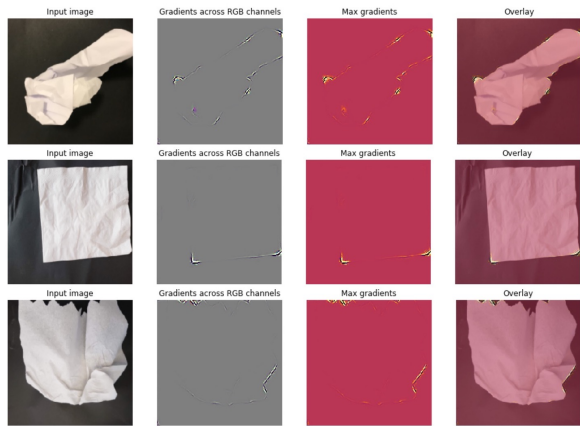


Fig. 3. Saliency maps from a baseline ResNet50 classifier. The images from top to bottom: crumpled piece of paper (paper), a napkin (landfill), crumpled piece of tissue (landfill).

Figure 2. 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 bin in which the piece of trash originated from and any trash-generating landmarks nearby. 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.

IV. METADATA

Metadata of all kinds can be collected through sensors in a smart trash bin where our model could be implemented, such as location of the trash bin and its distance to landmarks, or time of day. 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 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.

We used two fields of metadata in our network: location and distance, as well as time. The methodology implemented to transform these fields into inputs for our network is outlined in the following subsections: *Location and Distance* and *Time-Based Traffic Intensity*.

A. Location and Distance

The geographical location of a trash bin 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 bin. 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).

We acquired detailed, scaled blueprints, with trash bin locations marked, of the two floors where ISBNet was collected. The limited number of trash bins meant trash belonging to different landmarks was grouped together. Figure 4 showcases trash bin 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 bin 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

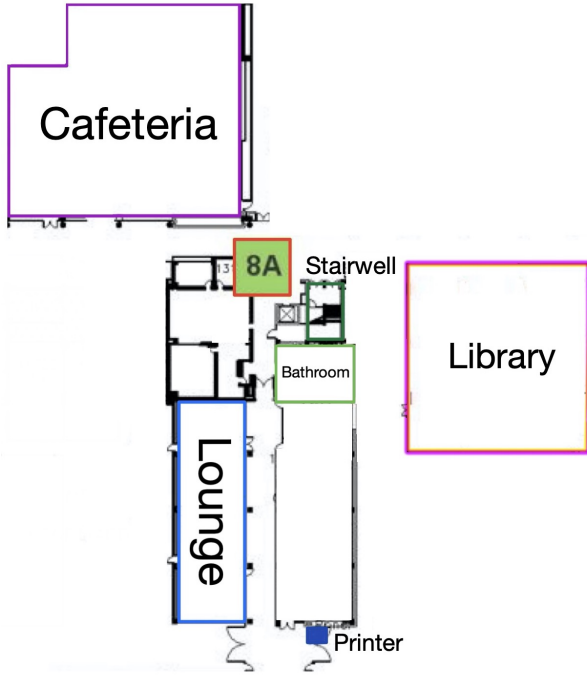


Fig. 4. Trash bin "8A" with a cafeteria, stairwell, bathroom, printer, lounge, and library as proximal landmarks.

- Printer
- Couch Area
- Library

A trash bin's *radius of proximity* is determined as half of the average of its neighboring bins. A bin that qualifies as a neighboring bin is one that is closest in each traffic direction. Thus, a trash bin's *proximal landmarks* are defined as the landmarks in the set of all landmarks listed above that are within its radius of proximity.

In order to input the location and distance information into the network, a few transformations were conducted. Each trash bin was annotated with a binary vector, $\vec{v}_{prox} \in \{0, 1\}^{11}$, that represents its proximal landmarks. Landmarks that exceeded its radius of proximity around the trash bin were indicated by a 0 entry. Another vector $\vec{v}_d \in \mathbb{R}^{11}$ describes the diagonal distance between the center of the trash bin and each landmark. Element-wise multiplication between these two vectors was performed so only distances from proximal landmarks are considered. This process is shown below:

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

The non-zero entries of this new vector, \vec{v}_c , represent the meter distances between the trash bin and its proximal landmarks. The vector \vec{v}_c is normalized into a unit vector using the L^2 norm. This converts the absolute meter distances into relative distances. This unit vector of \vec{v}_c is expressed as \hat{v}_c , where the non-zero relative distances are inversely proportional to their weight.

