ThanosNet: A Novel Trash Classification Method Using Metadata

Alan Sun
University of Maryland, College Park
asun17904@gmail.com

Harry Xiao
Columbia University
hx2310@columbia.edu

***Abstract*— Recent progress of deep neural networks has warranted significant development of literature for image-based trash classification. These methods predominately use transfer learning to achieve state-of-the-art results. In this contribution, a new methodology is introduced that uses metadata fields such as location and time-based traffic intensity to assist existing image-based classifiers. We curated ISBNet, a dataset which contains 889 images and their associated metadata, distributed over 5 classes (paper, plastic, cans, tetra pak, and landfill). This dataset was used to develop ThanosNet. We found that ThanosNet is superior to current state-of-the-art image-based, trash classification models. Although ISBNet is localized to one user community, the general methodology developed here applicable to a wide array of consumer contexts.**

***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 fine tune well-known CNN-based models. 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, 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]" [8]. 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 bin that the picture originated from. This exposes the location of the trash bin and 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 metadata 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. [9] to utilize transfer learning techniques with deep CNN models pre-trained on the ImageNet dataset [10]. 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.

Previous waste classification systems that incorporate CNNs [7], [9], [13], [17], [19], [20] 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, other useful information can and should be used to facilitate the classification process. 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.

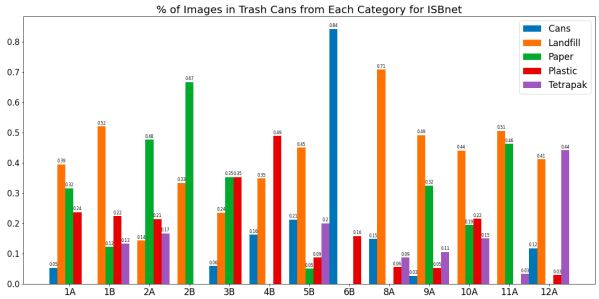

Fig. 1. Distribution of images belonging to each trash class within the individual trash bins. Values are denoted in percentages calculated as the percentage of total photos in the trash bin represented by the class.

In a response to these limitations, our model utilizes metadata that is associated with the physical trash bin including location and time of day. 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 synergies 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 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

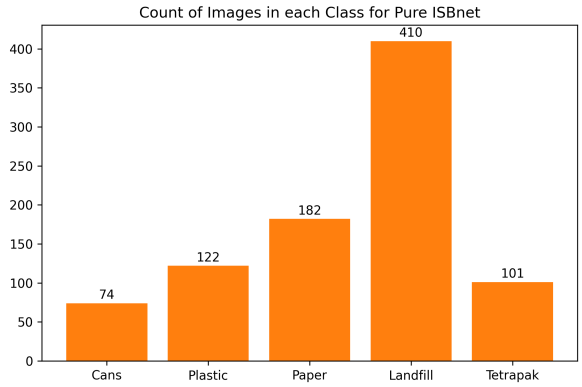

Fig. 2. Class Distribution of ISBNet

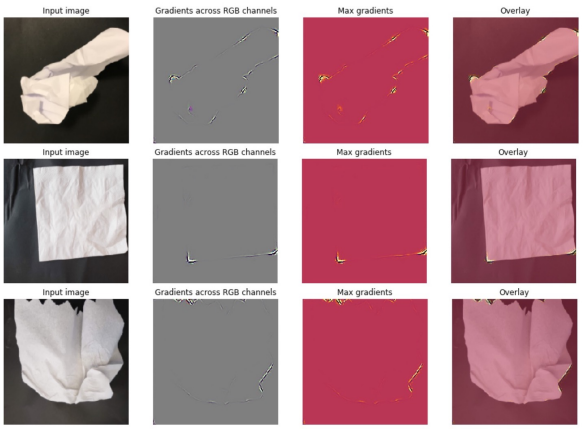

Fig. 3. Saliency maps from a baseline ResNet50 classifier. The images from top to bottom: crumpled piece of paper (paper), a napkin (landfill), cumpled piece of tissue (landfill).

