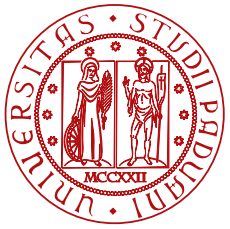

Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione

Master Thesis in Computer Engineering

DiffuWaste: Data Augmentation using Diffusion Model for Waste Semantic Segmentation

Master Candidate

Luca Sambin

Student ID 2056807

Supervisor

Prof. Alberto Pretto

University of Padova

Co-supervisor

Dott. Alberto Bacchin

University of Padova

Dott. Alberto Gottardi

IT+Robotics & University of Padova

Academic Year 2023/2024

Graduation Date 07/03/2024

To my parents
and friends

Abstract

Recent progress in text-to-image and image-to-image generation models have resulted in the development of sophisticated systems capable of generating highly realistic and detailed images. These models have become more accessible, enabling researchers to explore innovative techniques for enhanced manipulation and control. This thesis delves into the potential of Latent Diffusion Models (LDMs), specifically employing some exemplar-driven generation pipeline to address data scarcity issues in semantic segmentation tasks, within the waste sorting domain. The focus is on presenting results obtained from training the semantic model on synthetically generated images. My findings indicate that exemplar-driven augmentation stands out as a competitive technique, particularly beneficial in scenarios with limited data, where scarcity is a challenge. However, it is important to note that a significant gap still exists between synthetic and real images. These results emphasize the potential of zero-shot learning approaches, providing avenues to alleviate or eliminate the costs associated with creating extensive datasets while enhancing the capabilities of computer vision models. The code used in this thesis can be accessed at *https://github.com/lucasambin/DiffuWaste*

Sommario

I recenti progressi nei modelli di generazione text-to-image e image-to-image hanno portato allo sviluppo di sistemi sofisticati in grado di generare immagini altamente realistiche e dettagliate. Questi modelli sono diventati più accessibili, consentendo ai ricercatori di esplorare tecniche innovative per una migliore manipolazione e controllo. Questa tesi approfondisce il potenziale dei Latent Diffusion Models (LDMs), impiegando in particolare alcune pipeline di generazione guidata da esemplari per affrontare i problemi di scarsità dei dati nelle attività di segmentazione semantica, all'interno del dominio della raccolta differenziata. L'obiettivo è presentare i risultati ottenuti dall'addestramento del modello semantico su immagini generate sinteticamente. I miei risultati indicano che lincremento basato sugli esemplari si distingue come una tecnica competitiva, particolarmente vantaggiosa in scenari con dati limitati, dove la scarsità è una sfida. Tuttavia, è importante notare che esiste ancora un divario significativo tra le immagini sintetiche e quelle reali. Questi risultati sottolineano il potenziale degli approcci di apprendimento zero-shot, fornendo strade per alleviare o eliminare i costi associati alla creazione di set di dati estesi, migliorando al contempo le capacità dei modelli di visione artificiale. È possibile accedere al codice utilizzato in questa tesi all'indirizzo *https://github.com/lucasambin/DiffuWaste*

Contents

List of Acronyms

AI Artificial Intelligence

DMs Diffusion Models

GANs Generative Adversarial Networks

LDMs Latent Diffusion Models

mIoU mean Intersection over Union

NLP Natural Language Processing

RoWSS Robotic Waste Sorting System

SD Stable Diffusion

VAEs Variational AutoEncoders

1 Introduction

The latest advancements in image generation models have reached a level of maturity, demonstrating the ability to produce high fidelity and photorealistic images [16, 27].
These images now possess a quality that deems them suitable for practical applications, often rendering them indistinguishable from authentic photographs to the majority of observers [28].

Figure 1.1: On the left, a real images and on the right an image created by Artificial Intelligence (AI) [28].

In addition, text-to-image models have become more widely accessible, thanks to contributions from private companies, educational institutions and the open-

source community. Massive models like Stable Diffusion (SD) are now readily available for anyone interested in experimenting with them. This growing accessibility to such resources, empowers researchers globally to develop innovative techniques that enable improved manipulation and control of generative models.

Another important aspects covered in this thesis is semantic segmentation, which involves identifying objects within an image and generating a pixel-wise mask for each image. The accuracy of this task has significantly improved with the advent of machine learning-driven computer vision algorithms [18, 25, 40]. Furthermore, segmentation models are very useful for a large variety of tasks, such as autonomous driving and medical image diagnostics [32].
Nevertheless, these algorithms require large training datasets, which are often expensive to generate, due to the time and labor-intensive nature of providing pixel-wise annotations for all objects in an image. With the rise of generative image models, such as Latent Diffusion Models (LDMs) [48], a new approach for generating synthetic training data has been introduced. For instance, in the study detailed in this paper [6], the authors demonstrated the utility of generated images in enhancing image classification.

Since the focus of this thesis is on the waste sorting domain, the challenge becomes even more complex, as it is fraught with various difficulties. As the global population expands and urbanization progresses, waste production is anticipated to escalate to 2,6 billion tonnes annually by 2030, marking an increase from the current level of approximately 2,1 billion tonnes [36]. The detection of recyclable waste presents a unique computer vision challenge, demanding the detection of highly deformable and often translucent objects within cluttered scenes, all without the typical contextual information found in datasets focused on human-centric scenarios. As shown in Figure 1.2, a real-life scene is notably different and far more intricate compared to a single patch. Moreover, considering this challenging computer vision task, the availability of labeled patches and scenes for model input is limited. Finally, the challenge is also exacerbated by the need for extensive data and the demand for well-annotated datasets raises significant concerns. Therefore, this thesis attempts to produce synthetic patches using Diffusion Models (DMs), with the aim of reducing the intrinsic knowledge gap in the waste sorting domain.

Figure 1.2: An example of an actual waste environment.

The creation and maintenance of such vast datasets come with substantial expenses, often exceeding the resources available to most researchers [65]. As a result, the scientific community has predominantly focused on optimizing deep learning architectures rather than developing approaches to alleviate the costs associated with acquiring and sustaining large-scale datasets [20].

In summary, the primary objective of this thesis is to leverage state-of-the-art generative models to train a system capable of enhancing detection and semantic segmentation within the domain of waste sorting. Initially, employing exemplar-driven generation techniques, I established a pipeline for generating new synthetic patches and scenes employing Diffusion Models (DMs), comprising images and masks, from a dataset containing limited real waste patches, such as metal and plastic. Subsequently, using these synthetic generated scenes, I trained a semantic model capable of effectively performing semantic segmentation tasks within the waste sorting context.

Consequently, a pertinent question arises: *To what extent can images generated by image-to-image systems enhance the performance of computer vision models?* Specifically, I approached this issue from the perspective of data augmentation, focusing on its application to train a semantic segmentation model, with the goal of improving results beyond those achieved using only real images and the corresponding masks.
Furthermore, I investigated the impact of using solely generated images as input for training the semantic segmentation model. Additionally, I also explored whether synthetic images could address class imbalance in one of the datasets. Lastly, I examined the effects of using only a small fraction (10%) of the entire datasets. Through extensive experimentation, my findings indicate that

exemplar-driven augmentation proves to be a competitive technique under specific circumstances. In particular, exemplar-driven augmentation produces significant benefits in datasets characterized by a limited number of training images per class.

2

State of the Art

Text-to-image represents a burgeoning field within deep learning, where models demonstrate the ability to generate authentic and richly detailed images, based on textual prompts. Creating these models is a complex undertaking that necessitates the seamless integration of both computer vision and Natural Language Processing (NLP) techniques. The latest progress in text-to-image models has resulted in their capacity to produce high-quality images with nuanced semantic content. These advancements open up a plethora of applications, including virtual reality, video games, e-commerce and education. Despite the recent strides enabling commercial applications, generative models remain formidable challenges. This section seeks to delve into the current state-of-the-art in Diffusion Models (DMs) and explore the most promising approaches to address the previously discussed problem.

2.1 Image Generation

Image generation is the task of creating novel images based on a training set. The goal is to ensure that the synthesized images closely adhere to the input distribution. Various architectural approaches are suitable for generating visual content, with notable ones including Variational AutoEncoders (VAEs) [38], Generative Adversarial Networks (GANs) [21], flow-based models [47] and more recently, Diffusion Models (DMs) [27, 53].

While each method has demonstrated considerable success in generating high-quality samples, they also come with their respective limitations. GANs models, for instance, are prone to potentially unstable training and exhibit less diversity in generation, attributed to their adversarial training nature. On the other hand, VAEs relies on a surrogate loss and flow models necessitate specialized architectures to implement reversible transformations.
In contrast, DMs draw inspiration from non-equilibrium thermodynamics, establishing a Markov chain of diffusion steps, as can be seen in Figure 2.2. This process involves gradually introducing random noise to the data and subsequently learning to reverse the diffusion steps, constructing desired data samples from the noise. Unlike VAEs or flow models, DMs follow a fixed learning procedure and their latent variables have high dimensionality, matching that of the original data.

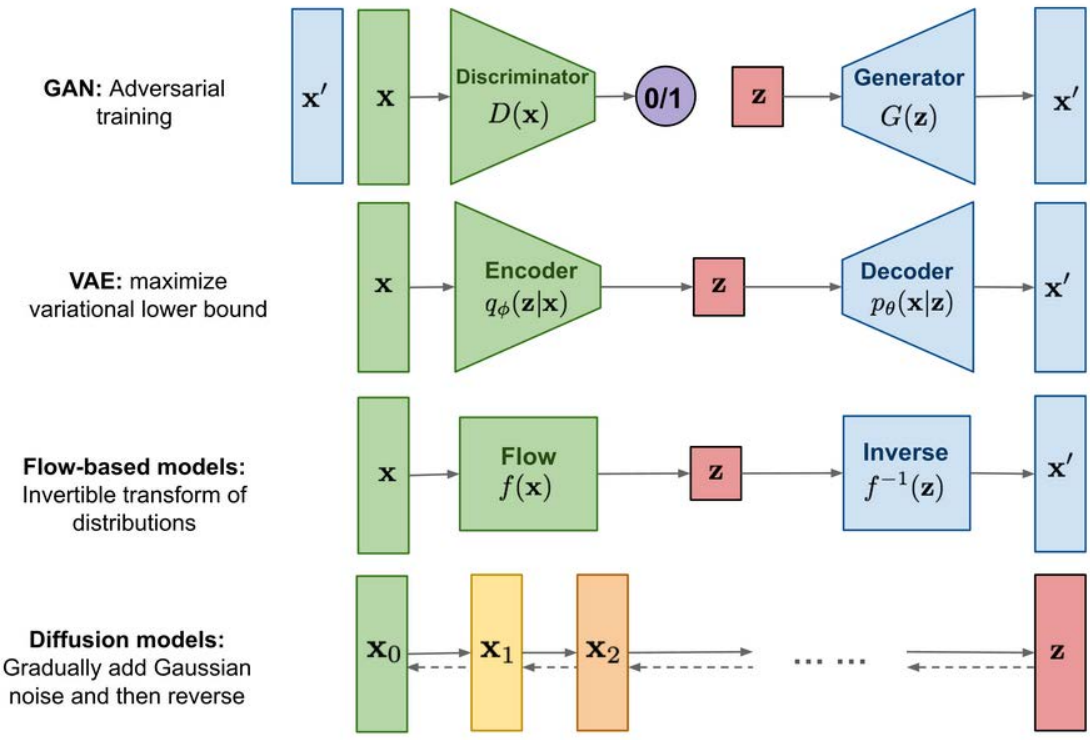

Figure 2.1: A brief overview of the different types of generative models [59].

