Evaluation of Data Augmentation in Waste Management

G002 (s2704516, s2713719, s2742860)

Abstract

Effective waste classification is crucial for recycling, energy recovery, and environmental sustainability but is hindered by limited labeled data. Data Augmentation (DA) can improve model performance, yet existing research often lacks transparency and systematic evaluation. This study assesses Image Manipulation, Image Erasing, and Diffusion Models using DenseNet-121, analyzing their impact on performance, generalization, image fidelity, and computational costs. We propose a structured framework for hyperparameter tuning for DA, with the current SOTA, Diffusion Models, achieving a 30% reduction in the generalization gap and 24% performance improvement on the test set and significantly outperforming simpler techniques. We also offer insights into selecting appropriate DA techniques while considering their inherent trade-offs. Our findings extend beyond waste classification, providing broader implications for DA's effectiveness.

1. Introduction

Waste management failures threaten the environment and society, causing pollution, ecosystem damage, health risks, and rising emissions. Global waste is projected to hit 3.40 billion tonnes by 2050, and reportedly, 33% of waste is currently being mismanaged (World Bank, 2025). The uncontrolled increase in solid waste poses serious risks to human and animal health, pollutes air and water, degrades land, and leads to illegal dumping in non-designated areas, including residential spaces (Abdu & Noor, 2022). In response, waste management has received increased attention (Abdu & Noor, 2022), as it plays a key role in garbage segregation—enabling efficient recycling and using certain waste materials for energy production (Williams, 2013). However, most on-site trash classification methods rely on human expertise, limiting accuracy due to subjectivity, scalability challenges, and labor-intensive processes (Single et al., 2023).

With advancements in Computer Vision (CV), enabling computers to analyze visual data (Voulodimos et al., 2018), these models have been increasingly adopted for waste classification, offering a more scalable and efficient alternative to manual sorting (Abdu & Noor, 2022). These models are often integrated into larger waste-sorting facilities for automatic sorting. Many different architectures have been studied in the literature for waste classification through comparative studies, such as DenseNets (Aral et al., 2018), EfficientNets (Fan et al., 2023), VGG architectures (Kumsetty et al., 2023), Inception models (Patrizi et al., 2021; Aral et al., 2018), and Visual Transformers (Dosovitskiy, 2020). The authors of (Majchrowska et al., 2022b), though, argue that variations in datasets and preprocessing techniques, model architectures, evaluation metrics, and data splitting strategies across studies hinder fair comparisons of methodologies and results in the field, ultimately limiting effective knowledge sharing between academia and industry. Additionally, progress in this field is limited by several other ongoing challenges.

1.1. Key challenges in the field

A key challenge highlighted in the literature is the shortage of sufficient and diverse datasets that accurately represent real-world conditions for waste classification (Abdu & Noor, 2022; Majchrowska et al., 2022b; Aral et al., 2018). To overcome this limitation, most researchers have used Data Augmentation (DA) techniques to expand datasets and enhance diversity by generating new images through transformations (Garcea et al., 2023). These techniques include basic transformations such as rotations, coloring, and shearing (Aral et al., 2018; Fan et al., 2023; Kumsetty et al., 2023; Adedeji & Wang, 2019), erasing techniques (Yang & Li, 2020; Nnamoko et al., 2022; Viegas, 2024), background replacement techniques (Patrizi et al., 2021), advanced generative models like Generative Adversarial Networks (GANs) (Alsabei et al., 2021) and their extensions Deep GANS (Fan et al., 2023).

However, while most studies report applying DA for waste classification, therefore acknowledging its effectiveness in improving performance, they often lack specific details on its implementation. Key aspects such as the specific augmentations implemented, their assigned probabilities, the possibility of applying multiple augmentations to a single image simultaneously ("stacking"), and whether these augmentations are fixed or vary dynamically across different batches and epochs are often unspecified. Additionally, most studies do not report the exact number of generated images, making it difficult to evaluate the scale of augmentation, the required computational resources, and the direct impact on classification performance. Without this information, comparing augmentation strategies or replicating results remains challenging for future research in the field. Another key insight is that most studies rely on a single dataset and do not assess whether their insights generally aid generalization to different datasets, which is essential for real-world deployment in the field.

1.2. Data Augmentation

Data augmentation techniques reportedly enhance image classification by improving model robustness and generalization while reducing overfitting through promoting invariant representations (Kumar et al., 2024; Garcea et al., 2023; Nanni et al., 2021). They are widely utilized to enhance performance by addressing challenges related to data scarcity, imbalance, and distribution shifts across various applications in image classification tasks, such as in medicine (Chlap et al., 2021), product defect detection (Rožanec et al., 2023), and wildlife species recognition (Tabak et al., 2019). Additionally, the emergence and development of advanced DA techniques, such as Diffusion Models, have shown improved image generation in various other tasks (Croitoru et al., 2023). The authors of (Dhariwal & Nichol, 2021) demonstrated that diffusion models outperform state-of-the-art models like GANs (Goodfellow et al., 2014) in image synthesis, establishing them as the current leading approach in the field.

As stated above, there has been limited research on comparing and identifying the most effective augmentation techniques for waste classification, and to the authors' best knowledge, when comparisons are made, as in (Yu & Grammenos, 2021), they typically involve only basic methods. Given that the effectiveness of augmentation techniques varies based on input type, visual task, and application (Garcea et al., 2023; Kumar et al., 2024; Nanni et al., 2021), making them highly domain-specific, it is *crucial* to explore which methods are best suited for the task of waste classification. Moreover, as data augmentation (DA) techniques are widely used in this field, it's crucial to understand their effects comprehensively. This involves adopting recent innovations like Diffusion Models and lesser-used techniques for this task, such as Advanced Image-Erasing Methods (Cutout, Hide-and-Seek, GridMask). Thoroughly evaluating these methods is crucial, as they will be systematically compared across key metrics, including classification performance, computational efficiency, generalization on an unseen different dataset, and generated image fidelity—an aspect that is novel in our task and adds a new dimension to waste classification research. The framework presented in this study can also be generalized for other tasks within image classification.

In this study, following one of the most recent taxonomies proposed by (Kumar et al., 2024), we focus on three basic image data augmentation (DA) techniques: Image Manipulation, which includes (non-)geometric transformations; and Image Erasing, which involves masking parts of the image, which are both grouped under the "Basic Image DA" category. Additionally, from the "Advanced Image DA" category, we incorporate diffusion-based data augmentation techniques. Recently introduced, Diffusion Models have seen minimal exploration for this task, with (Pachaiyappan et al., 2024) being one of the few examples, specifically in underwater waste detection. Therefore, we aim to further investigate their potential in this domain. We as authors, believe that using all these different DA techniques will enable a holistic comparison between commonly used, and computationally efficient methods ("Basic DA") and more advanced, computationally demanding approaches ("Advanced DA"), to determine the optimal balance between performance, generalization, image fidelity and resource requirements. The DA techniques used in this study are shown in Figure 1.

Therefore, this study aims to contribute to the field in the following ways:

- Evaluating state-of-the-art (SOTA) data augmentation techniques, which have not previously been explored, against standard models in the field of waste classification.
- Addressing the gap in comparative research by systematically analyzing these techniques across various evaluation metrics.
- Provide a structured framework for tuning the Data Augmentation pipelines based on task-specific requirements, including exploratory analysis of key characteristics and dimensionality reduction.
- Assess the generalization gap of these techniques to unknown datasets, further aiding in understanding their robustness and applicability beyond the training distribution.

