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 [?], while effective waste management reduces the pressure on landfills and can create beneficial economic and financial effects [?]. Reducing total waste is crucial in conserving and reusing resources, and becoming more sustainable [?], 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 [?], 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. [?], 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 [?].

In the same year, Yang and Thung [?] 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[?], 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]” [?]. 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 and Motivation

Yang and Thung [?] 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. [?] to utilize transfer learning techniques with deep CNN models pre-trained on the ImageNet dataset [?]. 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 [?] 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 [?].

Aral et al. [?] further experimented with the efficacy of various transfer learning architectures with established CNN-based models such as DenseNet [?], Inception-ResNet-V2 [?], and Xception [?]. Based on their experimental results, DenseNet121 and DenseNet169 performed the best, while Inception-V4 was a close second.

Vo et al. [?] continued the trend of transfer learning-based architectures with their DNN-TC model. DNN-TC utilizes ResNext-101 [?] 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. [?] 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.

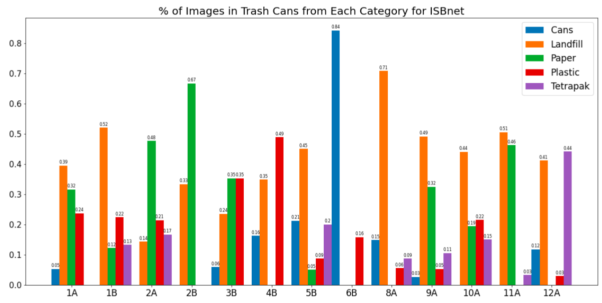

Fig. 1. Distribution of images belonging to each trash class within the individual trashcans. Values are denoted in percentages calculated as the percentage of total photos in the trash can represented by the class.

Previous waste classification systems that incorporate CNNs [?], [?], [?], [?], [?], [?] 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 can 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 metadata that is associated with the physical trash can 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 can at given points in time. As illustrated in Figure 1, there is a distinct difference in trash distribution between the various trash can locations. Therefore, a model that synergizes metadata and image-based features may exploit the information present in inter-trash can variance to enhance classification capabilities.

We envision the approach described in this study to be deployed in a modified consumer trash can. In turn, the smart trash can 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 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

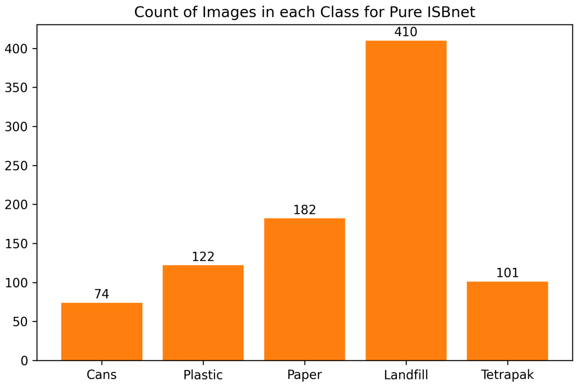

Fig. 2. Class Distribution of ISBNet

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 can 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.

Importantly though, it is not feasible to collect the exact time in which the piece of trash was disposed of. Trash that can be realistically imaged is retrieved from bins where it was disposed of at an unknown time. Regular trash cans also do not possess sensors that can detect trash being thrown into the bin and simultaneously record the time and associated image. Without sensors, the only workaround is to station a person at a trash can who is collecting the trash and recording its information in real time, which is simply not achievable in the vast majority of cases. As a result of this inherent shortcoming that exists when procuring a trash dataset possessing metadata, we approximated time distributions for trash cans that could be learned by the model. This methodology and its underlying theory are discussed in the *Time* subsection within the *Metadata* section.

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. 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.