2.2 Denoising Diffusion Probabilistic Models

The concept of DMs was initially presented in [27, 53] for the purpose of generating realistic textures and images. Denoising Diffusion Probabilistic models consists of two components: a forward diffusion process and a reverse diffusion process.

2.2.1 Forward Diffusion Process

Starting with a sample, denoted as $\mathbf{x}_0$, drawn from a distribution q ($\mathbf{x}_0 \sim q(\mathbf{x})$), one can systematically introduce Gaussian noise with increasing variance β_t defining a forward diffusion process. Iterating this process T times results in $\mathbf{x}_1, \ldots, \mathbf{x}_T$ noisy samples. Letting $T \to \infty$ will let the distribution of $\mathbf{x}_T$ approach the isotropic Gaussian distribution. In practice T is chosen large enough that $\mathbf{x}_T$ is close to being isotropically distributed and close to Gaussian. The authors of [27] have chosen to define this progression via

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$$

where $\{\beta_t \in (0,1)\}_{t=1}^{T}$ is a deterministic set of variance parameters.
Therefore, it follows that the probability distribution of $\mathbf{x}_t$ can be computed directly from the input $\mathbf{x}_0$ as

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$.

2.2.2 Reverse Diffusion Process

In the reverse process, isotropic Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ gets changed by progressively removing the added noise from a noisy sample. As the reverse probability distribution $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is not readily available it can be estimated by its approximation as

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

where $\boldsymbol{\mu}_\theta$ and $\boldsymbol{\Sigma}_\theta$ are approximated by neural networks. Subsequently, these are trained to lead to the generation of random samples that adhere to the original

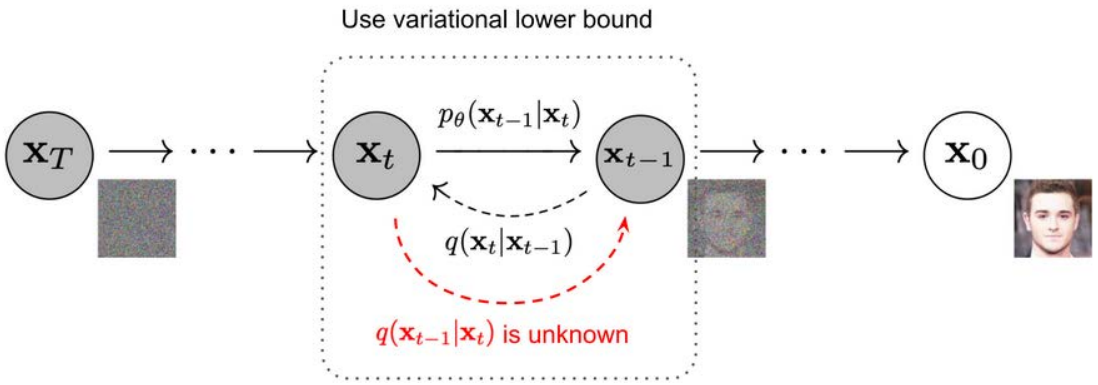

Figure 2.2: The Markov chain of forward (reverse) diffusion process of generating a sample by slowly adding (removing) noise [27, 59].

data distribution $q(\mathbf{x}_0)$. Various studies have proposed minor adjustments to the reverse process, significantly diminishing the required number of time steps T, to achieve good quality final samples [34, 42, 54].

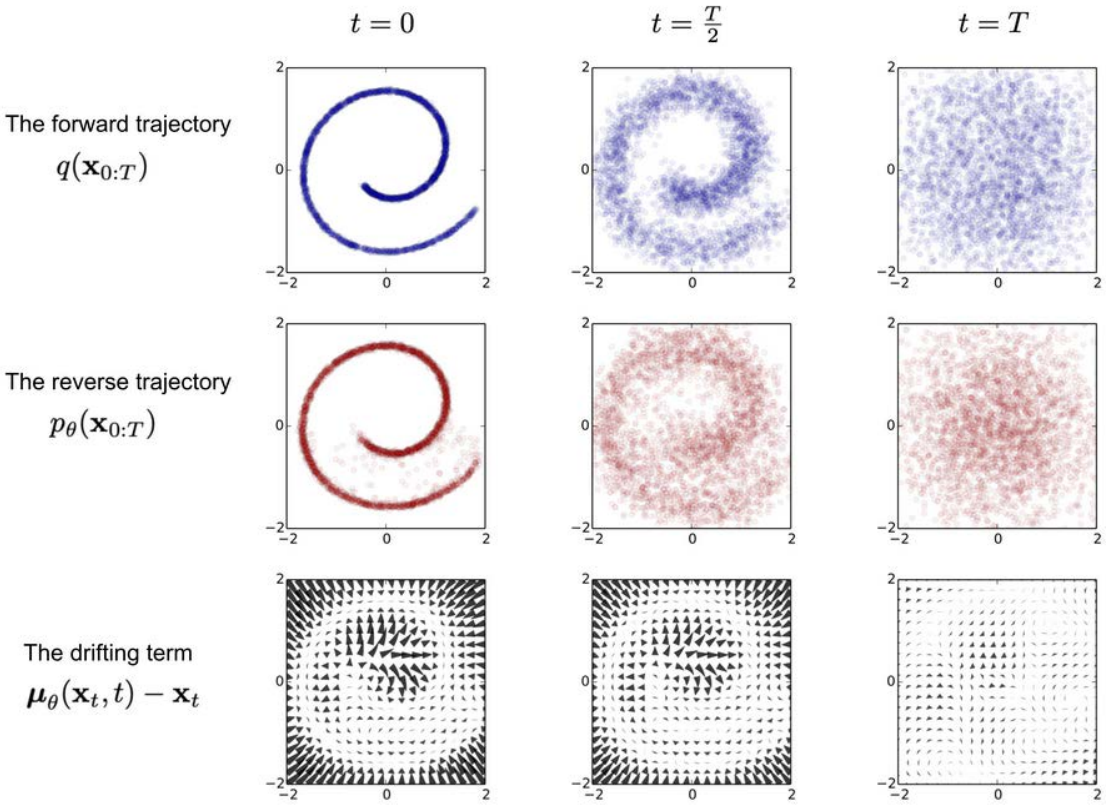

Figure 2.3: An example of training a diffusion model for modeling a 2D swiss roll data [53].

2.2.3 Improvements to Diffusion Models

When training the parameters of the model p_θ one can make use of a variational lower bound on the negative log likelihood, as shown in [27, 53]. Training of $\boldsymbol{\mu}_\theta$ will then proceed by estimating a noise factor $\boldsymbol{\epsilon}_\theta$ that gets added to the current image, while $\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t) = \sigma_t^2 \mathbf{I}$ can be assumed fixed. Fixed in the sense that σ_t can be substituted with β_t or $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}$. The authors of [27] reported that predicting $\boldsymbol{\epsilon}$ worked best and trained their network with a simplified loss term

$$L_t^{\text{simple}} = \mathbb{E}_{t\sim[1,T],\mathbf{x}_0,\boldsymbol{\epsilon}_t}\left[\|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|^2\right]$$

Other studies enhance these findings, as demonstrated in [42], where the authors highlight the benefits of incorporating the learning of variance $\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)$. The previous investigation [42], also introduced several enhancement techniques aimed at assisting DMs in achieving lower negative log likelihood. One of these improvements involves the implementation of a cosine-based variance schedule. In fact, the choice of the scheduling function can be arbitrary, as long as it yields a nearly linear decrease in the middle of the training process and exhibits subtle changes around $t = 0$ and $t = T$.

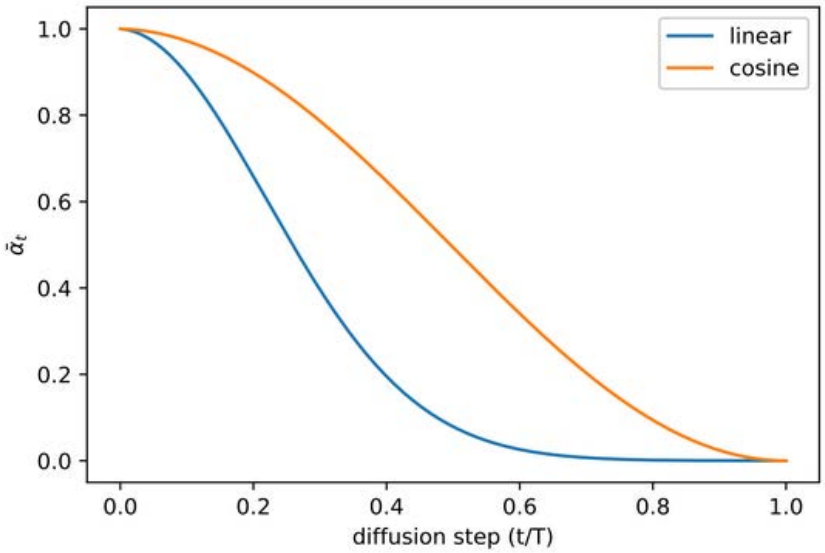

Figure 2.4: Comparison of linear and cosine-based scheduling of β_t during training [42].

Figure 2.5: Comparison of linear (top) and cosine-based (bottom) scheduling on a image [42].

2.3 Latent Diffusion Models

Although Denoising Diffusion Probabilistic Models have facilitated the generation of high-quality images, delivering state-of-the-art results, they are struggling with an inherent limitation that subsequent iterations of improvements have failed to address. The main challenge lies in their operation within pixel space, which involves adding and eliminating noise in a tensor of the same size as the input tensor. Indeed, this dependence on pixel space requires enormous computational resources for training these models.

Consequently, researchers from Ludwig Maximilian University of Munich and Runway ML proposed a solution in their publication *High-Resolution Image Synthesis with Latent Diffusion Models* [48]. The main idea involved a strategic shift from pixel space to latent space, leveraging previously trained autoencoders. This transition allowed for the acceleration of both training and inference calculations, as the latent space represents images in a lower-dimensional space. Furthermore, this approach also allowed to find a balance between preserving the quality of detail and reducing computational complexity.