However, the zero entries in this vector do not represent relative distance, rather they represent the absence of a

proximal landmark. Theoretically, these zero-values describe landmarks that are an infinite distance away. However, for numerical computation, we chose a value β that is sufficiently large to represent this behavior. This β value replaces the zero-values in the unit vector \hat{v}_c .

A component-wise negative logarithmic transformation, using the natural logarithm, was applied to this unit vector \hat{v}_c . This transformation is reflective of the tendency of consumers to throw away trash in the trash bin closest to the trash-generating landmark. Therefore, the relative influence of a landmark decays exponentially as the distance between the landmark and the trash bin increases linearly. This final location vector is concatenated with a time vector representing the time metadata, which is described in the next section.

B. Time-Based Traffic Intensity

Each landmark is associated with a time-based traffic intensity distribution. At different times of day, these landmarks experience different levels of traffic intensity. These relative changes in intensity were modelled with a probability distribution. We categorized all foot traffic distributions into three categories: multi-modal, normal, and uniform.

Landmarks characterized as multi-modal demonstrate increased foot traffic during multiple, regular, scheduled times of day. Consider the following example of the school cafeteria, which can be described using a multi-modal foot traffic distribution:

The cafeteria is mostly accessed during lunch periods and directly after school. These time periods are 12pm and 4pm respectively. Thus, the foot traffic distribution for the cafeteria is composed of two truncated normal distributions. A truncated normal distribution was used as we assumed the cafeteria experiences no traffic before it opens and after it closes, 8am and 6pm, respectively. The means of the two component foot traffic distributions were determined based on periods of highest activity, which were 12pm and 4pm. The variances were estimated by surveying the change in traffic around the cafeteria around the mean times. Let the random variables X_1 and X_2 describe the two truncated normal distributions that compose the foot traffic distribution of the cafeteria. These random variables are initialized such that:

$$X_1 \sim N(\mu = 12\text{pm}, \sigma^2 = (15\text{min})^2, a = 8\text{am}, b = 6\text{pm})$$

$$X_2 \sim N(\mu = 4\text{pm}, \sigma^2 = (15\text{min})^2, a = 8\text{am}, b = 6\text{pm})$$

where X has the condition of $a < X < b$ as a truncated normal distribution.

The probability density functions (PDF) of these two truncated normal distributions are defined as:

$$p_1(x) = \begin{cases} \frac{1}{15\text{min}} \frac{\phi(\frac{x-12\text{pm}}{15\text{min}})}{\Phi(\frac{6\text{pm}-12\text{pm}}{15\text{min}}) - \Phi(\frac{8\text{am}-12\text{pm}}{15\text{min}})} & 8\text{am} < x < 6\text{pm}, \\ 0 & \text{otherwise.} \end{cases}$$

$$p_2(x) = \begin{cases} \frac{1}{15\text{min}} \frac{\phi(\frac{x-4\text{pm}}{15\text{min}})}{\Phi(\frac{6\text{pm}-4\text{pm}}{15\text{min}}) - \Phi(\frac{8\text{am}-4\text{pm}}{15\text{min}})} & 8\text{am} < x < 6\text{pm}, \\ 0 & \text{otherwise.} \end{cases}$$

Here, $\phi(\cdot)$ is the probability density function of the standard normal distribution, while $\Phi(\cdot)$ is its cumulative distribution function.

The composite foot traffic distribution, $p_{\text{cafeteria}}$, of the cafeteria is determined by averaging the two PDFs, p_1 and p_2 . This is shown below:

$$p_{\text{cafeteria}}(x) = \frac{p_1(x) + p_2(x)}{2}$$

An average was used to combine these distributions to preserve the peaks of the respective truncated normal distributions, while maintaining constant area under the PDF.

Landmarks characterized as normal experience activity clustered around a central mean, adhering to a single truncated 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. Consider the following example of a printer: We assume that the foot traffic for all printers follow a uniform probability distribution during the 8am to 6pm school day. Let random variable X_{printer} describe the printer's uniform distribution. This random variable is initialized such that:

$$X_{\text{printer}} \sim \text{Uni}(t_i = 8\text{am}, t_f = 6\text{pm})$$

The probability distribution of this uniform distribution is shown below:

$$p_{\text{printer}}(x) = \begin{cases} \frac{1}{6\text{pm} - 8\text{am}} & 8\text{am} < x < 6\text{pm}, \\ 0 & \text{otherwise.} \end{cases}$$

From here, with the foot traffic distributions for each landmark characterized and determined, the distributions needed to be transformed into a viable input representative of the bin.