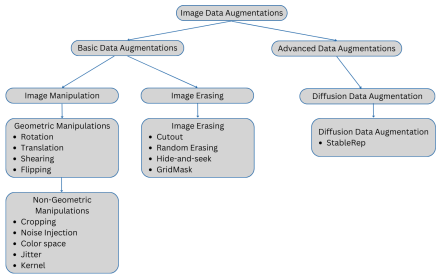

Figure 1. Adapted from (Kumar et al., 2024): Augmentation techniques employed in this study.

2. Data set and task

To address our research questions, we selected two publicly available datasets: RealWaste, introduced in (Single et al., 2023), for training, and the widely used TrashNet (Yang & Thung, 2016) as a test set to assess the generalization of our augmentations. This selection was motivated by several factors: both datasets share the six main waste categories, have similar data collection methods, are designed specifically for image classification tasks, and illustrate centered images in front of white backgrounds, ensuring consistency across samples. Examples of objects from the datasets are shown in Figure 2 and Appendix 6. Another key factor in our choice is the demonstrated compatibility of both datasets with similar CNN architectures, such as DenseNets121, which demonstrated notable performance in (Aral et al., 2018) and (Single et al., 2023). With that said, the datasets also differ in several aspects, including lighting conditions, object variations within categories (e.g. The "cardboard"

category in the RealWaste dataset includes objects made of cardboard like containers, shoe-boxes whereas in TrashNet, this category contains only actual packaging cardboard), color schemes, notably the presence of dirt, and noise, as can be seen in Appendix 6 and the quantitative analysis in 6. These differences allow us to assess whether our model has truly learned generalizable features rather than overfitting to dataset-specific patterns. RealWaste was chosen for training due to its larger size (3587 images) and since being relatively novel has received less attention in the research community. It was further subdivided into an 80% training (2868) and 20% validation set (719) for hyperparameter tuning, using stratified sampling to preserve class distribution. TrashNet was chosen as the test dataset due to its status as a well-established benchmark in the field (Abdu & Noor, 2022; Majchrowska et al., 2022a).

The datasets were further processed by mapping waste labels, removing unmatched categories from RealWaste ("Textile Trash," "Vegetation," and "Food Organics"), and resizing images to 256×256 before converting them into tensors to ensure consistency in input size. All original images were in RGB color space. The RealWaste dataset had an original resolution of 524×524, while TrashNet images were 512×384 before resizing. The datasets descriptions are shown in Table 1 with the corresponding training and validation from RealWaste and Test of TrashNet.

Figure 2. Realwaste vs TrashNet

Table 1. Category-wise distribution of waste images.

Category	RealWaste Training	RealWaste Validation	TrashNet Test
Cardboard	368	93	403
Glass	336	84	501
Metal	632	158	410
Paper	400	100	594
Plastic	736	185	482
Trash	396	99	137
Totals	**2868**	**719**	**2527**

2.1. Task and evaluation metrics

As mentioned earlier, the core task is image classification, where the input consists of the image pixels, and the output is one of the corresponding six class labels. Since the primary goal is to evaluate data augmentation techniques, we first measure the computational time for generating augmentations, followed by the training time, which will be reported as the ratio of $\frac{T_{\text{aug}}}{T_{\text{vanilla}}}$ compared to the non-augmented baseline. This follows prior work, such as (Kumar et al., 2024), which noted the trade-offs between complexity and performance, and (Liang et al., 2023), which analyzed training time ratios. Finally, classification performance is assessed using the F1-score, a standard and appropriate metric given the class imbalance in Table 1. Finally, another novel contribution of this study is the assessment of the similarities between augmented and original images using the Learned Perceptual Image Patch Similarity (LPIPS) metric (Zhang et al., 2018). Chosen for its ability to measure perceptual similarity through VGG-based deep features, LPIPS evaluates distances between feature representations using learned weights that mimic human vision, offering advantages over traditional pixel-based metrics like PSNR and SSIM. Augmented images should maintain close similarity to originals to ensure label consistency (i.e., lower LPIPS scores) while introducing sufficient variation for enhanced diversity and generalization. To the best of our knowledge, this metric has not been employed in the context of waste classification, despite its relevance in ensuring realistic augmented images.

3. Methodology

The key insight of this paper is the systematic evaluation of data augmentation techniques for waste classification across model performance, computational cost, image fidelity, and generalization to unseen data.

3.1. Data Augmentation techniques

The Data Augmentation techniques chosen to explore are *Image Manipulation*, *Image Erasing*, and *Diffusion Models*, following the taxonomy of (Kumar et al., 2024). These were applied using the albumentations library. The Image Manipulation category consists of both geometric and non-geometric transformations. These techniques are often preferred in the literature due to their ease of implementation (Shorten & Khoshgoftaar, 2019) and low computational complexity (Kumar et al., 2024). The geometric transformations include rotations, which rotate the image by a specified angle; translations, which shift the image in horizontal or vertical directions; shearing, which skews the image along an axis; and flipping, which reflects the image horizontally. For the non-geometric transformations, these include cropping, which extracts a specific region of the image; Gaussian noise injection, which adds random noise to simulate real-world variations; color space changes, which alter the image's color representation; jitter, which applies small random distortions to intensities; and kernel filtering, which modifies image sharpness or blurriness. These are the most widely used transformations for waste classification, as seen in, for example, (Aral et al., 2018; Adedeji & Wang, 2019) and many more.

Additionally, we experimented with Image Erasing, a technique that hides part of the image, encouraging the model to learn holistic features instead of specific details (Kumar et al., 2024). This method improves generalization,

especially for waste images, which often feature damage, poor angles, or low lighting (Yang et al., 2023b). This DA technique includes several methods, with the most common being Random Erasing, Cutout, Hide-and-Seek, and GridMask (Kumar et al., 2024). Random Erasing replaces a randomly selected rectangular region with a black box, while Cutout removes a fixed-size subregion, ensuring that a specific portion of the image is consistently removed (DeVries & Taylor, 2017). Hide-and-Seek divides the image into random regions and removes some, masking multiple parts at the same time, in contrast with the two above methods. Finally, GridMask addresses the randomness of these methods by overlaying a transparent grid, blacking out the rest of the image (Kumar et al., 2024). Random Erasing has been previously explored in the field, showing performance gains (Yang & Li, 2020; Nnamoko et al., 2022; Viegas, 2024), but the other three more advanced methods have reportedly been underexplored.

The final DA technique selected was Diffusion Models, the SOTA in image generation, which has gained significant traction in research in recent years (Yang et al., 2023a). These models have demonstrated outstanding performance in generating novel characteristics in images, including structural elements, textures, and perspective changes, surpassing traditional augmentation techniques (Trabucco et al., 2023). Recent studies also indicate that Diffusion Models have outperformed the long-dominant Generative Adversarial Networks (GANs) in image generation tasks (Dhariwal & Nichol, 2021). The core principle of diffusion models involves a two-stage process: first, data is progressively corrupted by adding noise according to a Markov process; then, during the reverse process, the model learns to remove this noise step-by-step to reconstruct the original data and generate new images. This approach allows diffusion models to generate high-quality images by refining noisy inputs over multiple timesteps. Mathematically, the forward process can be expressed with equation 1, where at each timestep t, a small amount of noise is injected, gradually destroying the structure of the original data. The degree to which the original data is retained is controlled by $\bar{\alpha}_t$, a cumulative weighting term that determines how much of x_0 remains at time t. The reverse process, responsible for denoising and reconstructing the data, is parameterized by a deep neural network and can be expressed with equation 2. This step approximates the true posterior distribution of the data by learning a mapping from noisy samples back to clean ones. During training, the noise prediction loss is minimized, ensuring that the model learns to estimate the noise component added at each step. Finally, during inference, new images are generated by starting from pure Gaussian noise and iteratively applying the learned denoising transformation, as formulated in equation 3.