Hence, the functioning of Latent Diffusion Models (LDMs) can be succinctly summarised in the diagram provided in Figure 2.6. In the initial training phase, the goal is to acquire a representation of the given image in the latent space $\mathcal{Z}$ through the encoder $\mathcal{E}$. Subsequently, Gaussian noise is introduced into the diffusion process until $\mathcal{Z}_t$ is attained.
For the inverse process, a U-Net network [49] is employed. Nevertheless, the real strength of this approach lies in its capacity to condition the generation. This is accomplished by utilizing a dedicated encoder τ_θ, which maps the conditioning factors onto the intermediate layers of the U-Net through cross-attention layers. Ultimately, the outcome in the latent space is transformed back to pixel space

via the decoder $\mathcal{D}$.

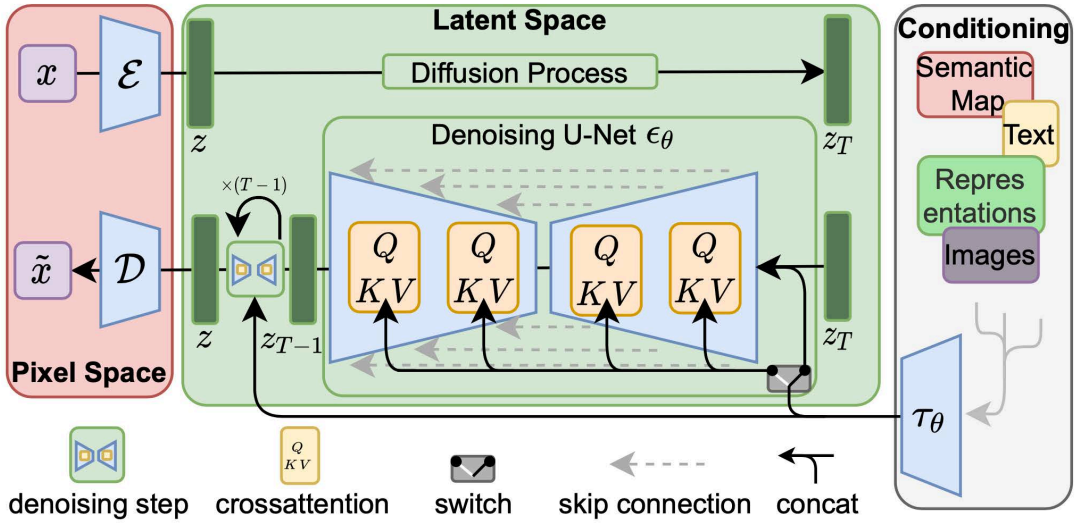

Figure 2.6: The architecture of Latent Diffusion Models (LDMs) [48].

2.4 Stable Diffusion

As already said previously, the realm of text-to-image models experienced a significant surge in popularity and capabilities throughout 2022. A pivotal catalyst in shaping this transformation in public perception has been Stable Diffusion (SD), an open-source model with publicly released weights and architecture. This development has prompted extensive efforts from researchers and enthusiasts to optimize and expand the project's functionalities. The forefront of these endeavors is spearheaded by Stability AI, a generative AI startup. Thanks to the open-source philosophy, this model is compatible with consumer-available hardware, empowering the community to harness its capabilities across a diverse array of applications. SD, adhering to the LDMs architecture established in the paper [48], as detailed in section 2.3, serves as the foundation for this advancement. The proposed technique is versatile and extends its applicability to tasks such as inpainting, outpainting, image-to-image generation and image upscaling, which are all tasks that can be performed from SD.

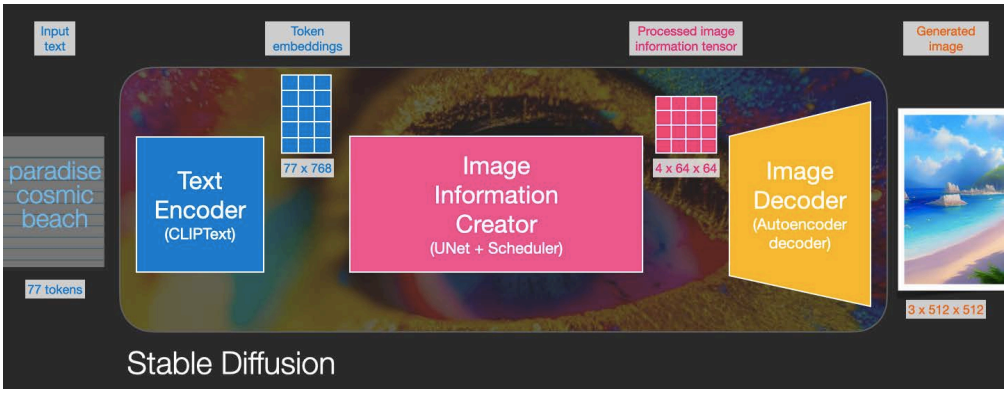

Figure 2.7: The Stable Diffusion main components [3].

Figure 2.7 presents a high-level diagram illustrating the principal components of the model:

- **Text Encoder**: It generates an encoded representation based on the textual description of the data, aiming to guide the diffusion process and guarantee alignment between the resulting image and the provided description. The initial iteration of SD employs CLIP [44], while its second version incorporates OpenCLIP [10]. In both cases, the text encoder collaborates with an image encoder. CLIP and OpenCLIP are designed to optimize the similarity between these two encoding, facilitating the model's ability to establish associations between images and their corresponding descriptions;

- **Image Information Creator**: It seeks to employ the diffusion process for noise reduction in the image through the manipulation of latent space information. As the process progresses, additional information is introduced to improve the similarity between the image and the description provided. Notably, this operation occurs in the latent space, leading to enhanced efficiency and serving as a pivotal advancement. Figure 2.8 offers a visual depiction of the denoising process guided by the text encoder;

- **Image Decoder**: The final image is generated by extracting and utilizing the compressed information stored in the latent space. This step is executed just once to form the final image.

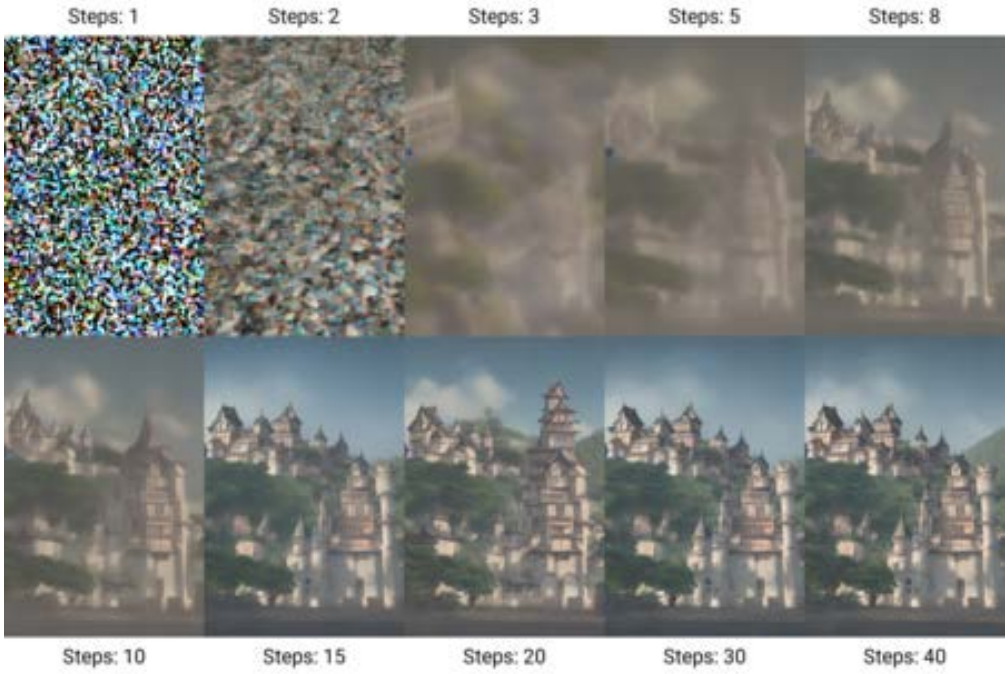

Figure 2.8: The denoising process used by SD [60].

2.5 Data Augmentation

Data augmentation serves as a method in the realm of machine learning to enhance model performance. The fundamental idea is to broaden the variety of training data, enabling the model to handle real-world data more accurately. Essentially, the goal is to enhance the generalizability of machine learning models. Despite its considerable potential, research has predominantly concentrated on refining model architectures rather than enhancing existing data augmentation techniques [13].
In practice, most data augmentation techniques, employed for real-world issues, are custom-designed. This is due to the fact that not all available transformations are applicable in every scenario. For instance, the horizontal flipping transformation is irrelevant in the MNIST digit recognition task. As a result, the development of augmentations necessitates expertise from machine learning professionals, to counteract the rising costs of creating computer vision models. This challenge was significant enough that in 2018, OpenAI identified the automatic search for augmentations as an unresolved problem [30].

Addressing this challenge, Google Brain researchers introduced *AutoAugment* in 2019 [13]. This data augmentation method autonomously explores optimal combinations of transformations to create a policy that yields favorable outcomes without requiring meticulous, ad-hoc design.

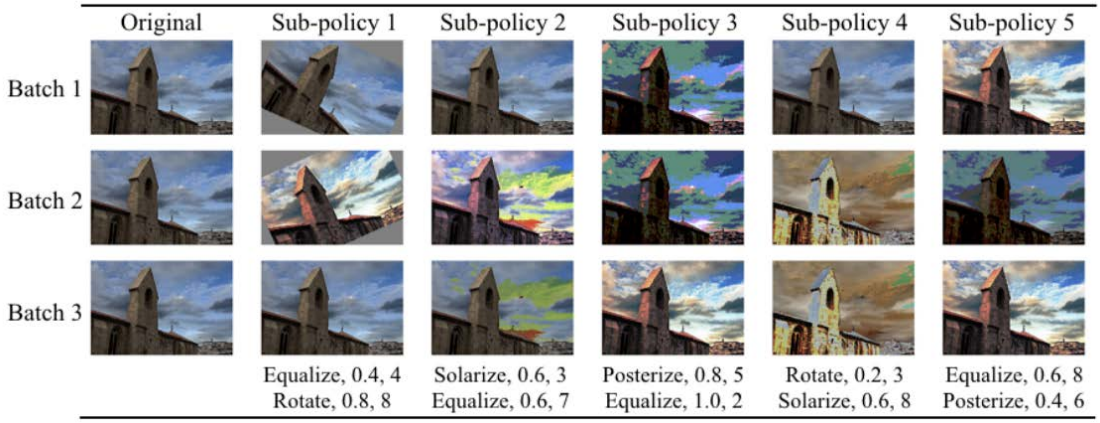

Figure 2.9: An example of images augmented by AutoAugment [13].

Despite the promising outcomes achieved by automated augmentation policies [13], their computational cost makes their massive use in training deep learning models impractical. The higher computational costs arise from an additional search phase, making the dual learning process, which essentially consists in training the network while simultaneously conducting a search in the augmentations space, unfeasible for many tasks. In the original *AutoAugment* publication, an attempt was made to mitigate this by performing the search task on a smaller dataset before transferring the results to the larger original dataset.