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
- Printer
- Couch Area
- Library

A trash bin's *radius of proximity* is defined as half of the average distance to 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

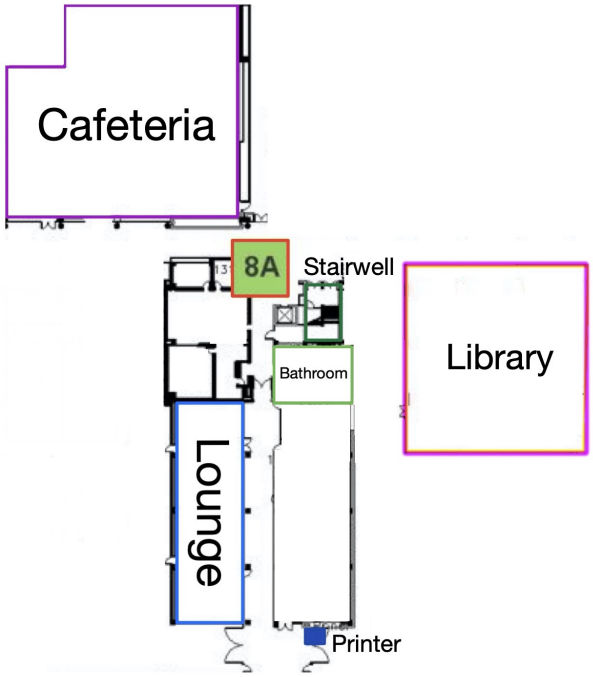

Fig. 4. Trash bin "8A" with a cafeteria, stairwell, bathroom, printer, lounge, and library as proximal landmarks.

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, $\vec{v}_{\text{prox}} \in \{0,1\}^{11}$, that represents its proximal landmarks. Landmarks that have exceeded or are within the bin's radius of proximity are represented by 0 and 1, respectively. Another vector, $\vec{v}_{\text{d}} \in \mathbb{R}^{11}$, describes the diagonal distance between the centers 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:

$$\vec{v}_{\text{x}} = \vec{v}_{\text{d}} \odot \vec{v}_{\text{prox}}.$$

The non-zero entries of this new vector, $\vec{v}_{\text{x}}$, represent the meter distances between the trash bin and its proximal landmarks. The vector $\vec{v}_{\text{x}}$ is normalized into a unit vector using the L^2 norm. This converts the absolute meter distances into relative distances. This unit vector of $\vec{v}_{\text{x}}$ is expressed as $\hat{\boldsymbol{v}}_{\mathbf{x}}$. The non-zero relative distances are negatively correlated, in the sense that shorter distances relate to larger significance, while longer distances related to smaller significance. 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 constant β that is sufficiently large to capture this behavior. This β value replaces the zero-values in the unit vector $\hat{\boldsymbol{v}}_{\mathbf{x}}$. In the sequel, $\hat{\boldsymbol{v}}_{\mathbf{x}}$ denotes the β-replaced unit vector.

A component-wise negative logarithmic transformation, using the natural logarithm, was applied to the unit vector $\hat{\boldsymbol{v}}_{\mathbf{x}}$. The transformed vector is denoted by $\hat{\boldsymbol{v}}_{\mathbf{c}}$. 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 as the distance between the landmark and the trash bin increases. This final location vector, which is denoted by $\hat{\boldsymbol{v}}_{\mathbf{c}}$ in latter sections, is concatenated with a time vector representing the time metadata, which is described in the next section.

B. Time-Based Traffic Intensity

At different times of day, a landmark experiences different levels of traffic intensity. This relative traffic intensity is modeled with a sampled probability distribution. We categorize all foot traffic distributions into three categories: multi-model, normal, and uniform.