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{1}$$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{2}$$

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right) + \sigma_t\epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{3}$$

In this study, Diffusion Models were selected due to their demonstrated ability to generate high-quality and diverse samples, stemming from their reverse noising process used to refine predictions (Croitoru et al., 2023), which can be highly beneficial for evaluating generalization and model robustness to other datasets (Kumar et al., 2024). Additionally, given their well-documented computational challenges—stemming from the iterative nature of the denoising process—we aimed to further analyze this trade-off. For our study, we used the Stable Diffusion Img2Img pipeline from the Hugging Face Diffusers library (von Platen et al., 2022), enabling image-to-image transformation paired with text prompts. We enhanced waste images using category-specific prompts from the RealWaste dataset. The pipeline also allowed tuning the balance between prompt influence and image realism for controlled augmentation.

Finally, an interesting finding regarding DA techniques is their enhanced performance when applied in combination (Pawara et al., 2017; Yang et al., 2022). To assess this effect and provide a comprehensive research perspective, we developed a pipeline that integrates images from multiple augmentation techniques using random downsampling, allowing us to evaluate their combined impact.

On another note, DA is an effective way to address class imbalance by generating synthetic samples for the minority class (Shorten & Khoshgoftaar, 2019). As shown in (de la Rosa et al., 2022), such techniques can significantly improve performance. To reduce bias from our highly imbalanced dataset, we used DA to generate additional images, expanding the minority class to twice the size of the majority class. Importantly, the original images were retained. So in total, our final dataset consists of 1472 images per class (736 images are in the majority "plastic" class in the *training* set), in total 8832 images. These images were generated before training and remained unchanged across epochs. See Appendix 6 for hyperparameter details and Appendix 6 for sample generated and transformed images.

3.2. Model Architecture

The pre-trained DenseNet-121 (Huang et al., 2017) was utilized, as it performed well on both benchmark datasets (Aral et al., 2018; Single et al., 2023). Its dense connectivity improves feature reuse and gradient flow while keeping parameters low. The architecture includes an initial convolution and pooling layer, followed by four dense blocks (6, 12, 24, and 16 layers), separated by transition layers with 1×1 convolutions and 2×2 pooling. The network uses the ReLU activation function and applies batch normalization throughout. Max pooling is used initially, while average pooling is applied in the transition layers. Only this final layer is trainable, and the rest of the model is frozen to retain pre-trained parameters. For training, we used the

Stochastic Gradient Descent (SGD) optimizer with a learning rate of 0.0001, momentum of 0.9, and cross-entropy loss. All models were trained for 100 epochs.

To summarize our methodology, this paper highlights the importance of task-specific data augmentation (DA)—an aspect often overlooked in prior work. Following (Nanni et al., 2021), we began with a per-class exploratory analysis to guide the selection and tuning of three chosen DA techniques. The generation of the augmented images was done using an iterative pipeline to assess alignment with real data (this will be described in Section 4 and Figure 4). To address class imbalance, we expanded each class to twice the size of the majority class, including both original and augmented images in the training set. We also built a combined pipeline using downsampling to evaluate its performance. All datasets were used to train a DenseNet-121 model for 100 epochs with the same hyperparameters. The final evaluation was conducted on the unseen TrashNet dataset using the F1 score, LPIPS for fidelity, generalization gap, and computation time (augmentation and training). A visual overview of the full methodology is provided in Figure 3.

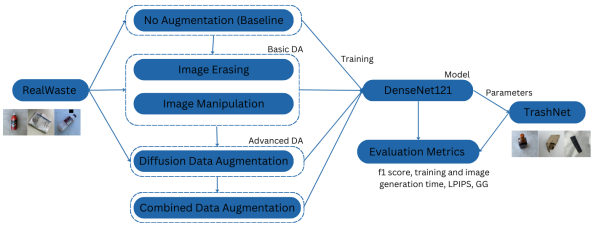

Figure 3. Methodology

4. Experiments

All experiments presented in this study were conducted on an RTX 4070 with 8GB VRAM. As mentioned in Section 3, our main motivation is to systematically evaluate Data Augmentation (DA) in waste classification, a field where its impact remains underexplored. Many studies apply DA heuristically, making it difficult to assess its true benefits. To systematically evaluate the role of Data Augmentation (DA) in waste classification, we test widely used, computationally efficient techniques such as *Image Manipulation* and *Erasing*, belonging int he "Basic DA" category and compare them against state-of-the-art yet underexplored *Diffusion Models* in the "Advanced DA" category, all applied to DenseNet-121. Our goal is to hollistically evaluate them and identify the best trade-offs between classification performance, generalization, image fidelity, and computational cost.

Through our distinct DA pipelines, we aim to:

- Identify if data augmentation is useful for waste classification, and if so, identify the most effective augmentation technique for this task under our experimental setup.
- Evaluate whether diffusion-based methods offer a meaningful performance advantage over traditional augmentations and whether their computational overhead is justified.
- Analyze the trade-offs between generation time, training time, image fidelity, generalization gap and overall performance.
- Examine the trade-off between validation and test performance, particularly when evaluating on an unseen and different dataset, and explore how generalization gaps vary across pipelines.
- Investigate whether combining all augmentation techniques through downsampling yields further benefits, as is commonly done in the literature.

Ultimately, understanding the role of DA is essential for enhancing model robustness to unseen data and building more scalable, reliable waste classification systems.

4.1. Image Generation process

Building on the insights of (Nanni et al., 2021), which emphasize the importance of understanding the data for augmentation, we adopted a quantitative approach to hyperparameter tuning across our augmentation pipelines. This analysis proved critical: we observed a 4–6% performance improvement on the test set when comparing carefully tuned pipelines to those with randomly selected parameters. For example, in the Erasing pipeline, the F1 score improved from 47% (not reported here) to 52%, as shown in Table 2. This was achieved using our *Iterative Pipeline*, which involves three steps: image generation, image evaluation, and iterative refinement based on quantitative metrics. After generating images, we applied t-SNE to compare the distributions of original and augmented training data. Misalignment signaled the need for further tuning, quantified using average KL Divergence across classes—a measure of divergence between augmented and non-augmented distributions. In the example given above with the Erasing pipeline, a high KL score indicated poor alignment in our initial pipeline. Visual inspection showed that excessive erasure removed important features, which we addressed by reducing the grid size to preserve relevant content. We also applied each of the four techniques—CoarseDropout, GridDropout, Hide-and-Seek, and RandomErasing—independently rather than stacking them, which degraded image integrity. This highlights a key insight: incorporating specific metrics for hyperparameter tuning within the pipelines is crucial and can lead to significant performance gains—challenging the notion that data augmentation is merely an "art" or a "heuristic". Importantly, these gains were not due to overfitting: improvements appeared in both validation and test sets, and the t-SNE/KL analysis was conducted on the training set. Similar tuning effects were observed with other hyperparameters. The final pipelines are detailed in Appendix 6, and the code repository includes results from earlier iterations.