However, Google Brain researchers, in the publication *RandAugment* [14], presents findings that contradicts this approach. Consequently, they propose a novel automated augmentation technique that circumvents these issues by eliminating the search task. Their solution involves significantly reducing the search space, enabling the identification of the optimal combination through grid search on the two proposed hyperparameters.

On the other hand, the scientific community has delved into methods for generating new training data, not exclusively relying on the transformation of existing data. Some efforts have focused on the automatic creation of entirely new data. A particularly intriguing approach in recent years has been the adoption

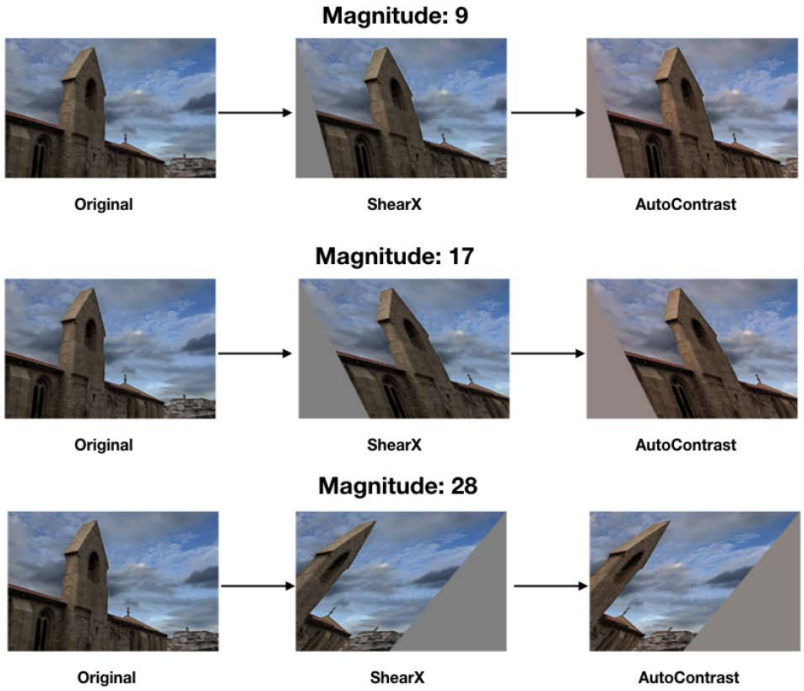

Figure 2.10: An example of images augmented by RandAugment [14].

of copy & paste augmentation for segmentation tasks. This technique involves extracting objects from certain images and inserting them into the backgrounds of other images, effectively generating new training images at no additional cost. Furthermore, this approach offers numerous possibilities for combinations, allowing for exploration of diverse implementation methods. One notable success in this domain is outlined in the publication *Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation* [20]. The researchers demonstrated that a straightforward strategy, involving the selection of random objects and pasting them into random locations, produced results that enhance the baselines of various problems. The outcomes of applying this technique to two images can be seen in Figure 2.11.

Figure 2.11: An example of Simple Copy-Paste augmentation [20]. Random scale jittering is applied on two random training images and then a subset of instances from one image is randomly selected to paste onto the other image.

The subsequent logical advancement in the realm of generating new training data involves the incorporation of text-to-image models. The surge in capabilities of such models in recent years has given rise to a trend in generative data augmentation [16, 48]. This trend advocates for the utilization of highly advanced image generation models to generate entirely new images within the training dataset. Consequently, this approach entailed artificially enhancing the diversity of the data. Recent works, notably in April 2023, such as *Synthetic Data from Diffusion Models Improves ImageNet Classification* [6], are notable for their compelling results. These studies demonstrated that augmenting training data with images generated by text-to-image models yields models that significantly surpass previous benchmarks. Moreover, additional research has indicated the feasibility of training classification models using solely synthetic images, achieving competitive results [52].

While the primary objective slightly differs from the outcomes of the aforementioned publications, the current investigation seeks to evaluate the effectiveness of exemplar-driven generation techniques in improving the performance of semantic segmentation models within the domain of waste sorting. Specifically, in this study, the primary application of DMs lies in attempting to generate diverse and varied waste patches. Subsequently, these patches will be used for image composition and the training of the semantic segmentation model.

3 Methods

A particularly intriguing aspect of this thesis, is to alter or entirely generate the object in question, so as to preserve the visual characteristics corresponding to its object class, thus achieving significant diversity.
Considering the impressive results produced by generative models, one might contemplate creating images depicting objects belonging to a specific class. The authors of [6] demonstrate that these generated images can contribute to image classification. However, focusing solely on classification may prove insufficient when addressing the challenge of semantic segmentation. In fact, in order to generate precise training data, it becomes essential to outline the exact position of the object.

This chapter introduces some techniques, such as image harmonization and inpainting, that show how to enhance a semantic segmentation dataset with Diffusion Models (DMs). The approach uses a pre-trained diffusion model, Stable Diffusion (SD), specifically using the following checkpoint: SD 1.4.

3.1 Dataset

The datasets used include ZeroWaste [7] and a private one called IT+Robotics-V (ITR-V). Both datasets contain realistic and complex scenes featuring a conveyor belt with a lots of waste on top. The ZeroWaste dataset includes four types of waste materials: soft plastic, rigid plastic, cardboard and metal. On the other

hand, the IT+Robotics-V dataset was collected in a real waste sorting plant and focuses on a single type of contaminant, namely plastic.

3.1.1 ZeroWaste

The purpose for which this dataset was collected is the following: to reduce environmental pollution. In fact, less than 35% of recyclable waste is actually recycled in the United States [2], leading to increased land and sea pollution, representing a major concern for environmental researchers and the general public.

The crux of the problem lies in the inefficiencies of the waste sorting process, which involves the separation of paper, plastic, metal, glass and other materials. This challenge is compounded by the messy and intricate nature of the waste stream. Recyclable waste detection presents a particular challenge for computer vision, requiring the identification of highly deformable and often translucent objects in cluttered scenes, lacking the typical contextual information found in human-centric datasets.

Figure 3.1: On the left, there is an example of two images extracted from the ZeroWaste dataset. The corresponding ground truth instance segmentation is displayed on the right [7].

As can be seen from Figure 3.1, the ZeroWaste dataset includes removable objects of four material types: soft plastic, rigid plastic, cardboard and metal. These objects form the foreground, while the background consists exclusively of the conveyor belt. At the end of this conveyor belt only paper or cardboard objects should remain. Therefore, the background includes the conveyor belt and paper/cardboard objects. Table 3.1 provides a concise overview of all images and objects within the dataset.

Split	#Images	Cardboard	Soft Plastic	Rigid Plastic	Metal	#Objects
Train	3002	12940	4862	1160	263	19225
Validation	572	2167	855	305	60	3387
Test	929	3428	1236	315	63	5042

Table 3.1: Statistics detailing the overall count of objects in the training, validation and test splits of the ZeroWaste dataset [7].

3.1.2 IT+Robotics-V

As can be seen from Figure 3.2, the IT+Robotics-V dataset contains removable objects of only two types: contaminants and non-contaminants.

All the contaminants objects constitute the foreground, while in the background there are only the conveyor belt and the non-contaminants objects. Only non-contaminants objects must remain at the end of the conveyor belt. Table 3.2 provides a brief summary of all images and objects within the dataset.

Split	#Images	Contaminant	Not Contaminant
Train	907	447	460
Validation	232	103	129
Test	362	138	224

Table 3.2: Statistics detailing the overall count of objects in the training, validation and test splits of the IT+Robotics-V dataset.

3.2 Image Composition

The practice of extracting the foreground from one image and integrating it into another, with the aim of creating a realistic composite, is a widely employed

Figure 3.2: On the left, there is an example of two images extracted from the IT+Robotics-V dataset. The corresponding ground truth instance segmentation is displayed on the right.

operation in photo editing. Several methods [11, 12, 15, 31, 55, 58] have been proposed, with a particular emphasis on image harmonization to improve the realism of the composite. More traditional approaches [11, 31, 56] often rely on handcrafted features to match color distributions, while more recent work [8, 22] use deep semantic features to enhance robustness. A recent breakthrough, *DCCF* [63], introduced four human comprehensible neural filters in a pyramid structure, achieving state-of-the-art color harmonization results.
However, these methods presume semantic harmony between the foreground and background, adjusting the composite mainly in the low-level color space while preserving the structure. In contrast, this thesis addresses the semantic image composition, also taking into account the challenging aspect of semantic disharmony. As mentioned previously, in fact, waste detection represents a unique challenge for computer vision, since the same object can be in very different conditions, starting from the shape (deformable) up to the appearance which can be more or less worn. Furthermore, it is also important to consider the type of material an object is made of, given that when dealing with translucent

objects, the illumination given by the lights and the arrangement on the conveyor belt play a fundamental role.

3.2.1 Image vs Textual prompt

Within the realm of semantic image editing, there is an increasing emphasis on text-guided image manipulation. Early works [1, 4, 61] employed pre-trained GANs generators [33] and text encoders [44] to progressively refine images according to textual prompts. Nevertheless, these methods based on GANs, have encountered difficulties when handling complex scenes or diverse objects, due to the limited modeling capabilities of GANs.

The swift rise and advancement of DMs [45, 46, 51] is demonstrated by their ability to synthesize diverse and high-quality images. Several studies [5, 26, 35, 37, 39, 43, 50] leverage DMs for text-guided image editing. In particular, *DiffusionCLIP* [37], *Imagic* [35] and *DreamBooth* [50] fine-tune DMs for specific text prompts. Another example is *Blended Diffusion* [5], which introduced a multi-step blended process to execute local manipulation using masks provided by the user.

Although these methods yield noteworthy outcomes, I think that the language guidance still lacks precise control, while images offer a more effective way to express concrete ideas and highlight even the smallest details. As the proverb goes, *"a picture is worth a thousand words"*. In fact, as can be seen from Figure 3.3, the results obtained using images as references significantly exceed those obtained through textual prompts. Therefore, this thesis focuses on exemplar-based image editing.

Figure 3.3: A comparison between progressively precise textual description and image as guidance. Using image as reference helps preserve finer details [64].

3.3 Copy & Paste

The first approach to address the main problem of this thesis, which is the generation of new scenes starting from a limited number of patches, is one of the simplest but still effective. Obviously, with this first attempt, the main objective is not to reduce the intrinsic semantic disharmony in the domain of waste sorting, but to have an almost zero-cost approach, which does not require the use of generative models. The salient parts of the procedure are shown below:

- Firstly, I extracted all individual patches from both datasets, ensuring a clear separation throughout the entire process to prevent any contamination;
- Secondly, I randomly selected a negative image, i.e. a scene used as a background and no positive object to be removed from the conveyor belt. An example of a negative image can be seen in Figure 3.4;
- Thirdly, I randomly selected a specific number of patches for image composition, constituting the foreground. The number of objects chosen varies depending on the dataset. Specifically, for the ZeroWaste dataset, I opted for a maximum of four positive patches per image, while for the IT+Robotics-V dataset I limited myself to a maximum of two positive patches. This choice is due to the fact that this model will be used in a Robotic Waste Sorting System (RoWSS), in which the number of positive objects to be removed is strictly linked to the speed with which the robot is able to remove the object in question. Furthermore, I also tried to make the scenes consistent with those present in both original datasets. An example of some patches is illustrated in Figure 3.4;
- Finally, I copied the selected patches and then pasted them into the background image at random positions. An example of a final image composition can be found in Figure 3.4.