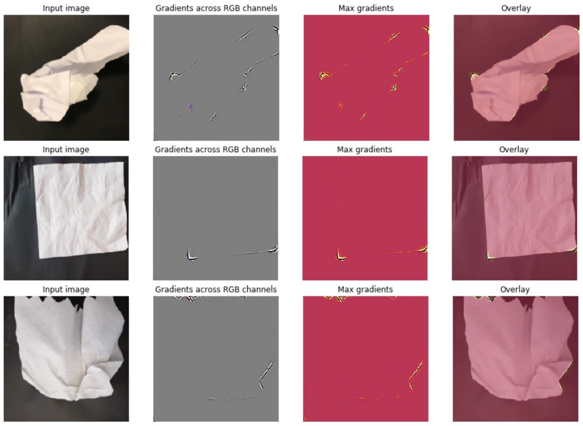

Fig. 3. Saliency maps from a baseline ResNet50 classifier. The images from top to bottom: crumpled piece of paper (paper), a napkin (landfill), cumpled piece of tissue (landfill).

Saliency maps [?] 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*.

A. *Location and Distance*

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).

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

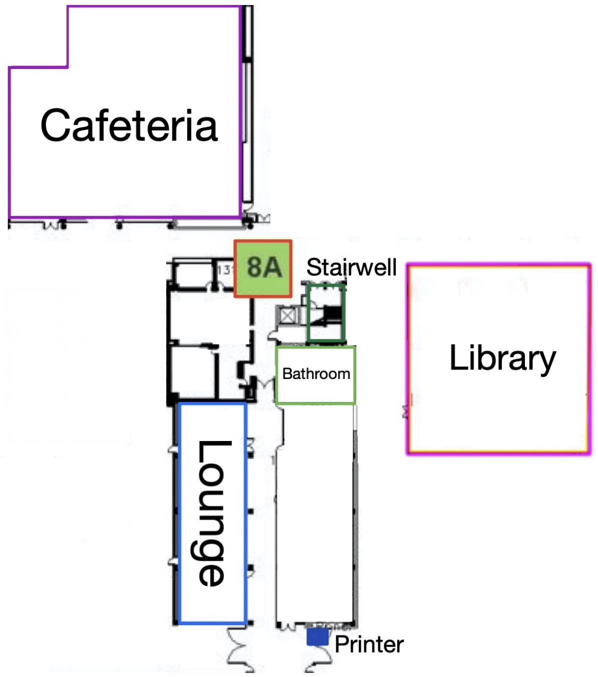

Fig. 4. Trash can "8A" with a cafeteria, stairwell, bathroom, printer, lounge, and library as proximal landmarks.

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 *radius of proximity* is determined as the average of the distances between the trash can and its nearest neighboring bins. Thus, a trash can'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 can 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 can were indicated by a 0 entry. Another vector $\vec{v}_{\mathrm{d}} \in \mathbb{R}^{11}$ describes the diagonal distance between the center of the trash can 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:

$$\vec{v}_{\mathrm{c}} = \vec{v}_{\mathrm{d}} \odot \vec{v}_{prox}$$

The non-zero entries of this new vector, $\vec{v}_{\mathrm{c}}$, represent the meter distances between the trash can and its proximal landmarks. The vector $\vec{v}_{\mathrm{c}}$ is normalized into a unit vector using the L^2 norm. This converts the absolute meter distances into relative distances. This unit vector of $\vec{v}_{\mathrm{c}}$ is expressed as $\hat{\boldsymbol{v}}_{\mathbf{c}}$, 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 $\hat{\boldsymbol{v}}_{\mathbf{c}}$.

A component-wise negative logarithmic transformation, using the natural logarithm, was applied to this unit vector $\hat{\boldsymbol{v}}_{\mathbf{c}}$. This transformation is reflective of the tendency of consumers to throw away trash in the trash can closest to the trash-generating landmark. Therefore, the relative influence of a landmark decays exponentially as the distance between the landmark and the trash can 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

Time of day refers to the time when a piece of trash was thrown into the bin. During noon, we would expect that the large amount of trash generated is associated with lunch time. More likely than not, these pieces of trash would belong to the landfill or tetra pak categories. A piece of trash disposed during noon is highly likely to be associated with the cafeteria. Thus, the significance of a trash can's proximal landmarks is associated with time.