Finally, as illustrated in Figure 4, our pipeline follows a

generalizable, iterative loop of generation and evaluation to refine the DA process. The goal is to strike a *delicate balance* between preserving key features and introducing enough diversity to improve generalization. Imbalance can lead to overfitting —either to artifacts introduced by poor augmentations or to particularities of the training set itself if the augmented data does not introduce sufficient variation. Through repeated cycles of refinement, the pipeline progressively refines the augmentation hyperparameters and produces more effective and adaptable datasets. Metrics and thresholds should be selected per task to avoid overfitting, and future research may help define what constitutes "optimal" thresholds and how metric choice affects generalization and performance.

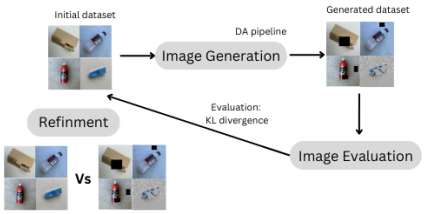

Figure 4. Iterative pipeline for Data Augmentation

4.2. Key Results

A preliminary inspection of Table 2 and Table 3 shows that the two "Basic DA" methods—Image Manipulation and Image Erasing—have comparable image creation and training times, demonstrating high computational efficiency compared to Diffusion Models, which require *significantly* more time for image creation. Additionally, these basic methods effectively reduced the generalization gap from 39% (without DA) to 19% and 15%, respectively, highlighting their impact on improving model generalization while still improving results on the test set. The LPIPS scores of 0.59 and 0.51 indicate that images generated through Image Manipulation were perceptually more different from the original training set than those generated through Image Erasing. Since lower LPIPS values reflect greater perceptual similarity, Image Erasing demonstrates better alignment with the original data. These LPIPS scores also correlate with performance on the validation set: Image Erasing outperformed Image Manipulation. However, the increased diversity introduced by both pipelines contributed positively to generalization, making these differences beneficial and further supporting the effectiveness of the augmentations. Between them, they had comparable performance on the test set, scoring 53% and 52%, respectively, but with a notable difference on the validation set, 72% and 77% (perhaps due to the LPIPS score), suggesting that Image Erasing worked better overall.

On the other hand, Diffusion Models—classified under the "Advanced DA" category—were significantly more computationally expensive in terms of image generation time. However, since all pipelines produced an equal number of images, training times remained comparable across methods. Diffusion models successfully reduced the generalization gap by 30% compared to the baseline non-augmented model, to 9%, while achieving strong performance on the validation set and the strongest performance on the test set of 66%, an impressive 24% increase to the baseline, with just a 6% decrease in performance on the validation set, thus making them by far the best-performing model overall. This aligns with current research as they are the current SOTA. They also produced diverse images that deviated from the originals, as reflected by the LPIPS score—though less diverse than Image Manipulation, which is key to generalization. However, a key question remains: Is this significant performance gain worth the added computational complexity? According to various studies, this overhead must be justified by corresponding performance improvements (Ulhaq & Akhtar, 2022; Kumar et al., 2024), which, in this case, we as authors believe is justified. Even within our limited setup, they achieved a 24% improvement on the test set compared to the second-best, more computationally efficient method, which reached 11%. We believe that with additional resources to generate more images, this performance gap could increase even further.

Table 2. Weighted F1 scores in the RealWaste [Val] and TrashNet [Test], trained on Densenet-121 (DN)

DA technique	**RealWaste**	**TrashNet**
DenseNet-121 (DN)	81%	42%
DN+Image Manipulation	72%	53%
DN+Image Erasing	77%	52%
DN+Diffusion Models	75%	66%
DN+Combined DA	75%	63%

Table 3. Measures per technique

DA Technique	**Creation***	**Train***	**GG***	**LPIPS**
Image Manipulation	~ minutes	2.3	19%	0.61
Image Erasing	~ minutes	2.5	15%	0.51
Diffusion Models	~ days	2.6	9%	0.55
Combined DA	-	2.3	13%	0.57

Note: *Creation* is the time taken to generate augmented images. *Train* is the training time as a multiple of the baseline model (no augmentation). *GG* is the Generalization Gap between validation and test sets. *LPIPS* measures the distributional difference of the augmented data for image fidelity.

Finally, the combined data augmentation (DA) pipeline performed well on the validation set, matching the performance of the diffusion models. However, it achieved a lower test set score of 63% and a relatively high LPIPS score of 0.60, suggesting that simply combining multiple augmentation methods to increase diversity does not necessarily lead to optimal performance. In terms of computation, training time was comparable across methods, and since we used downsampling, image creation time was negligible. This highlights the need for a thoughtful integration strategy, as certain techniques are more effective at capturing the underlying structure given a task and a dataset. For the LPIPS scores: while higher LPIPS scores were sometimes linked to improved test performance—particularly for Diffusion Models and the combined DA pipeline—this trend was inconsistent. For instance, Image Manipulation, with a higher LPIPS score of 0.61, underperformed on the test set, possibly because the augmentations deviated "too much" and did not produce realistic images from the original

data. On the validation set, lower LPIPS scores—indicating better alignment—generally corresponded to stronger performance.

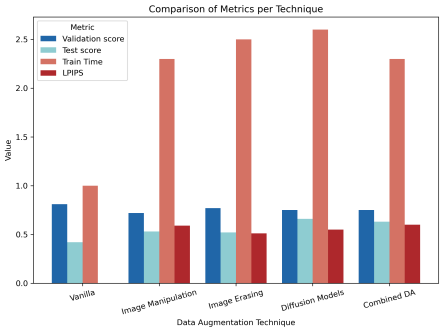

Figure 5. Results of pipelines

4.3. Results per-category

The per-category performance across validation and test set is shown in Tables 4 and 5. Firstly, Table 4 shows that data augmentation (DA) techniques generally led to a reduction in per-category performance in the RealWaste validation dataset, with Image Manipulation causing the most significant drop, followed by Image Erasing and Diffusion Models. Plastic had the most reduction in performance consistently across the DA techniques, with an average drop of approximately 10%, followed by the category of Trash. Trash, in particular, had a drop of 11% with Image Manipulation and 14% for Diffusion models, likely due to its high intrinsic variability and the difficulty of generating meaningful transformations that preserve class-specific features. However, in the combined pipeline, the performance drop was smaller for Trash, suggesting that for such complex categories in our given limited setup, a combination of augmentation techniques may be necessary to better capture the diversity while retaining discriminative signals. Cardboard, Metal, and Paper showed more moderate loss in performance, consistently across the DA pipelines. Some classes, for example, Glass, remained largely unaffected by certain augmentations, such as, for example Diffusion models achieving 85%, nearly matching the baseline, indicating strong class separability even with transformations. It is worth noting that Image Erasing concluded in more stable declines, as it also is the best-performing validation score overall, closely followed by Diffusion models. The combined pipeline also performed strongly, often acting as an average of the individual pipelines in categories like Metal and Cardboard, while in some cases matching the best performance of any single method.