Using this approach I was also able to save the related semantic masks for each patch, which will be used later for training the semantic model.

From now on, all subsequent approaches are based on a zero-shot image generation, since the computational resources I was able to use, did not allow to retrain or fine-tune the entire model.

Figure 3.4: On the top left, an example of negative image (background). On the top right, an example of patches retrieved from the ZeroWaste dataset (foreground). On the bottom, the result obtained after copy & paste, with the relative segmentation mask.

3.4 Image Harmonization

The second approach extends the first one. In addition to a simple copy & paste image composition, I also incorporated a pre-trained DMs, specifically SD 1.4, to achieve a more photorealistic image. I made this choice, that is, to use a pre-trained version of SD, instead of fine-tuning or retraining the entire model, because my computational resources were not sufficient. In this way, I could exploit the potential of DMs with the main purpose of making the new composed scene more faithful to reality, without putting too much emphasis on diversity.

This was possible using the approach presented in the *Cross-domain Compositing with Pretrained Diffusion Models* paper [23]. The main goal of this paper was to broaden the capabilities of DMs, with a specific emphasis on tasks such as image blending, object immersion, texture replacement and even CG2Real translation or stylization. The authors used a localized, iterative refinement scheme that in-

corporates contextual information from the background scene into the inserted objects, allowing finer control over the degree and types of changes the object can undergo.
As previously stated, this method places a greater emphasis on realism rather than diversity. An example showing this approach applied to both datasets is shown in Figure 3.5.

Figure 3.5: On the left the copy & paste version, on the center the segmentation mask and finally, on the right, the final harmonized result. Above an example from the ZeroWaste dataset, while below an example from the IT+Robotics-V dataset.

3.5 Image Inpainting

The third approach further expands the first two approaches. In fact, in addition to a simple copy & paste image composition and harmonization, this approach is able to exploit all the potential of SD, using inpainting, with the aim of obtaining a more photorealistic and varied image.

Inpainting is the task of filling specific parts of an image delimited by a mask. DMs inherently possess capabilities for inpainting masked areas. Previous work [53], demonstrated the potential of inpainting with DMs and *RePaint* [41] was able to achieve realistic outcomes even with extreme masks, by modifying the reverse diffusion process. However, *RePaint* primarily concentrated on pixel space rather than the latent space utilized in SD. Although it can be applied in the latent space of SD, it results in a loss of pixel-level control, leading to

inaccuracies in mask-object correspondence.
Other studies, such as *GLIDE* [43] and *Blended diffusion* [5], demonstrated the effectiveness of text-guided image inpainting with DMs. Fine-tuning additional parameters for SD proved beneficial in developing a reliable inpainting method, building upon the trained SD model [48]. *SmartBrush* [62] showcased that with additional regularization, it is possible to fine-tune a diffusion model to fill the entire masked area with the inpainted object while being guided by text and mask, which it's something that standard SD inpainting struggles with.

This was possible using the approach presented in the *Paint by Example: Exemplar-based Image Editing with Diffusion Models* paper [64]. In fact, for the first time, in this paper the authors investigated exemplar-guided image editing for more precise control than a simple textual prompt. They were able to achieve this by leveraging self-supervised training to disentangle and rearrange the original image and the exemplar. The entire framework involves a single forward of the DMs without any iterative optimization. Importantly, this approach works effectively even without the need to fine-tune or retrain the entire SD model.
As mentioned before, this approach places a greater emphasis on patch variety rather than fidelity to the reference patch, provided as input. An example of this approach applied to both datasets is presented in Figure 3.6.

Figure 3.6: From left to right: the negative image (background), the segmentation mask of the patch to be inpainted, the reference patch and finally, the image inpainting result. Above an example from the ZeroWaste dataset, while below an example from the IT+Robotics-V dataset.

3.6 Image Inpainting + Harmonization

This final approach seeks to exploit the advantages of the two previous models, while minimizing their disadvantages. As expected, the main limitation of image harmonization lies in the degree of variety achievable for each patch. Essentially, image harmonization aims to improve the realism of the scene without introducing substantial additional knowledge. On the other hand, image inpainting is able to achieve significant variety, but encounters challenges with mask-object correspondence, resulting in inaccuracies. Furthermore, the results of this approach can vary greatly, ranging from a partially incorrect mask to a completely blank scene. Compared to the image harmonization approach, the final results of image inpainting are also not that photorealistic. An example showing these problems can be seen in Figure 3.7.

Figure 3.7: Above an example of the image harmonization approach where the lack of variety can be seen. Below an example from the image inpainting approach where you can see the limit in the mask-object correspondence.

Indeed, inspired by the methodology presented in *Effective Data Augmentation With Diffusion Models* [57], which aims to tackle the issue of limited diversity in data augmentation through image-to-image transformations using pre-trained text-to-image DMs, I tried to integrate the strengths of the aforementioned approaches.
Initially, after collecting all patches from both datasets, I augmented individual objects using the inpainting approach, as done in section 3.5. Subsequently, following the standard image composition applied in the copy & paste approach, I further used the harmonization approach (section 3.4) to enhance realism. This approach allowed me to achieve both wide variety and high fidelity, avoiding the limitations of previous approaches. Figure 3.8 shows an example of patches augmented employing the inpainting approach, while Figure 3.9 shows the final results incorporating image harmonization after standard image composition.

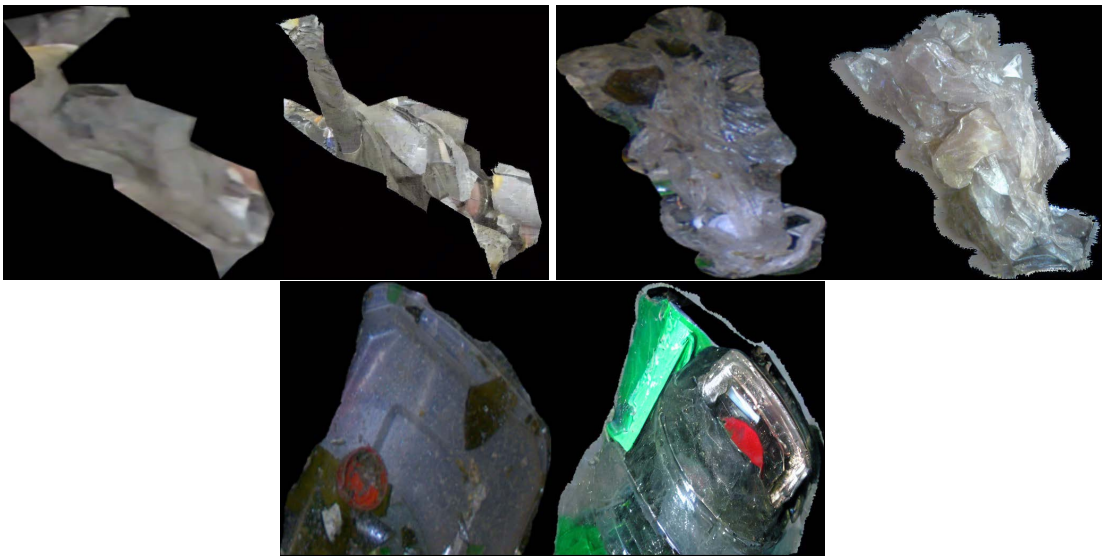

Figure 3.8: An example of some patches augmented using the image inpainting approach.

Figure 3.9: On the left the copy & paste version with the patch augmented by image inpainting, on the center the segmentation mask and finally, on the right the final harmonized result. Above an example from the ZeroWaste dataset, while below an example from the IT+Robotics-V dataset.

4

Experiments

This chapter is dedicated to empirically assessing the effectiveness of exemplar-driven generation techniques for data augmentation in improving the performance of computer vision models. My approach involves evaluating the competitiveness of this data augmentation technique in real-world tasks, such as semantic segmentation within the waste sorting domain.
Additionally, I examined the dynamics of the relationship between real and synthetic images, evaluating the amount of information that can be gleaned from synthetic images alone. Through these investigations, I sought to demonstrate that the proposed augmentation approaches can improve the performance of segmentation models.

My results support the hypothesis that exemplar-driven generation is a competitive data augmentation technique in real-world tasks, particularly showing its importance in scenarios with sparse training data. However, it can be noted that including synthetic images in a small dataset only makes sense up to a certain threshold, especially when sufficient real training images are available. Moreover, I will present evidence showing the gap between the results obtained using only synthetic versus real images in training a computer vision task. In the subsequent test, various semantic segmentation models were evaluated and the most promising ones are reported below:

- ***U-Net*** [49] as segmentation model, ***ResNet-50*** [24] as encoder and ***ImageNet*** as weights;

- ***DeepLabV3+*** [9] as segmentation model, ***mobilenet_v2*** [17] as encoder and ***ImageNet*** as weights;
- ***DeepLabV3+*** [9] as segmentation model, ***mobilevitv2_200*** [19] as encoder and ***ImageNet*** as weights.

Finally, all the results obtained are fully replicable, since the training and testing of these models are performed using the *Segmentation Models* library [29].

4.1 ZeroWaste: Balance vs Imbalance

In this section, I will analyze the distribution of data within the ZeroWaste dataset, since it is the only dataset containing more than one positive class (soft plastic, hard plastic and metal). As can be seen from the Table 4.1, resembling the one presented in Chapter 3, it is evident that the distribution of patches is not uniform.

Split	**#Objects**	**Soft Plastic**	**Rigid Plastic**	**Metal**
Train	6285	4862	1160	263
Real Distribution		77,36%	18,46%	4,18%
Uniform Distribution		33,33%	33,33%	33,33%

Table 4.1: Statistics detailing the overall distribution of objects in the ZeroWaste training dataset [7].

Indeed, each class possesses a distinct number of patches and since certain classes have more patches than others, the outcomes may be influenced by this imbalance. As explained in the Chapter 3, in the sampling phase, during which the patches that will subsequently be used to compose the scenes to be generated will be selected, I aimed to employ a more uniform distribution, deviating from the distribution in the original dataset.
The results of this comparison are presented in the following table.

Unet+resnet50	Copy & Paste	Harmonization	Inpainting
Imbalance	48,791%	47,93%	47,405%
Balance	49,537%	45,429%	52,267%
Baseline		**48,861%**	

Table 4.2: Results achieved using different sampling to create scene to generate for the ZeroWaste dataset [7].