Each landmark is associated with a foot traffic distribution. This foot traffic distribution varies between landmarks. We categorized all foot traffic distributions into three categories: multi-model, 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 modelled through the 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, 9am and 6pm respetively. The means of the two component foot traffic distributions are 12pm and 4pm, while the variances were estimated by surveying the traffic around the cafetera. 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:

$$X_1 \sim N(\mu = 12\text{pm}, \sigma^2 = (15\text{min})^2, a = 9\text{am}, b = 6\text{pm})$$
$$X_2 \sim N(\mu = 4\text{pm}, \sigma^2 = (15\text{min})^2, a = 9\text{am}, b = 6\text{pm})$$

where X conditional on $a < X < b$ has a truncated normal distribution.

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

$$p_1(x) = \begin{cases} \frac{1}{15\text{min}} \frac{\phi(\frac{x-12\text{pm}}{15\text{min}})}{\Phi(\frac{6\text{pm}-12\text{pm}}{15\text{min}})-\Phi(\frac{9\text{am}-12\text{pm}}{15\text{min}})} & 9\text{am} < x < 6\text{pm}, \\ 0 & \text{otherwise.} \end{cases}$$

$$p_2(x) = \begin{cases} \frac{1}{15\text{min}} \frac{\phi(\frac{x-4\text{pm}}{15\text{min}})}{\Phi(\frac{6\text{pm}-4\text{pm}}{15\text{min}})-\Phi(\frac{9\text{am}-4\text{pm}}{15\text{min}})} & 9\text{am} < x < 6\text{pm}, \\ 0 & \text{otherwise.} \end{cases}$$

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

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

With the foot traffic distributions for each landmark characterized and determined, this 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 $f_{\text{m}}(x)$. b_{m}'s probability density function with respect to time can be described using the following:

$$f_{\text{m}}(x) = \sum_{i}^{n} \frac{1}{n} f_{\text{i}}(x)$$

where $f_i(\cdot)$ is the truncated probability density function of foot traffic of the ith proximal landmark.

Consider the following example, where bin b is in proximity to the cafeteria and a printer whose foot traffic distributions are defined previously. The PDF of the trash-generating probability distribution, p_{bin} with respect to time is expressed as:

$$p_{\text{bin}}(x) = \frac{p_{\text{cafeteria}}(x) + p_{\text{printer}}(x)}{2}$$

where $p_{\text{bin}}(x)$ conditional on $9\text{am} < x < 6\text{pm}$ is a truncated probability

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

$$\vec{v}_{\text{t}}[i]$$

where $\vec{v}_{\text{t}}[0]$ is the first componet in $\vec{v}_{\text{t}}$ and so forth.

The components of this vector are initialized by determining the area under this newly defined trash can convolutional PDF, f_{m}. This process is described below:

$$\vec{v}_{\text{t}}[0] = \int_{8\text{am}}^{9\text{am}} f_{\text{m}}(x)dx$$
$$\vec{v}_{\text{t}}[1] = \int_{9\text{am}}^{10\text{am}} f_{\text{m}}(x)dx$$
$$\vdots$$
$$\vec{v}_{\text{t}}[10] = \int_{5\text{pm}}^{6\text{pm}} f_{\text{m}}(x)dx$$

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 $\vec{v}_{\text{t}}$ is then normalized into the unit vector using the L^2 norm. The unit vector of $\vec{v}_{\text{t}}$ is denoted by $\hat{\boldsymbol{v}}_{\text{t}}$.

Approximating the time distribution for specific trash cans through observing foot traffic is perferable over inputting a real-time, time, expressed as a scalar. The limited size of the dataset decreases the probability that the collected data will encompass the probability distribution for each bin. Moreover, although collecting this data during production-time is intuitive, it is unfeasible to acquire this information without existing sensors and infrastructure in place.

An alternative and more generalizable method, described in this section, to account for this a priori probability distribution is needed. By identifying multiple peaks for each bin the model will be able to learn the correlation between each of these peaks and its respective proximal landmarks.

Thus, during production time, the arrival time expressed as a one-hot vector will contain a singular peak. This is then used by the pre-trained model to discriminate between relatively important and unimportant proximal landmarks.