Landmarks that characterize 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 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 that characterize 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 composite 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) $f_{\text{m}}(x)$ is defined as

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

where $f_i(\cdot)$ is the truncated PDF of foot traffic for the ith proximal landmark. An average is used to consolidate the time-based traffic intensity distributions of the bin's proximal landmarks into one composite distribution. This consolidation may appear to reduce the dimensionality of the input; however, it does not. Simply, in all practical scenarios, trash bins outnumber landmarks. Each trash bin has one composite time-based traffic intensity distribution. Collectively, over all bins, the dimensionality of the input probability distribution is preserved. For the same argument, equal weighting rather than relative traffic intensity weighting suffices. Furthermore, this composite PDF, which models the bin's time-based traffic intensity, encompasses multiple peaks, such that each peak corresponds to an influx of traffic from one or more of the bin's proximal landmarks. The average reduces noise for less significant time intervals.

This PDF, modeling the trash-generating distribution for bin b_{m}, is discretized by sampling over one-hour intervals 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 component in $\vec{v}_{\text{t}}$, and so forth.

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

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

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

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

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

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

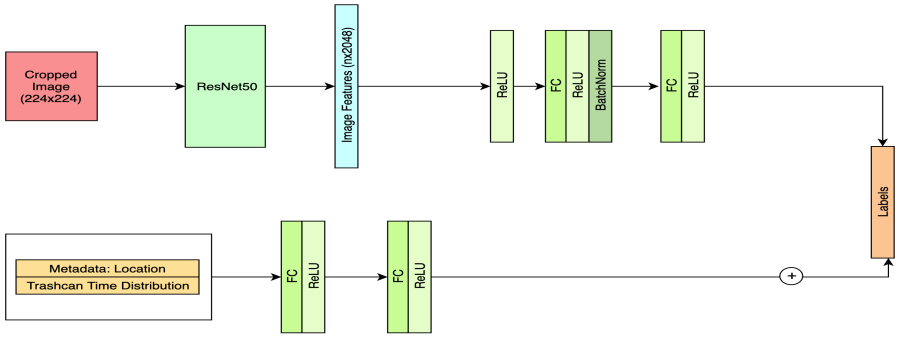

Fig. 5. ThanosNet with additive metadata attachment

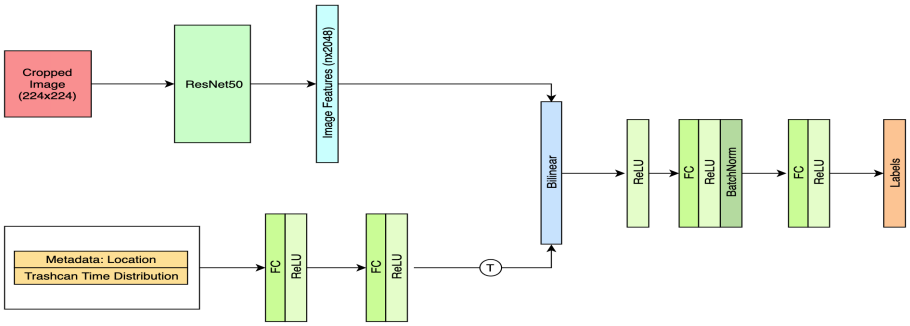

Fig. 6. Thanosnet with bilinear metadata attachement

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 $\vec{y}_{\text{predict}}$ 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

$$\omega_i = \frac{\sum_j \|j\|}{\|i\|},$$

where the summation is over all possible j. $\|j\|$ 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 [13], [17], [19], [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 $\times$ 256, then center-cropped to 224 $\times$ 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 $\sigma_r^2 = 0.229, \sigma_g^2 = 0.224, \sigma_b^2 = 0.225$.

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 traffic intensity 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, its extracted features are correlated to the context provided by location and traffic intensity metadata. 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.

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. We believe that the improvements of BilinearThanosNet come from correlating metadata features with extracted image features. Further, placing this bilinear layer directly after the feature extractor increases the distance between the feature layer and prediction layer. This gives the network more opportunity to correct errors from the bilinear layer and reduce the effect of noise from the data.