Table 4.2 shows the results obtained using copy & paste, image harmonization and image inpainting approaches, which were presented in Chapter 3. The **Baseline** represents the results obtained using only real images from the ZeroWaste dataset, without any augmentation. Consequently, **Imbalance** results arise from using the original probability distribution of the dataset, while **Balance** results emerge from using a more uniform probability distribution.
Examining the results summarized in Figure 4.1, which provides a concise overview of the data presented in the previous table, it becomes evident that using a more balanced sampling approach tends to produce superior overall results.

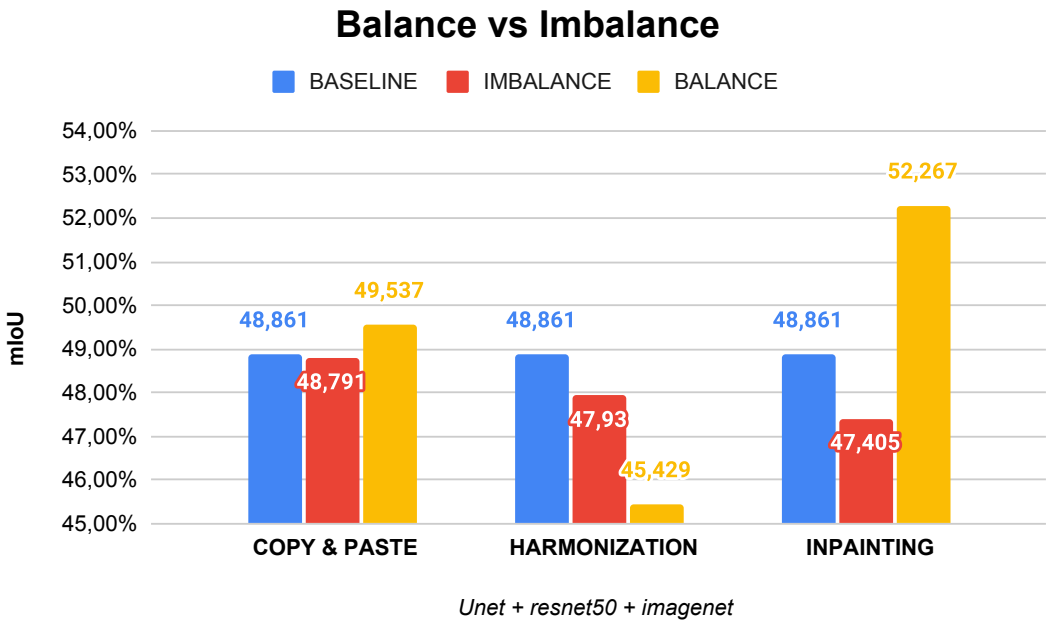

Figure 4.1: A simple graph showing the results of using different sampling distribution for the ZeroWaste dataset [7].

4.2 Synthetic vs Real

Within this section an examination will be conducted on the quality and degree of realism exhibited by synthetic images compared to their real counterparts. The findings are derived from the preceding section's outcomes, where patches are uniformly sampled from the ZeroWaste dataset. The following table shows the mean Intersection over Union (mIoU) results of this comparative analysis.

Unet+resnet50	**ZeroWaste**	**IT+Robotics-V**
Baseline	48,861%	84,684%
Copy & Paste	28,763%	75,521%
Harmonization	26,340%	51,73%
Inpainting	26,062%	60,584%
Inpainting + Copy & Paste	30,666%	75,552%
Inpainting + Harmonization	27,206%	52,931%

Table 4.3: Model comparison results from only synthetic and real images for both datasets.

Table 4.3 displays the outcomes generated through various approaches outlined in Chapter 3. The **Baseline** shows the results obtained exclusively from real images in both datasets, without any augmentation. Conversely, all other approaches, as elucidated in the preceding chapter, show the results derived exclusively from synthetic images in both datasets.
Examining the results summarized in Figure 4.2, which provides a summary overview of the data presented in the previous table, it becomes apparent that a disparity persists between models trained with synthetic images (31% for ZeroWaste and 75% for ITR-V) and those trained with real images (49% for ZeroWaste and 85% for ITR-V). In fact, the findings suggest that while satisfactory outcomes can be achieved, the synthetic images lack the fidelity to reality necessary for competitive results.

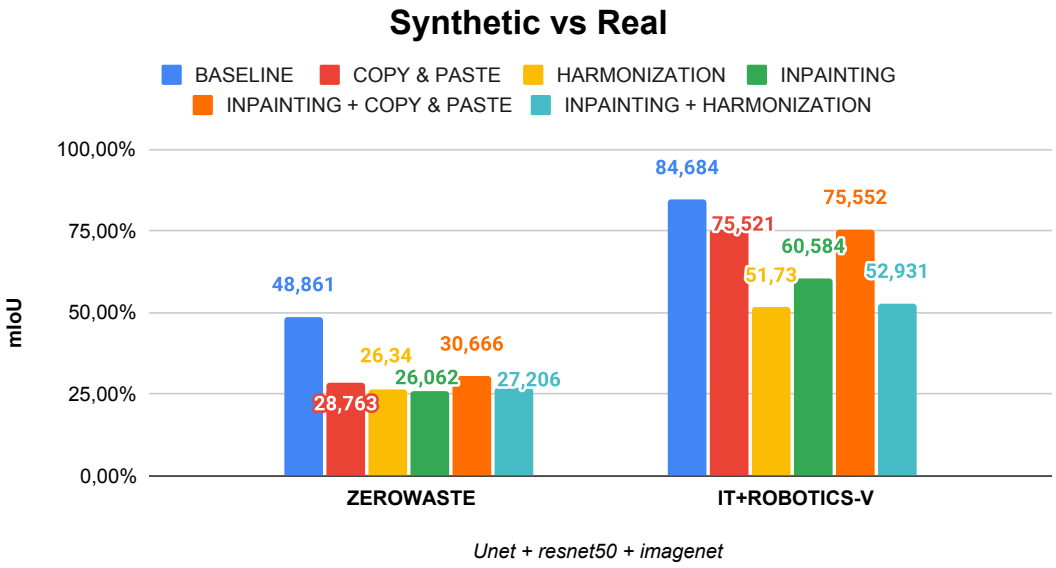

Figure 4.2: A simple graph showing the results of using only synthetic in comparison with real images (Baseline) for both datasets.

4.3 Synthetic + Real: Augmentation Rate

In this section, I will examine the impact of combining synthetic images with real ones. Essentially, I added a percentage of synthetic images to the original training set of both datasets, which is exclusively composed of real scenes. To achieve this, I used only a small portion of the complete datasets, as augmenting entire datasets (with over 7000 patches) is computationally demanding. As you will see in the next section, the results indicate that using only a small portion of the dataset produces similar results compared to using the complete datasets. So, I decided to limit both datasets to 10% of its total size, reducing the number of patches from over 6000 to 600 for the ZeroWaste dataset and from over 600 to 60 patches for the IT+Robotics-V dataset. Sampling for the ZeroWaste dataset follows a uniform distribution, as discussed in the previous section.
Consequently, as just said, I chose a limited number of real images (10% of the entire dataset) and then added a predefined percentage of synthetic images generated, in order to better understand the contribution and the trend shown by the semantic segmentation model as the number of synthetic generated scenes increases. Therefore, as reported in the following tables, I chose to use three

augmentation steps for each dataset, to get a more complete picture of how the approaches, presented in Chapter 3, can contribute to the goal of my thesis.

- Table 4.4 summarizes the results obtained using ***U-Net*** [49] as segmentation model, ***ResNet-50*** [24] as encoder and ***ImageNet*** as weights;
- Table 4.5 summarizes the results obtained using ***DeepLabV3+*** [9] as segmentation model, ***MNet*** [17] as encoder and ***ImageNet*** as weights;
- Table 4.6 summarizes the results obtained using ***DeepLabV3+*** [9] as segmentation model, ***MViT*** [19] as encoder and ***ImageNet*** as weights.

U-Net	**ZeroWaste**			**IT+Robotics-V**		
+ResNet-50	**100%**	**300%**	**450%**	**100%**	**500%**	**1000%**
Baseline		**44,76%**			**80,252%**	
Copy & Paste	45,51%	44,59%	41,66%	80,105%	81,713%	79,265%
Harmonization	43,25%	39,5%	42,22%	80,095%	80,395%	80,161%
Inpainting	43,02%	42,91%	42,18%	78,161%	77,889%	78,656%
Inp. + C&P	44,321%	42,607%	42,889%	80,792%	81,03%	80,72%
Inp. + Harm.	41,747%	41,029%	39,689%	79,273%	80,234%	80,791%

Table 4.4: Results of combining synthetic images, using different augmentation rate, with both reduced real dataset (10%).

DeepLabV3+	**ZeroWaste**			**IT+Robotics-V**		
+MNet	**100%**	**300%**	**450%**	**100%**	**500%**	**1000%**
Baseline		**39,95%**			**76,603%**	
Copy & Paste	41,122%	41,827%	40,716%	76,101%	74,449%	78,136%
Harmonization	37,14%	41,22%	38,34%	78,185%	79,122%	79,408%
Inpainting	41,06%	43,04%	41,06%	77,291%	75,962%	78,656%
Inp. + C&P	43,31%	43,351%	41,08%	77,353%	78,21%	79,984%
Inp. + Harm.	42,268%	39,216%	37,09%	76,615%	79,136%	78,155%

Table 4.5: Results of combining synthetic images, using different augmentation rate, with both reduced real dataset (10%).

DeepLabV3+ +MViT	ZeroWaste			IT+Robotics-V		
	100%	**300%**	**450%**	**100%**	**500%**	**1000%**
Baseline		**44,01%**			**81%**	
Copy & Paste	43,714%	45,66%	45,641%	81,051%	74,955%	80,302%
Harmonization	41,963%	44,137%	41,021%	79,33%	80,27%	80,7%
Inpainting	42,931%	44,827%	46,223%	78,473%	77,986%	79,457%
Inp. + C&P	45,008%	45,247%	43,27%	79,563%	80,478%	80,602%
Inp. + Harm.	41,86%	42,041%	42,902%	79,064%	78,585%	79,361%

Table 4.6: Results of combining synthetic images, using different augmentation rate, with both reduced real dataset (10%).

The **Baseline** indicates the results obtained exclusively with real images from both reduced (10%) datasets, without any augmentation. Consequently, all other approaches, as detailed in the previous chapter, demonstrate the results obtained through the combination of both real and synthetic images.
Upon scrutinizing the results encapsulated in the subsequent figures, which provide a condensed overview of the data presented in the earlier tables, it is clear that the utilization of synthetically generated images proves advantageous for the objectives of this thesis. Additionally, it is noteworthy that the augmentation of synthetic images is not universally beneficial, as the ultimate outcomes can vary significantly based on the quantity of synthetic images employed and the chosen approach.