The time metadata $\hat{\boldsymbol{v}}_{\mathbf{t}}$ and location metadata $\hat{\boldsymbol{v}}_{\mathbf{c}}$ are concatenated such that:

$$\sum(\hat{\boldsymbol{v}}_{\mathbf{t}}, \hat{\boldsymbol{v}}_{\mathbf{c}}) = (\hat{\boldsymbol{v}}_{\mathbf{c}}[0], \dots, \hat{\boldsymbol{v}}_{\mathbf{c}}[11], \hat{\boldsymbol{v}}_{\mathbf{t}}[0], \dots, \hat{\boldsymbol{v}}_{\mathbf{t}}[10])$$

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 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. $\overrightarrow{y_{\text{predict}}} \in \mathbb{R}^5$ represent the models prediction of the input X. Component i of the prediction vector $\overrightarrow{y_{\text{predict}}}$ is indexed by $\overrightarrow{y_{\text{predict}}}[i]$. The standard cross-entropy loss function is defined as the following:

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

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:

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

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

B. Baseline Experiments

We experimented with VGG16, ResNet50 [?], and DenseNet169 as feature extractors for our image-based, baseline models. These are networks that have performed well in prior literature [?], [?], [?], [?] 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 [?] 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 this inherent bias by using the additive attachment as a bias parameter for the prediction layer.

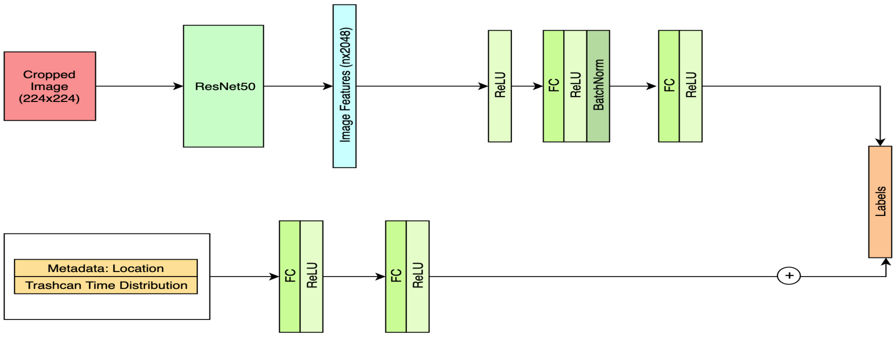

Fig. 5. ThanosNet with additive metadata attachment

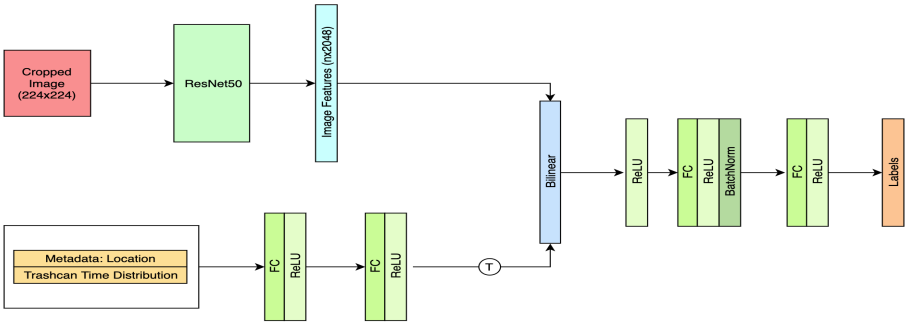

Fig. 6. Thanosnet with bilinear metadata attachement

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

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
BASELINE MODELS AND THANOSNET MODELS RESULTS

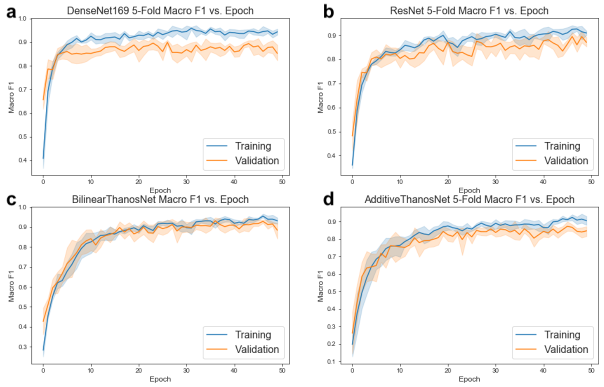

Fig. 7. The macro F_1 in the training (blue) and validation (orange) process for a) DenseNet169, b) ResNet50, c) BilinearThanosNet, and d) AdditiveThanosNet.