First, each bin was assigned a convolutional probability distribution, which is the average of all foot-traffic probability distributions from the bin's proximal landmarks. For a given bin b_m which has n proximal landmarks, its corresponding probability density function (PDF) is defined as $f_m(x)$. b_m 's PDF with respect to time can be described using the following:

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

where $f_i(\cdot)$ is the truncated PDF of foot traffic for the i th proximal landmark. An average is used to consolidate the time-based traffic intensity distributions of the bin's proximal landmarks into one convolutional distribution. This convolutional PDF, that models the bin's time-based traffic intensity, encompasses multiple peaks, such that each peak corresponds to a influx of traffic from one or more of the bin's proximal landmarks. It is intuitive that the total number of bins should drastically out-number the total number of landmarks. ThanosNet uses this to its advantage, and it is able to learn to discriminate between the individual peaks

of the bin's convolutional distribution and their respective proximal landmarks. Further, by averaging these individual proximal landmark traffic intensity distributions, we are able to smooth out the noise present in these estimated traffic intensity distributions.

This convolutional PDF, which models the trash-generating distribution for bin b_m , is discretized by sampling over one-hour intervals in the sample space [8am, 6pm]. These discretized values form a new vector \vec{v}_t , where component i of \vec{v}_t is indexed as:

$$\vec{v}_t[i]$$

where $\vec{v}_t[0]$ is the first component in \vec{v}_t , and so forth.

The components of this vector are initialized by determining the area under this newly defined trash bin convolutional PDF, f_m . This process is described below:

$$\begin{aligned} \vec{v}_t[0] &= \int_{8\text{am}}^{9\text{am}} f_m(x) dx \\ \vec{v}_t[1] &= \int_{9\text{am}}^{10\text{am}} f_m(x) dx \\ &\vdots \\ \vec{v}_t[10] &= \int_{5\text{pm}}^{6\text{pm}} f_m(x) dx \end{aligned}$$

This creates a vector of length 10, where each component corresponds to the probability that a piece of trash is thrown in a specific one-hour interval during an average school day (8am - 6pm).

The vector \vec{v}_t is then normalized into the unit vector using the L^2 norm. The unit vector of \vec{v}_t is denoted by \hat{v}_t .

The time metadata, \hat{v}_t , and location metadata, \hat{v}_c , are concatenated such that:

$$\sum(\hat{v}_t, \hat{v}_c) = (\hat{v}_c[0], \dots, \hat{v}_c[10], \hat{v}_t[0], \dots, \hat{v}_t[11])$$

This vector, of length 21, forms the metadata input for ThanosNet.

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

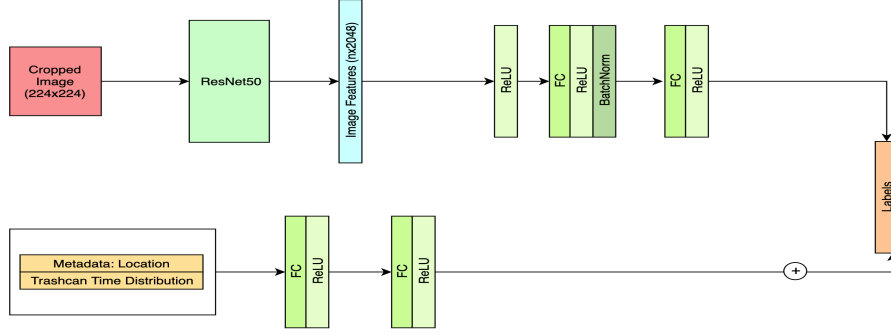


Fig. 5. ThanosNet with additive metadata attachment

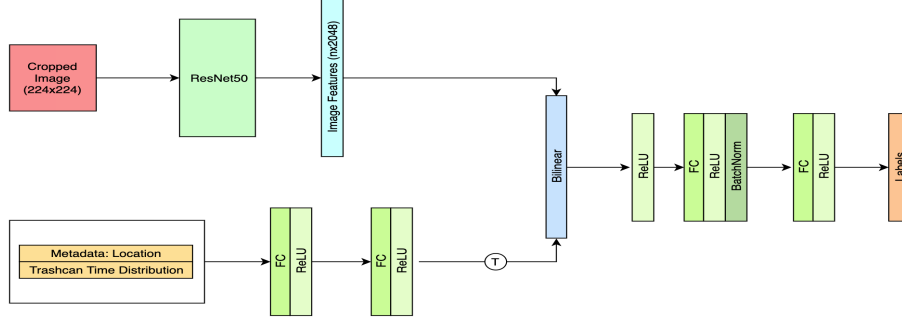


Fig. 6. Thanosnet with bilinear metadata attachment

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

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

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 ISBNet 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:

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

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 bin, provides information that inherently skews the prediction of a trash item. Therefore, the first variant, AdditiveThanosNet, replicates 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

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

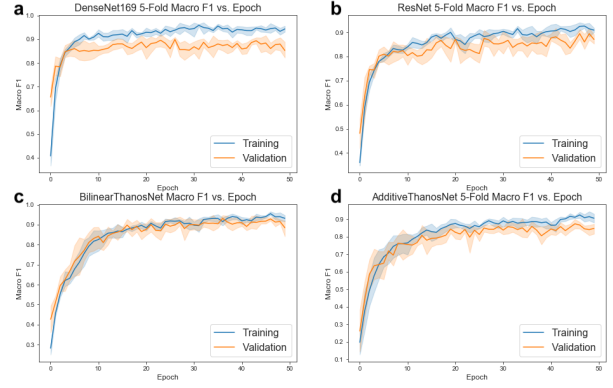


Fig. 7. The macro F_1 in the training (blue) and validation (orange) process for a) DenseNet169, b) ResNet50, c) BilinearThanosNet, and d) AdditiveThanosNet.

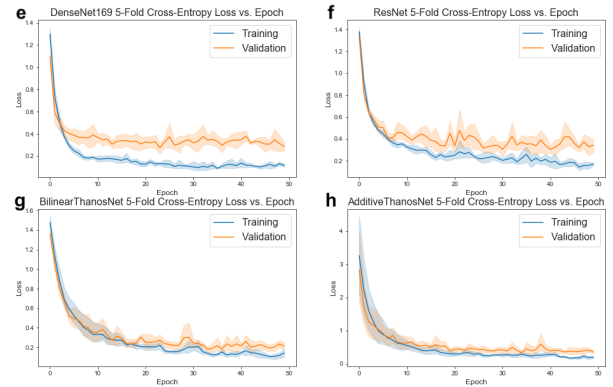


Fig. 8. The loss in the training (blue) and validation (orange) process for e) DenseNet169, f) ResNet50, g) BilinearThanosNet, and h) AdditiveThanosNet.

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 bin. 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 shows good stability, unlike ResNet50, the second highest-performing model, which struggled with volatility. We be-

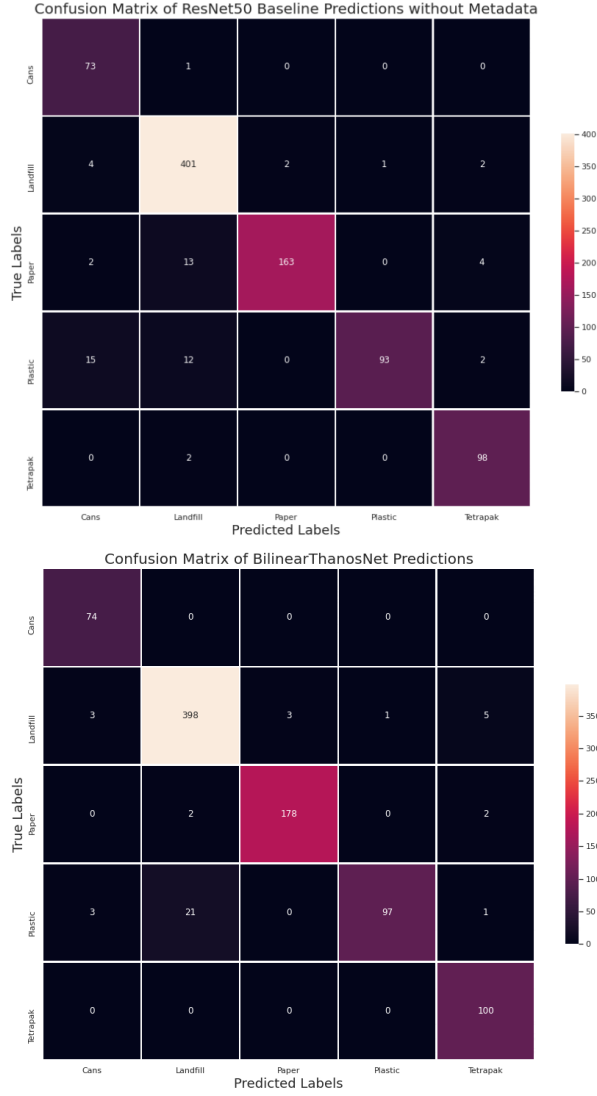


Fig. 9. Confusion matrix of ResNet50 model without incorporating metadata and BilinearThanosNet respectively, generated by combining validation predictions from each model in each fold.

lieve 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.

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 bin 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.