Table 4. F1 scores in Realwaste per category, trained on Densenet-121 (DN)

DA technique	Cardboard	Glass	Metal
DenseNet-121 (DN)	79%	86%	83%
DN+Image Manipul.	73%	75%	77%
DN+Image Erasing	77%	78%	81%
DN+Diffusion Models	74%	85%	79%
DN+Combined DA	75%	83%	78%

DA technique	Paper	Plastic	Trash
DenseNet-121 (DN)	79%	82%	80%
DN+Image Manipulation	74%	71%	69%
DN+Image Erasing	78%	73%	81%
DN+Diffusion Models	73%	71%	66%
DN+Combined DA	78%	72%	73%

Table 5 shows the results on the test set, where in all but one case, DA significantly improved F1 scores per class, with some categories benefiting more than others. Cardboard, previously identified as a key differentiating class in Section 2, showed significant performance gains across all pipelines, particularly with Diffusion Models, from 45% to 69%. This technique may have leveraged prompt variations to generate more generalizable images. Similarly, Paper, Glass, and Metal exhibited notable improvements, especially with Image Manipulation and Diffusion Models highlighting the effectiveness of these techniques for these categories. Notably, the Trash category, the hardest to label with a 3% score in the baseline, showed a three-fold improvement with Diffusion Models and Erasing, increasing to 9% and 11% with the combined pipeline. However, despite this relative gain, the low overall performance underscores the difficulty of distinguishing Trash from other categories, likely due to its high variability and lack of consistent class-specific features, which is perhaps why the combined pipeline managed to work better in this case. When comparing the pipelines, Image Manipulation improves most categories significantly (e.g., 15+% gains for Cardboard, Paper, and Glass) but is not helpful in Plastic (10% decrease). Image Erasing is always effective, with consistent but, in general, lower per-class gains. Diffusion Models, on the other hand, the current SOTA show significant gains (e.g. 24% gains in Cardboard, 15% Glass etc) and also when it comes to difficult to distinguish categories such as Trash. The combined pipeline also showed strong but not always the best performance, reflecting the uncertainty around which augmentations were applied and which are most effective for each category. In many cases, it performed close to the average of the individual pipelines, while in some classes, it matched or even exceeded the best-performing method. However, its inconsistent behavior highlights the need for a more controlled approach to combining augmentations.

Table 5. F1 scores in TrashNet per category, trained on Densenet-121 (DN)

DA technique	Cardboard	Glass	Metal
DenseNet-121 (DN)	45%	50%	60%
DN+Image Manipulation	60%	65%	70%
DN+Image Erasing	52%	62%	66%
DN+Diffusion Models	69%	65%	71%
DN+Combined DA	63%	65%	70%

DA technique	Paper	Plastic	Trash
DenseNet-121 (DN)	55%	40%	3%
DN+Image Manipulation	72%	30%	7%
DN+Image Erasing	61%	48%	9%
DN+Diffusion Models	75%	56%	9%
DN+Combined DA	72%	50%	11%

4.4. Discussion

Data augmentation (DA) clearly improved generalization across all pipelines, but not all techniques were equally effective. Diffusion Models achieved the strongest results, reflecting their ability to generate diverse, high-quality images, albeit at a significant computational cost. This confirmed their value as the most effective augmentation strategy for this task. The combined pipeline revealed an interesting insight: even limited use of advanced augmentations, when strategically mixed with simpler, more affordable techniques, can be highly effective—a commonly underexplored area. For instance, generating only a small number of images using an advanced technique and supplementing the rest with a cheaper technique, like Image Erasing. This particularly helped the category "Trash", that was the hardest to label, due to its inherent variability. However, this blend remains a black box, with unclear contributions from individual methods. In contrast, basic techniques like Image Manipulation and Erasing were far more efficient and still delivered solid performance—especially Image Erasing, which provided consistent generalization gains

across categories despite modest decrease on validation. Image Manipulation also performed well overall, but with more variation per category. We also observed a clear tradeoff between validation and test performance. In many cases, validation scores dropped slightly, while test performance improved significantly, especially for difficult classes like Trash. This suggests that higher image diversity can boost generalization to unseen domains even if in-distribution performance suffers slightly. LPIPS scores also partially suggest a trade-off: moderate diversity (as in Diffusion Models and combined pipeline) improved test performance, but excessive deviation from the original images, such as in the case of Image Manipulation can hurt results. In general also, lower LPIPS values aligned with better validation performance, highlighting the importance of perceptual realism. Across techniques, we found no signs of overfitting, supporting the idea that augmentations enhance robustness rather than simply "memorizing". Overall, our findings reinforce that DA is beneficial, but effective use requires balancing generation time, fidelity to the original distribution, and class-specific effects, especially in settings with limited data and class variability like waste classification.

5. Related work

Several studies have explored waste image classification using neural networks (Aral et al., 2018; Single et al., 2023; Vo et al., 2019), often applying data augmentation to address limited data. However, most do not assess its impact rigorously—they lack comparisons to non-augmented baselines, omit parameter details, and fail to evaluate different augmentation types. Since insufficient data remains one of the most common challenges in the waste classification task (Abdu & Noor, 2022), it is worth understanding the impact of DA. To the best of the author's knowledge, this is one of the first studies to quantitatively evaluate the impact of various data augmentation techniques holistically in waste classification. It provides insights into the computational complexity and performance gaps previously identified in the literature. Overall, our results demonstrate that DA is an effective approach for improving generalization, consistent with (Zhou et al., 2022; Kumar et al., 2024) on unseen datasets with variations, significantly reducing the generalization gap and increasing performance on the test set. The best-performing approach—Diffusion Models—achieved the highest results, albeit at a significant computational cost.

To our knowledge, some research has been conducted on identifying effective data augmentation strategies across various domains, including medicine (Safdar et al., 2020; Goceri, 2023). However, most of the studies tend to focus either on basic augmentation techniques (Safdar et al., 2020; Yu & Grammenos, 2021) or on direct comparisons between state-of-the-art methods, such as diffusion models and GANs (Dhariwal & Nichol, 2021). Even in cases where comparative studies between basic and advanced methods do exist—such as (Nanni et al., 2021)—they are typically limited to reporting performance metrics, lacking a more holistic evaluation across dimensions like generalization gap, training efficiency, or image fidelity. We, therefore, believe that our comprehensive analysis—incorporating a range of evaluation metrics—offers valuable insights that could be extended to other fields, helping to demystify data augmentation and move beyond treating it as a black-box technique.

In the context of existing literature, one limitation of this study is computational power. For instance, (Single et al., 2023) reported 89% accuracy on RealWaste, likely due to training for over 180 epochs and generating a larger volume of augmented images. Despite these differences, our results remain consistent with their findings. Likewise, higher scores on TrashNet—such as those in (Aral et al., 2018)—were achieved using training-validation splits rather than true generalization testing, along with significantly longer training and more extensive augmentation.

5.1. Future Work and Limitations

This study had limitations, notably fewer computational resources compared to others like (Single et al., 2023), which enabled larger models, longer training, and more extensive augmentation. Nevertheless, our methodology allowed for a thorough analysis of each method's strengths and weaknesses. A promising future direction is inspired by (Xue et al., 2021), where a selective mechanism evaluates the realism of *each* generated image—rather than the entire image set as we applied —based on entropy and feature similarity before deciding whether to include it in the dataset. Such filtering could boost performance, especially if paired with threshold tuning to balance image fidelity and generalization. Although this approach was considered, time constraints prevented implementation.A promising direction is to tune augmentation hyperparameters per class, as methods like Image Manipulation showed varying effectiveness across categories—indicating that class-specific settings may yield better results than a uniform approach.