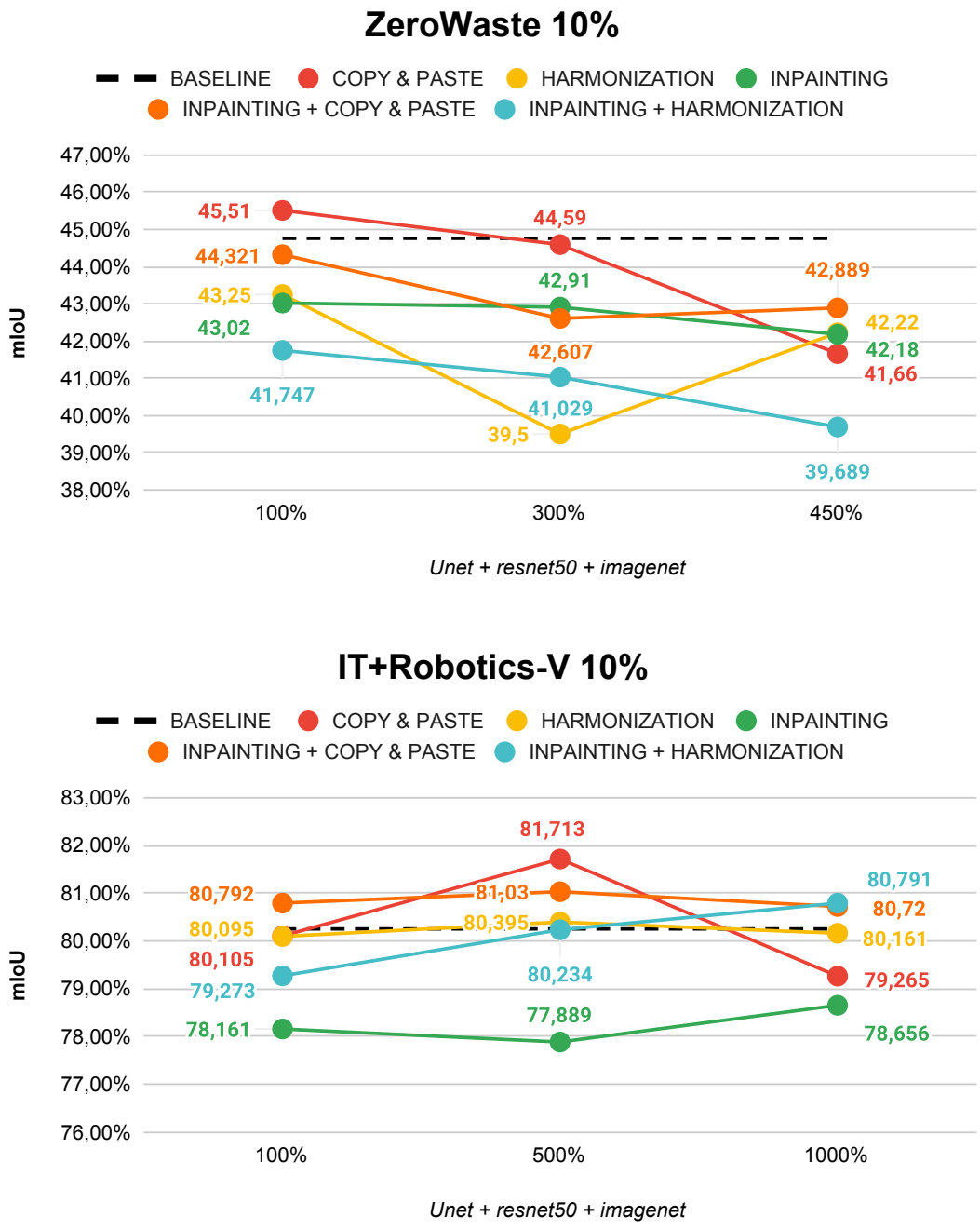

Figure 4.3: A simple graph showing the results of combining synthetic images, using different augmentation rate, with both reduced real dataset (10%).

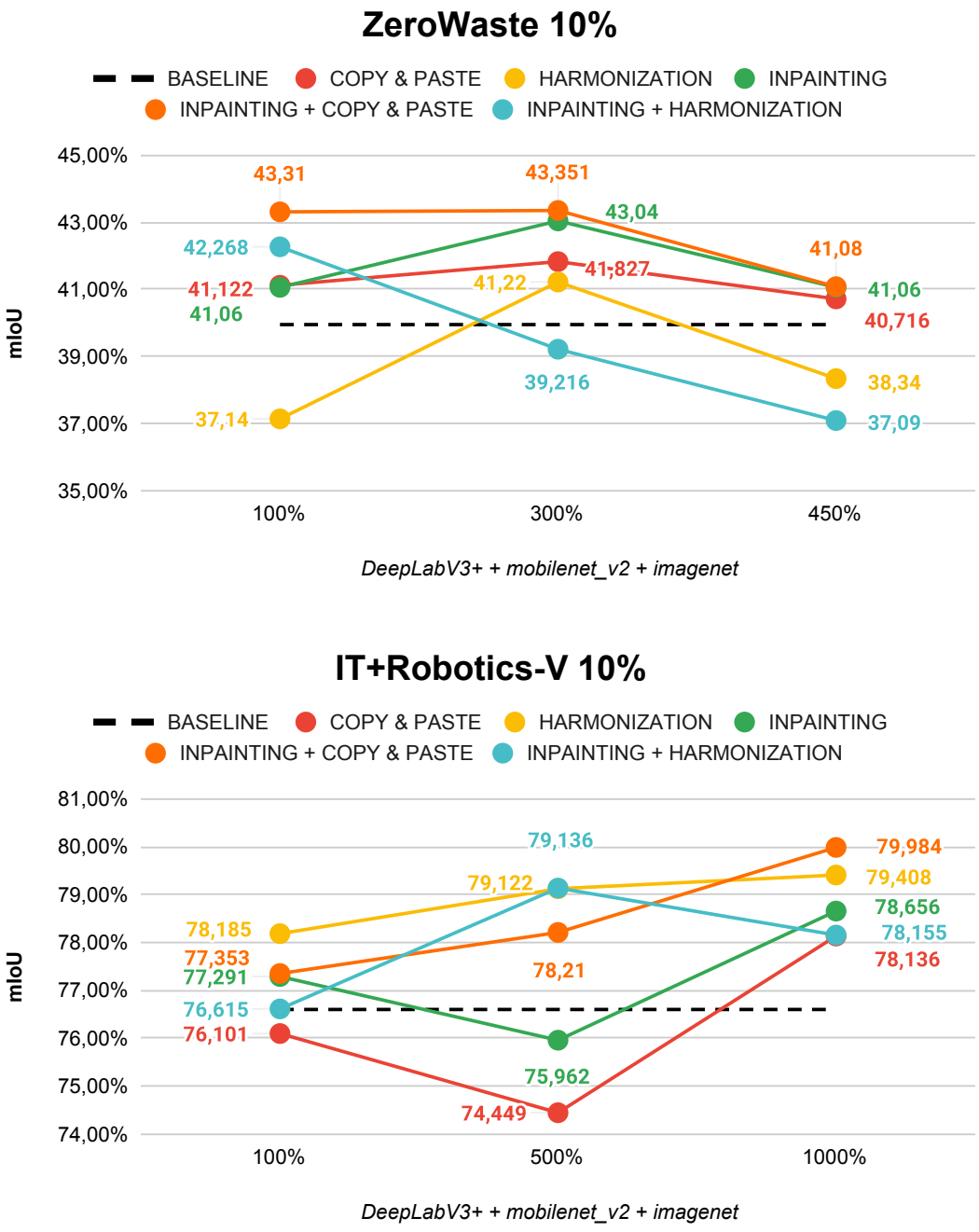

Figure 4.4: A simple graph showing the results of combining synthetic images, using different augmentation rate, with both reduced real dataset (10%).

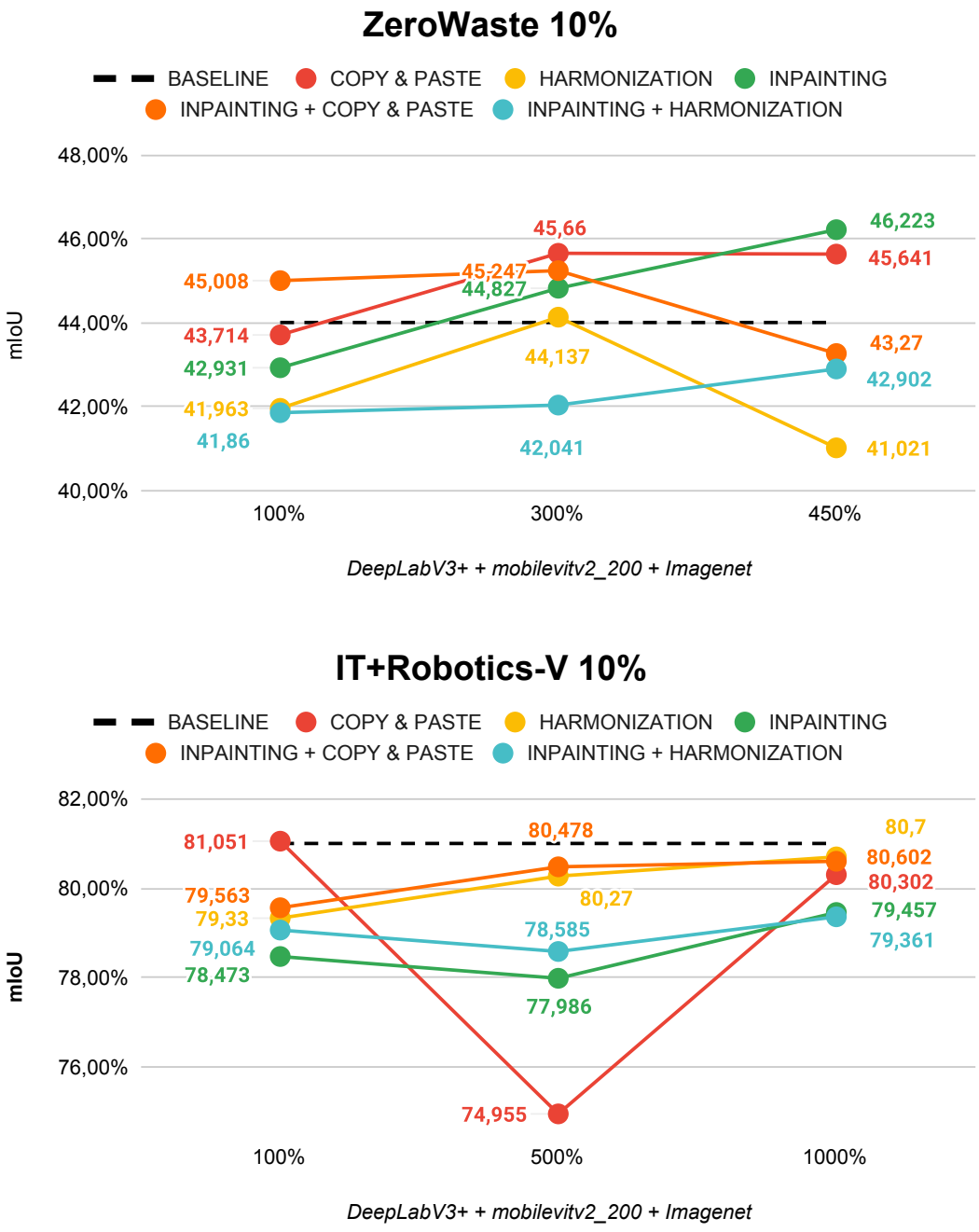

Figure 4.5: A simple graph showing the results of combining synthetic images, using different augmentation rate, with both reduced real dataset (10%).