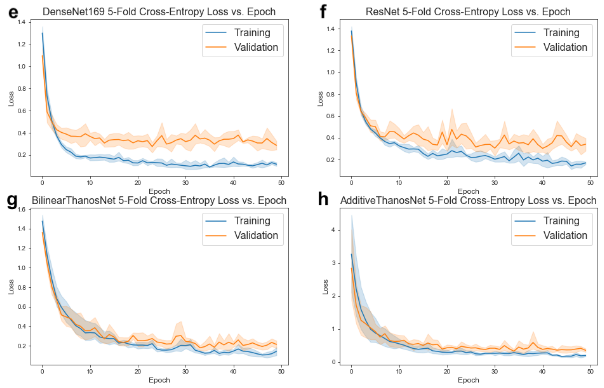

Fig. 8. The loss in the training (blue) and validation (orange) process for e) DenseNet169, f) ResNet50, g) BilinearThanosNet, and h) AdditiveThanosNet.

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.

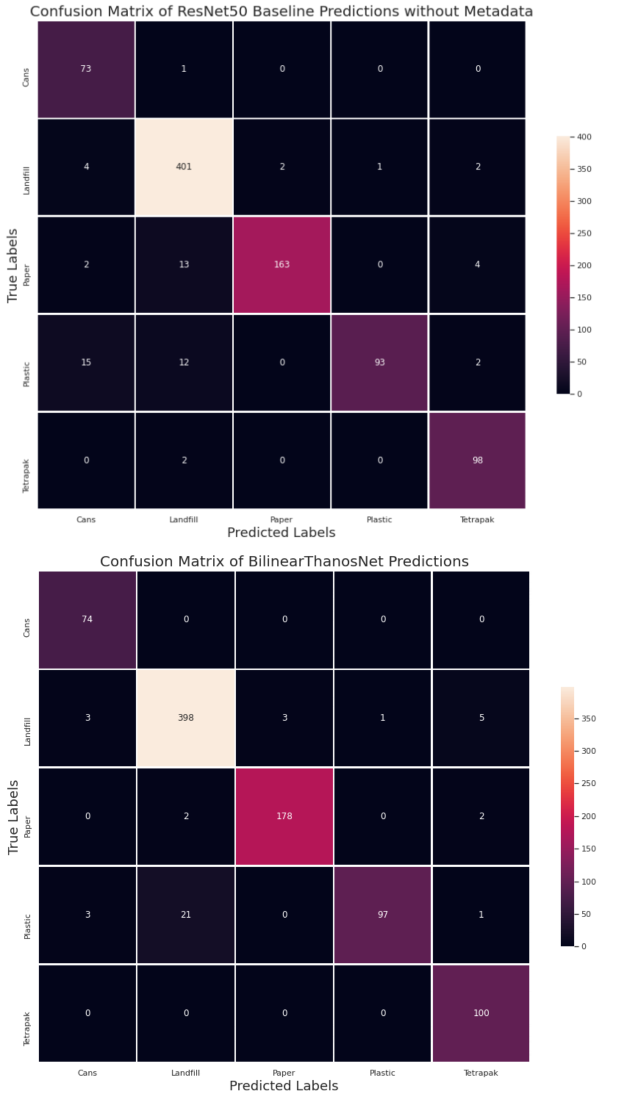

Fig. 9. Confusion matrix of ResNet50 model without incorporating metadata and BilinearThanosNet respectively, generated by combining validation predictions from each model in each fold.

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.