6. Conclusions

This paper presents several important conclusions regarding the role of data augmentation (DA) in image-based waste classification. First, we demonstrated that quantitative analysis of the augmentations -through an iterative pipeline (Figure 4)—can significantly improve model performance. Our iterative pipeline promotes a balanced approach to augmentation that enhances generalization while avoiding overfitting, helping to demystify the "black-box" nature of DA, and can be adapted and applied to other tasks as well. Second, we confirmed that Diffusion Models—currently state-of-the-art—perform exceptionally well in this task, encouraging further research and adoption in the waste classification domain. Despite their high computational cost, our results show they are worthwhile, especially for real-world deployment. Moreover, our combined pipeline demonstrates that Diffusion Models can be effectively integrated with lightweight augmentation techniques to achieve strong performance even under limited resource settings. We also highlighted the effectiveness of Image Erasing and its variants—an underexplored technique in this field, where typically only "Random Erasing" is used. This finding challenges current trends in the literature for this task, where Image Manipulation dominates, and underscores the need for further research into the potential of other erasing-based augmentations. Given our resource constraints, our experiments revealed a trade-off between validation and test performance, emphasizing the importance of evaluating generalization to other datasets when assessing augmentation methods. Notably, the baseline without augmentation achieved a high validation score —without signs of overfitting—yet failed to generalize effectively to the test set. This insight is critical for real-world deployment and further underscores the role of data augmentation in enhancing generalization. Additionally, we analyzed the trade-offs between performance and computational cost, concluding that, in our setup, the performance gains of Diffusion Models justify the added complexity—especially when generalization is a priority. Overall, our findings highlight that the effectiveness of data augmentation depends heavily on the characteristics of the dataset and task at hand; meaningful improvements require analyzing the data closely rather than applying augmentation methods randomly. As such, we advocate for the use of adaptable, iterative frameworks that guide augmentation choices based on dataset characteristics and performance feedback. We hope this study encourages more comprehensive and holistic evaluations of DA techniques—across various datasets and domains—moving beyond simple performance metrics to include factors such as generalization, image fidelity, and efficiency.

References

Abdu, Haruna and Noor, Mohd Halim Mohd. A survey on waste detection and classification using deep learning. *IEEE Access*, 10:128151–128165, 2022.

Adedeji, Olugboja and Wang, Zenghui. Intelligent waste classification system using deep learning convolutional neural network. *Procedia Manufacturing*, 35:607–612, 2019.

Alsabei, Amani, Alsayed, Ashwaq, Alzahrani, Manar, and Al-Shareef, Sarah. Waste classification by fine-tuning pre-trained cnn and gan. *International Journal of Computer Science & Network Security*, 21(8):65–70, 2021.

Aral, Rahmi Arda, Keskin, Şeref Recep, Kaya, Mahmut, and Hacıömeroğlu, Murat. Classification of trashnet dataset based on deep learning models. In *2018 IEEE International Conference on Big Data (Big Data)*, pp. 2058–2062. IEEE, 2018.

Chlap, Phillip, Min, Hang, Vandenberg, Nym, Dowling, Jason, Holloway, Lois, and Haworth, Annette. A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, 65(5):545–563, 2021.

Croitoru, Florinel-Alin, Hondru, Vlad, Ionescu, Radu Tudor, and Shah, Mubarak. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (9):10850–10869, 2023.

de la Rosa, Francisco López, Gómez-Sirvent, José L, Sánchez-Reolid, Roberto, Morales, Rafael, and Fernández-Caballero, Antonio. Geometric transformation-based data augmentation on defect classification of segmented images of semiconductor materials using a resnet50 convolutional neural network. *Expert Systems with Applications*, 206:117731, 2022.

DeVries, Terrance and Taylor, Graham W. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Dhariwal, Prafulla and Nichol, Alexander. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Dosovitskiy, Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Fan, Jinhao, Cui, Lizhi, and Fei, Shumin. Waste detection system based on data augmentation and yolo_ec. *Sensors*, 23(7):3646, 2023.

Garcea, Fabio, Serra, Alessio, Lamberti, Fabrizio, and Morra, Lia. Data augmentation for medical imaging: A systematic literature review. *Computers in Biology and Medicine*, 152: 106391, 2023.

Goceri, Evgin. Medical image data augmentation: techniques, comparisons and interpretations. *Artificial Intelligence Review*, 56(11):12561–12605, 2023.

Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, and Bengio, Yoshua. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Huang, Gao, Liu, Zhuang, Van Der Maaten, Laurens, and Weinberger, Kilian Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Kumar, Teerath, Brennan, Rob, Mileo, Alessandra, and Bendechache, Malika. Image data augmentation approaches: A comprehensive survey and future directions. *IEEE Access*, 2024.

Kumsetty, Nikhil Venkat, Nekkare, Amith Bhat, Sowmya Kamath, S, and Anand Kumar, M. An approach for waste classification using data augmentation and transfer learning models. In *Machine Vision and Augmented Intelligence: Select Proceedings of MAI 2022*, pp. 357–368. Springer, 2023.

Liang, Wen, Liang, Youzhi, and Jia, Jianguo. Miamix: Enhancing image classification through a multi-stage augmented mixed sample data augmentation method. *Processes*, 11(12):3284, 2023.

Majchrowska, S., Mikołajczyk, A., Ferlin, M., Klawikowska, Z., Plantykow, M. A., Kwasigroch, A., and Majek, K. Deep learning-based waste detection in natural and urban environments. *Waste Management*, 138:274–284, 2022a. doi: 10.1016/j.wasman.2021.12.036.

Majchrowska, Sylwia, Mikołajczyk, Agnieszka, Ferlin, Maria, Klawikowska, Zuzanna, Plantykow, Marta A, Kwasigroch, Arkadiusz, and Majek, Karol. Deep learning-based waste detection in natural and urban environments. *Waste Management*, 138:274–284, 2022b.

Nanni, Loris, Paci, Michelangelo, Brahnam, Sheryl, and Lumini, Alessandra. Comparison of different image data augmentation approaches. *Journal of imaging*, 7(12):254, 2021.

Nnamoko, Nonso, Barrowclough, Joseph, and Procter, Jack. Solid waste image classification using deep convolutional neural network. *Infrastructures*, 7(4):47, 2022.

Pachaiyappan, Prabhavathy, Chidambaram, Gopinath, Jahid, Abu, and Alsharif, Mohammed H. Enhancing underwater object detection and classification using advanced imaging techniques: A novel approach with diffusion models. *Sustainability*, 16(17): 7488, 2024.

Patrizi, Arianna, Gambosi, Giorgio, and Zanzotto, Fabio Massimo. Data augmentation using background replacement for automated sorting of littered waste. *Journal of imaging*, 7(8): 144, 2021.

Pawara, Pornntiwa, Okafor, Emmanuel, Schomaker, Lambert, and Wiering, Marco. Data augmentation for plant classification. In *Advanced Concepts for Intelligent Vision Systems: 18th International Conference, ACIVS 2017, Antwerp, Belgium, September 18-21, 2017, Proceedings 18*, pp. 615–626. Springer, 2017.

Rožanec, Jože M, Zajec, Patrik, Theodoropoulos, Spyros, Koehorst, Erik, Fortuna, Blaž, and Mladenić, Dunja. Synthetic data augmentation using gan for improved automated visual inspection. *Ifac-Papersonline*, 56(2):11094–11099, 2023.