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

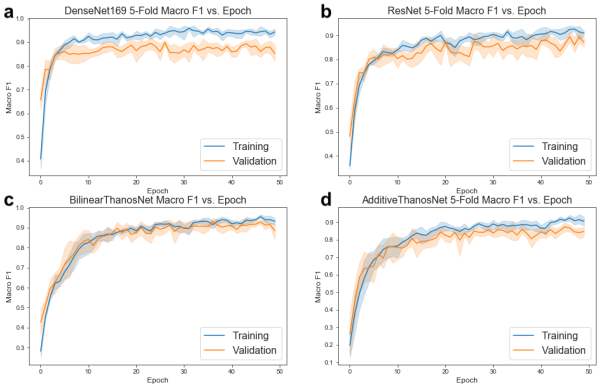

Fig. 7. The macro F_1 in the training (blue) and validation (orange) process for a) DenseNet169, b) ResNet50, c) BilinearThanosNet, and d) AdditiveThanosNet.

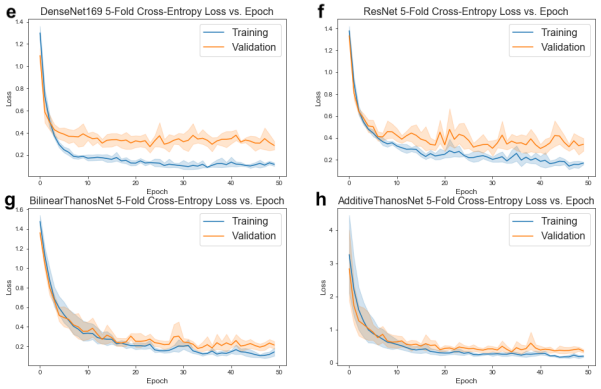

Fig. 8. The loss in the training (blue) and validation (orange) process for e) DenseNet169, f) ResNet50, g) BilinearThanosNet, and h) AdditiveThanosNet.

As shown in Figures 7 and 8, BilinearThanosNet 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

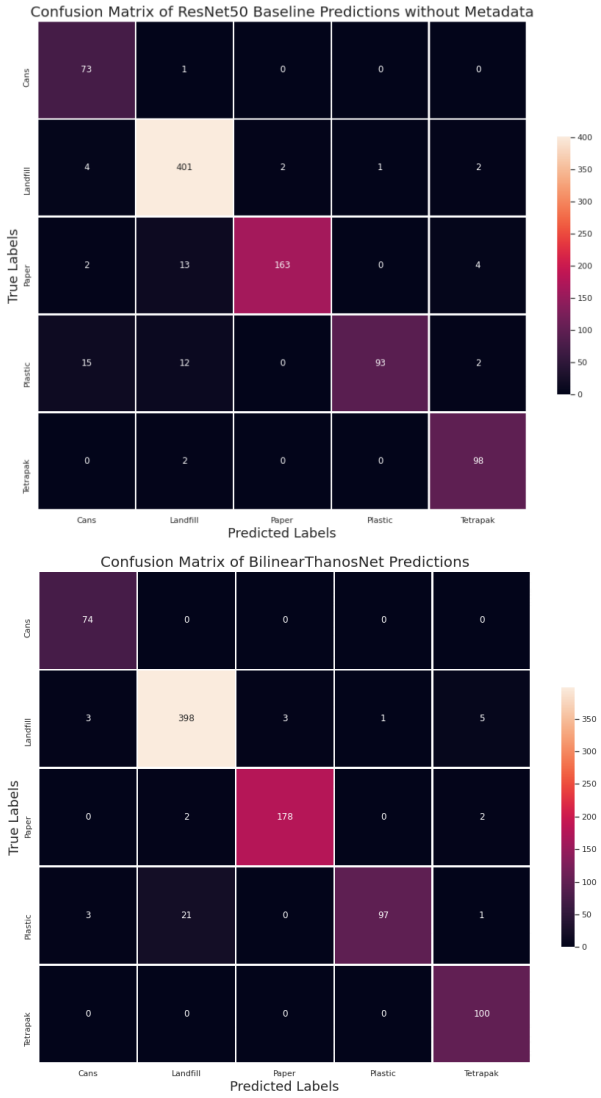

Fig. 9. Confusion matrix of ResNet50 model without incorporating metadata and BilinearThanosNet respectively, generated by combining validation predictions from each model in each fold.