Safdar, Muhammad Farhan, Alkobaisi, Shayma Saad, and Zahra, Fatima Tuz. A comparative analysis of data augmentation approaches for magnetic resonance imaging (mri) scan images of brain tumor. *Acta informatica medica*, 28(1):29, 2020.

Shorten, Connor and Khoshgoftaar, Taghi M. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1): 1–48, 2019.

Single, Sam, Iranmanesh, Saeid, and Raad, Raad. Realwaste: A novel real-life data set for landfill waste classification using deep learning. *Information*, 14(12):633, 2023.

Tabak, Michael A, Norouzzadeh, Mohammad S, Wolfson, David W, Sweeney, Steven J, VerCauteren, Kurt C, Snow, Nathan P, Halseth, Joseph M, Di Salvo, Paul A, Lewis, Jesse S, White, Michael D, et al. Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology and Evolution*, 10(4):585–590, 2019.

Trabucco, Brandon, Doherty, Kyle, Gurinas, Max, and Salakhutdinov, Ruslan. Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*, 2023.

Ulhaq, Anwaar and Akhtar, Naveed. Efficient diffusion models for vision: A survey. *arXiv preprint arXiv:2210.09292*, 2022.

Viegas, Tamires Ribeiro. Detection and classification of polluting waste in public spaces. 2024.

Vo, Anh H, Vo, Minh Thanh, Le, Tuong, et al. A novel framework for trash classification using deep transfer learning. *IEEE Access*, 7:178631–178639, 2019.

von Platen, Patrick, Patil, Suraj, Lozhkov, Anton, Cuenca, Pedro, Lambert, Nathan, Rasul, Kashif, Davaadorj, Mishig, Nair, Dhruv, Paul, Sayak, Liu, Steven, Berman, William, Xu, Yiyi, and Wolf, Thomas. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

Voulodimos, Athanasios, Doulamis, Nikolaos, Doulamis, Anastasios, and Protopapadakis, Eftychios. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018(1):7068349, 2018.

Williams, Paul T. *Waste treatment and disposal*. John Wiley & Sons, 2013.

World Bank. Trends in solid waste management. *What a Waste 2.0: A Global Snapshot of Solid Waste Management to 2050*, 2025. URL https://datatopics.worldbank.org/what-a-waste/trends_in_solid_waste_management.html. Accessed: 2025-01-24.

Xue, Yuan, Ye, Jiarong, Zhou, Qianying, Long, L Rodney, Antani, Sameer, Xue, Zhiyun, Cornwell, Carl, Zaino, Richard, Cheng, Keith C, and Huang, Xiaolei. Selective synthetic augmentation with histogan for improved histopathology image classification. *Medical image analysis*, 67:101816, 2021.

Yang, Ling, Zhang, Zhilong, Song, Yang, Hong, Shenda, Xu, Runsheng, Zhao, Yue, Zhang, Wentao, Cui, Bin, and Yang, Ming-Hsuan. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4): 1–39, 2023a.

Yang, Mingxing and Thung, Gary. Classification of trash for recyclability status. *Stanford University*, 2016.

Yang, Suorong, Xiao, Weikang, Zhang, Mengchen, Guo, Suhan, Zhao, Jian, and Shen, Furao. Image data augmentation for deep learning: A survey. *arXiv preprint arXiv:2204.08610*, 2022.

Yang, Suorong, Li, Jinqiao, Zhang, Tianyue, Zhao, Jian, and Shen, Furao. Advmask: A sparse adversarial attack-based data augmentation method for image classification. *Pattern Recognition*, 144:109847, 2023b.

Yang, Zhihu and Li, Dan. Wasnet: A neural network-based garbage collection management system. *IEEE access*, 8: 103984–103993, 2020.

Yu, Youpeng and Grammenos, Ryan. Towards artificially intelligent recycling improving image processing for waste classification. *arXiv preprint arXiv:2108.06274*, 2021.

Zhang, Richard, Isola, Phillip, Efros, Alexei A, Shechtman, Eli, and Wang, Oliver. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Zhou, Kaiyang, Liu, Ziwei, Qiao, Yu, Xiang, Tao, and Loy, Chen Change. Domain generalization: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(4): 4396–4415, 2022.

Appendix 1: Examples of Images and Comparisons

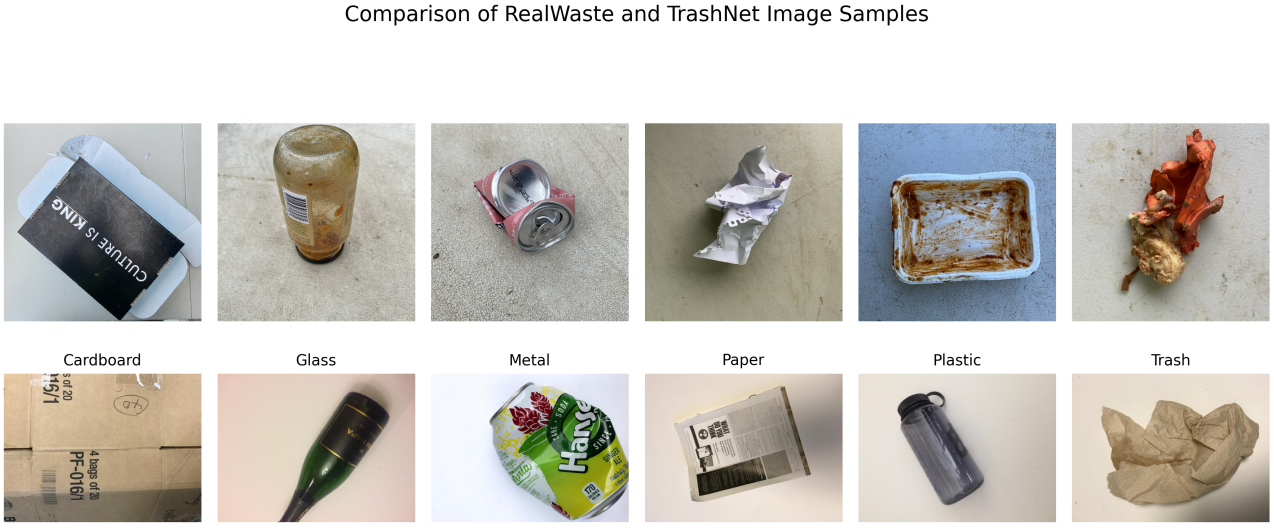

Figure 6. Sample images from the RealWaste (top row) and TrashNet (bottom row) datasets for each waste category.

Appendix 2: Augmented and Original Images

Figure 7. Original and Augmented Images per Category

Appendix 3: Quantitative Differences between Datasets

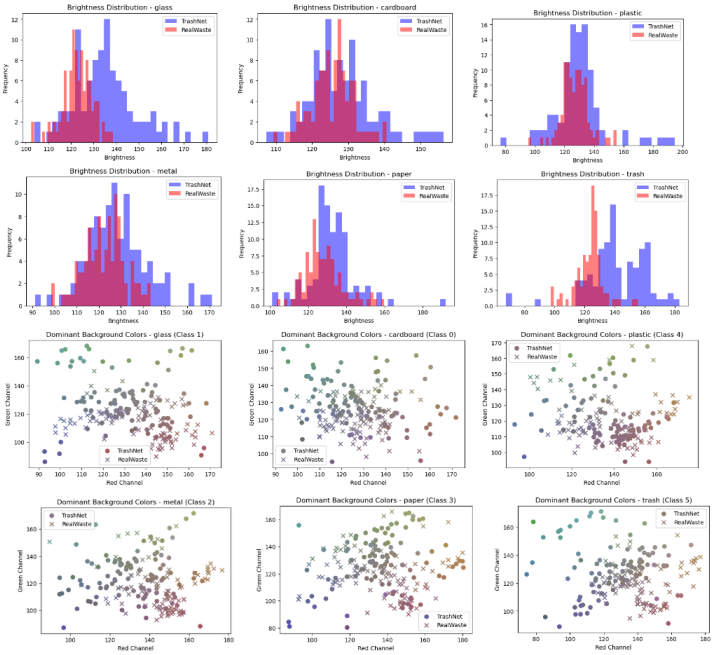

Figure 8. Differences of datasets

The images of the datasets might seem similar at first to the naked eye but clearly, they differ in various metrics, for example, brightness and color composition. Using them, we can truly test for generalization.

Appendix 4: Confusion Matrices

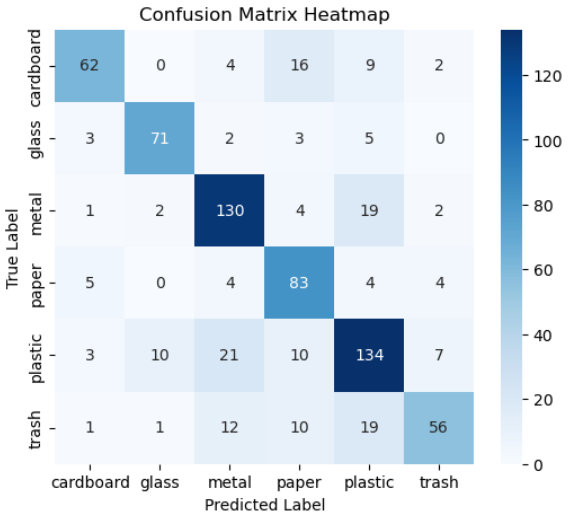

Figure 9. Confusion Matrix for RealWaste [Val]

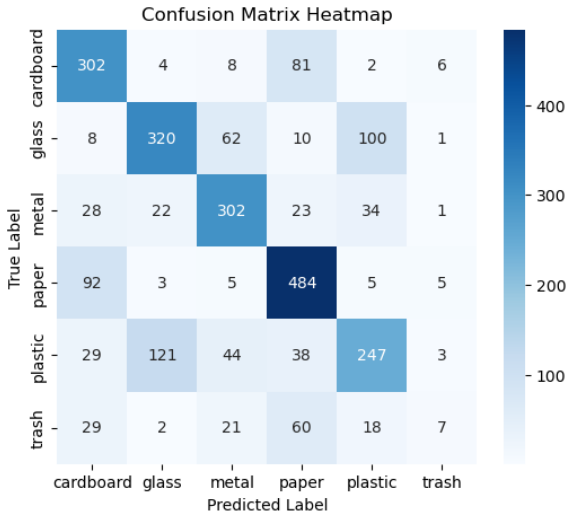

Figure 10. Confusion Matrix for TrashNet [Test]

These figures present the confusion matrices of the best-performing pipeline of Diffusion Models. On the validation set, the model demonstrates strong performance, with only minor misclassifications—primarily between metal-plastic and paper- cardboard. On the test set, similar confusion appears between paper-cardboard, along with additional misclassifications between glass-plastic.

Appendix 5: t-SNE analysis of Iterative pipeline

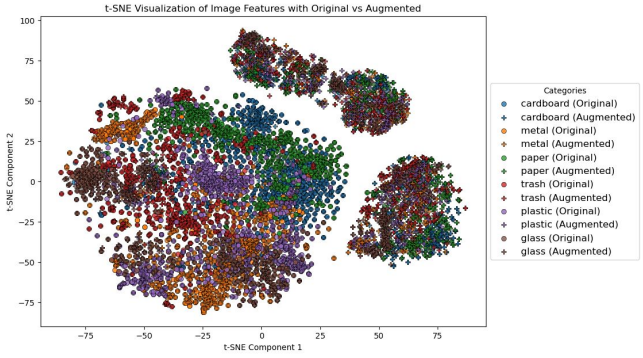

Figure 11. T-SNE Analysis on the Final Erasing Pipeline

This is the graph of our final erasing pipeline, where we see that the generated images (depicted with '+' per class) are generally very close to their non-augmented counterparts. This is also supported by a considerably lower weighted average KL Divergence compared to the initial Erasing pipeline. These metrics guided the hyperparameter tuning in our iterative pipeline.

Appendix 6: Data Augmentation Techniques and Hyperparameters

This appendix outlines the data augmentation techniques and corresponding hyperparameters used in the experiments. The methods are grouped into three DA Techniques: *Image Manipulation*, *Image Erasing*, and *Diffusion Models*.

Image Manipulation	
Rotate	limit=25°, p=0.5
ShiftScaleRotate	shift=0.3, scale=0.2, p=0.3
Affine Shear	shear=5°, p=0.5
HorizontalFlip	p=0.5
RandomResizedCrop	size=524×524, scale=(0.7–1.0), p=0.5
GaussNoise	var=(2,10), p=0.2
CLAHE	clip=2.0, grid=(8,8), p=0.2
ColorJitter	b=0.2, c=0.2, s=0.2, h=0.05, p=0.3
GaussianBlur	blur=(3,7), p=0.2
Resize	resize to 524×524
Image Erasing	
CoarseDropout	max_holes=10, size=(50×50), p=1.0
GridDropout	ratio=0.45, p=1.0
Hide-and-Seek	grid=(12–48), hide_prob=0.2
Erasing	scale=(0.05, 0.5), ratio=(0.25, 4.0), p=1.0
Diffusion Models	
Cardboard	strength=0.8, guidance=9, size=512
Glass	strength=0.6, guidance=12, size=512
Metal	strength=0.8, guidance=14, size=512
Paper	strength=0.75, guidance=14, size=512
Plastic	strength=0.75, guidance=14, size=512
Trash	strength=0.8, guidance=14, size=512

Table 6. Best set of hyperparameters for each of the Data Augmentation Techniques

Parameter Descriptions:

- *p*: Probability of applying the augmentation.
- *limit*, *scale*, *shear*, *shift*: Control geometric transformations (e.g., rotation, scaling).
- *clip*, *grid*, *var*, *blur*: Define the intensity or range for color/contrast, noise, or smoothing effects.

- *CLAHE* (Contrast Limited Adaptive Histogram Equalization): Improves local contrast in images by adjusting brightness in small regions. The *clip* value controls how much contrast is added, and *grid* sets how finely the image is divided.
- *max_holes*, *size*, *ratio*, *hide_prob*: Define the size and distribution of masked regions in image erasing.
- *strength*, *guidance*, *size*: Used in diffusion-based image synthesis. *Strength* controls how much the generated image deviates from the input — higher values allow more creative variation, while lower values keep the output closer to the original image. *Guidance* determines how strongly the model follows the text prompt — high values lead to outputs that match the prompt closely (but with less creativity), while low values allow more diverse or imaginative results. *Size* sets the resolution of the generated image, with higher values producing larger outputs.

Appendix 7

The GitHub repository to reproduce the key insights from this work can be accessed here: Github.