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 and Future Work

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. Through our experiments, we demonstrated how incorporating metadata can result in significant improvements in the classification ability of a model, as ThanosNet outperformed other purely image-based models with an F_1 score of 0.952.

The methodology derived in this study is implementation-friendly, as all the metadata fields can be collected through sensors and simple inspection. Meanwhile, the same idea of a metadata-based trash classification system can also be applied at the industrial level. Recycling plants could geo-tag incoming trash based on the routes of their respective garbage trucks with landmark labels such as: hospital, residential area, office park, etc. These metadata fields would further improve already-implemented, image-based trash sorting systems.

Varying trash bin layouts will result in different overall performances. Since ThanosNet is exposed to location and traffic intensity data, the model has the ability to quantify that difference. Thus, ThanosNet has the ability to optimize trash bin topology. Such a system or application could be the subject of further experimentation and research.

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 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. DOI: 10.21744/ijls.v1i1.24.

[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, ISSN: 23525398. DOI: 10.2991/agc-18.2019.47.

[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. DOI: 10.5281/zenodo.1339604.

[4] Y. Luo, I. Zelenika, and J. Zhao, "Providing immediate feedback improves recycling and composting accuracy," *Journal of Environmental Management*, pp. 445–454, 2019. DOI: 10.1016/j.jenvman.2018.11.061.

[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)*, 2018, pp. 378–383.

[6] J. Donovan, *Auto-trash sorts garbage automatically at the techcrunch disrupt hackathon*, 2016. [Online]. Available: https://techcrunch.com/2016/09/13/auto-trash-sorts-garbage-automatically-at-the-techcrunch-disrupt-hackathon/.

[7] M. Yang and G. Thung, "Classification of trash for recyclability status," pp. 1–6, 2016. DOI: 10.1145/2971648.2971731. [Online]. Available: http://cs229.stanford.edu/proj2016/report/ThungYang-ClassificationOfTrashForRecyclabilityStatus-report.pdf.

[8] M. Ming, *China focus: Beijing to implement citywide mandatory household garbage sorting*, 2019. [Online]. Available: http://www.xinhuanet.com/english/2019-11/27/c_138587683.htm.

[9] J. Knowles, S. Kennedy, and T. Kennedy, *Oscarnet: Using transfer learning to classify disposable waste*, 2016. [Online]. Available: https://cs230.stanford.edu/projects_spring_2018/reports/8290808.pdf.

[10] 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*, 2009, pp. 248–255.

[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*.

[13] R. A. Aral, S. R. Keskin, M. Kaya, and M. Haciömeroğlu, "Classification of trashnet dataset based on deep learning models," *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, pp. 2058–2062, 2019. DOI: 10.1109/BigData.2018.8622212.

[14] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, *Densely connected convolutional networks*, 2016. arXiv: 1608.06993.

[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. arXiv: 1602.07261.

[16] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *CoRR*, vol. abs/1610.02357, 2016. arXiv: 1610.02357.

[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. 178 631–178 639, 2019, ISSN: 21693536. DOI: 10.1109/ACCESS.2019.2959033.

[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. arXiv: 1611.05431.

[19] G. White, C. Cabrera, A. Palade, F. Li, and S. Clarke, "Wastenet: Waste classification at the edge for smart bins," 2020. arXiv: 2006.05873.

[20] Z. Yang and D. Li, "Wasnet: A neural network-based garbage collection management system," *IEEE Access*, vol. 8, pp. 103 984–103 993, 2020, ISSN: 21693536. DOI: 10.1109/ACCESS.2020.2999678.

[21] A. Z. Karen Simonyan Andrea Vedaldi, *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*, 2013. arXiv: 1312.6034.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. arXiv: 1512.03385.

[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)*, Dec. 2015.