4.4 Synthetic + Real: Semantic Segmentation

In this section, unlike the previous sections, I will examine the impact of combining synthetic images with real images on both entire datasets. The following tables show the results obtained after the training on both entire datasets, using the three semantic segmentation models discussed in the previous section.

U-Net	**ZeroWaste**		**IT+Robotics-V**	
+ResNet-50	**10%**	**100%**	**10%**	**100%**
Baseline	44,76%	48,861%	80,252%	84,684%
Copy & Paste	45,51%	49,537%	81,713%	81,204%
Harmonization	43,25%	45,43%	80,395%	84,909%
Inpainting	43,02%	52,26%	78,656%	82,882%
Inp. + C&P	44,321%	45,953%	81,032%	83,512%
Inp. + Harm.	41,747%	43,313%	80,791%	82,17%

Table 4.7: Results of combining synthetic images with both real dataset.

DeepLabV3+	**ZeroWaste**		**IT+Robotics-V**	
+MNet	**10%**	**100%**	**10%**	**100%**
Baseline	39,95%	47,28%	76,603%	80,06%
Copy & Paste	41,827%	45,605%	78,136%	82,032%
Harmonization	41,22%	44,02%	79,408%	78,34%
Inpainting	43,02%	48,94%	75,962%	80,133%
Inp. + C&P	43,351%	43,569%	79,9839%	80,708%
Inp. + Harm.	42,267%	43,614%	79,136%	80,288%

Table 4.8: Results of combining synthetic images with both real dataset.

DeepLabV3+	**ZeroWaste**		**IT+Robotics-V**	
+MViT	**10%**	**100%**	**10%**	**100%**
Baseline	44,01%	47,23%	81%	84%
Copy & Paste	45,66%	50,072%	81,051%	80,382%
Harmonization	44,137%	47,583%	80,7%	84,27%
Inpainting	46,223%	50,481%	79,457%	82,53%
Inp. + C&P	45,247%	50,481%	80,602%	82,052%
Inp. + Harm.	42,902%	48,571%	79,361%	81,79%

Table 4.9: Results of combining synthetic images with both real dataset.

The **Baseline** indicates the results obtained exclusively with real images from both datasets, without any augmentation. Consequently, all other methodologies demonstrate the results obtained through the combination of both real and synthetic images.
Examining the results summarized in the following figures, which provide a summary of the data presented in the previous tables, it becomes evident that the utilization of synthetically generated images proves advantageous for the objectives of this thesis. Additionally, the results obtained from both dataset may change a lot with respect to the semantic models used.
Finally, it is interesting to note that the use of synthetic images is not always advantageous, in fact, some results are even lower than the baseline, even if real images are also found in the training set.

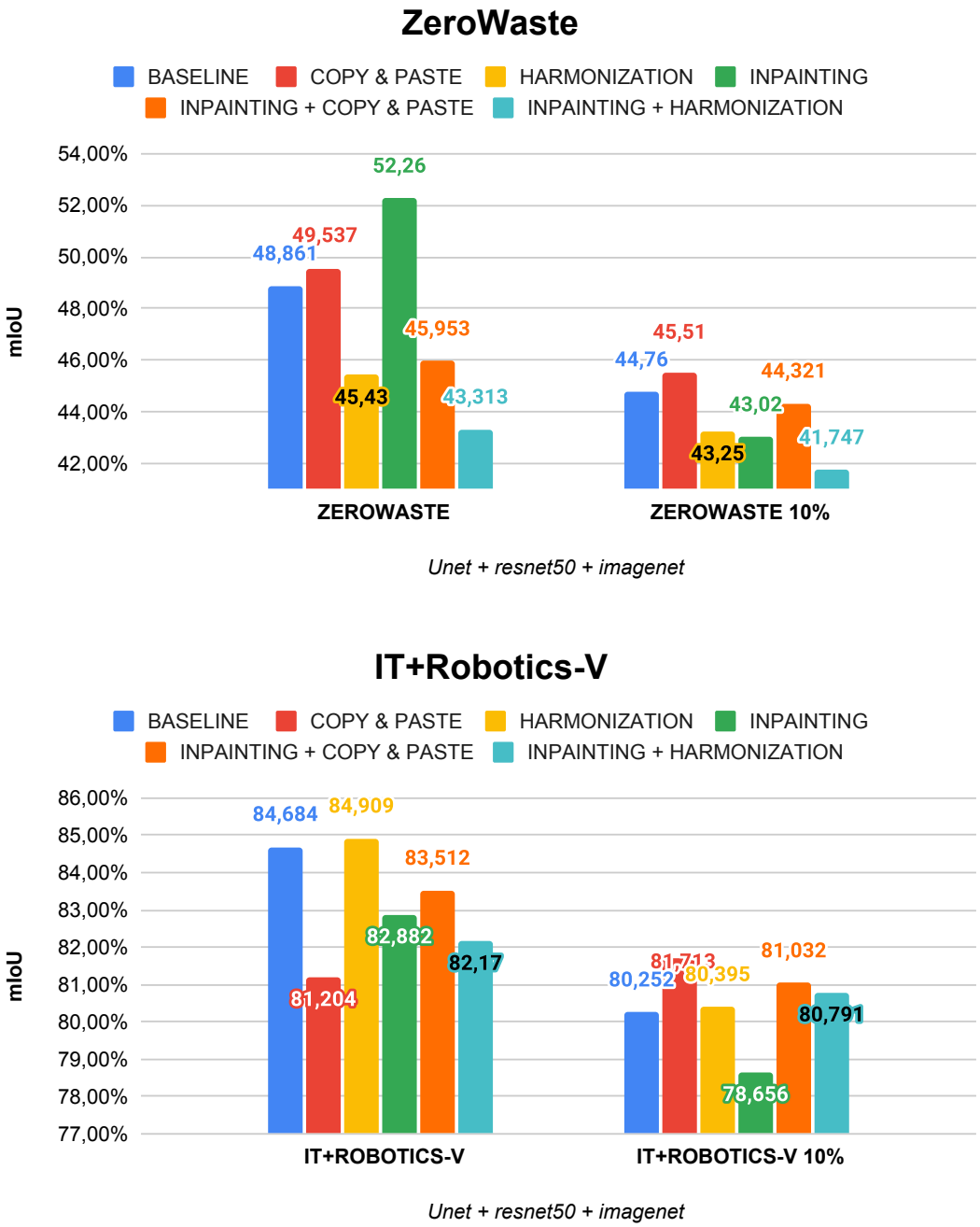

Figure 4.6: A simple graph showing the results of combining synthetic images with both real dataset.

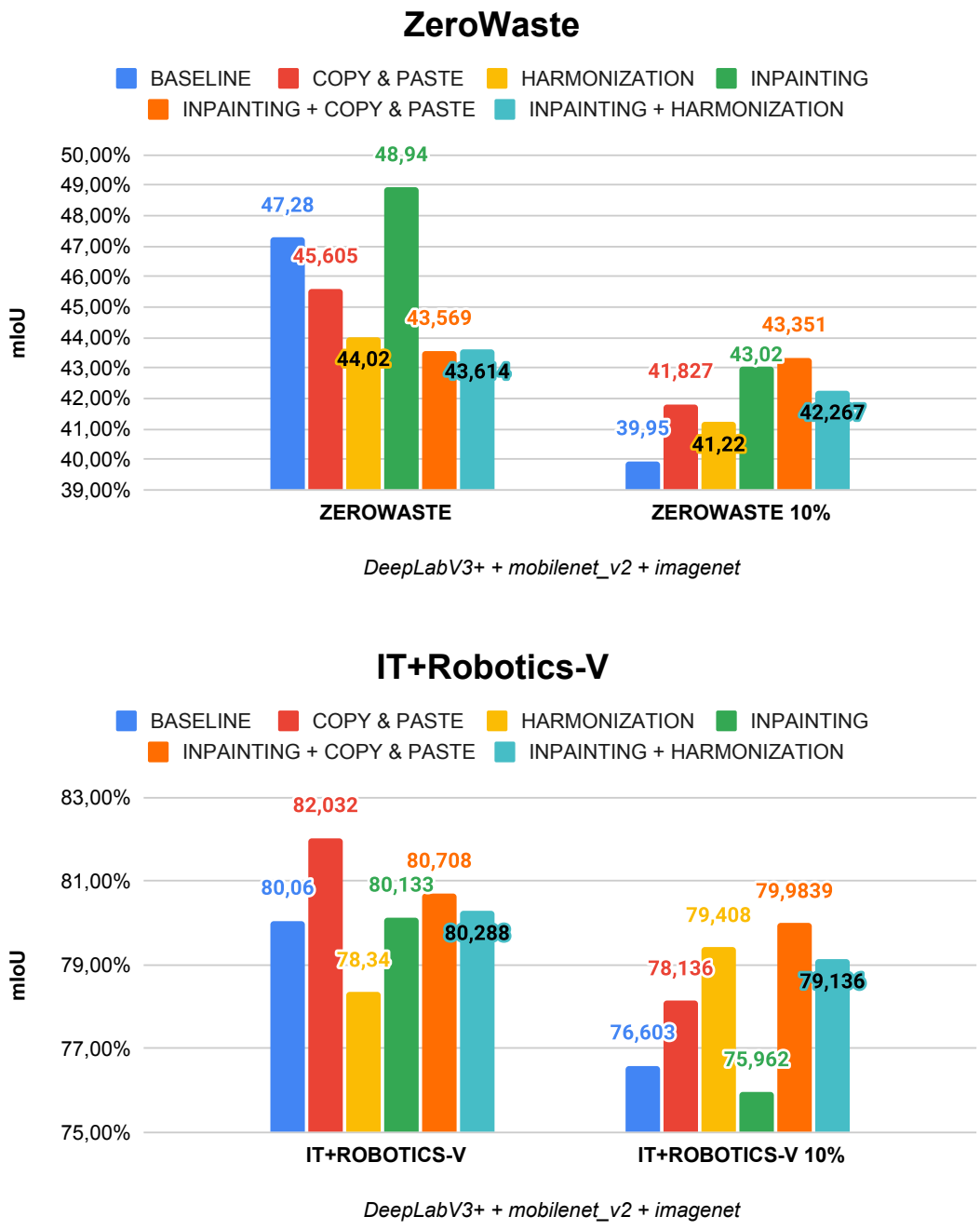

Figure 4.7: A simple graph showing the results of combining synthetic images with both real dataset.

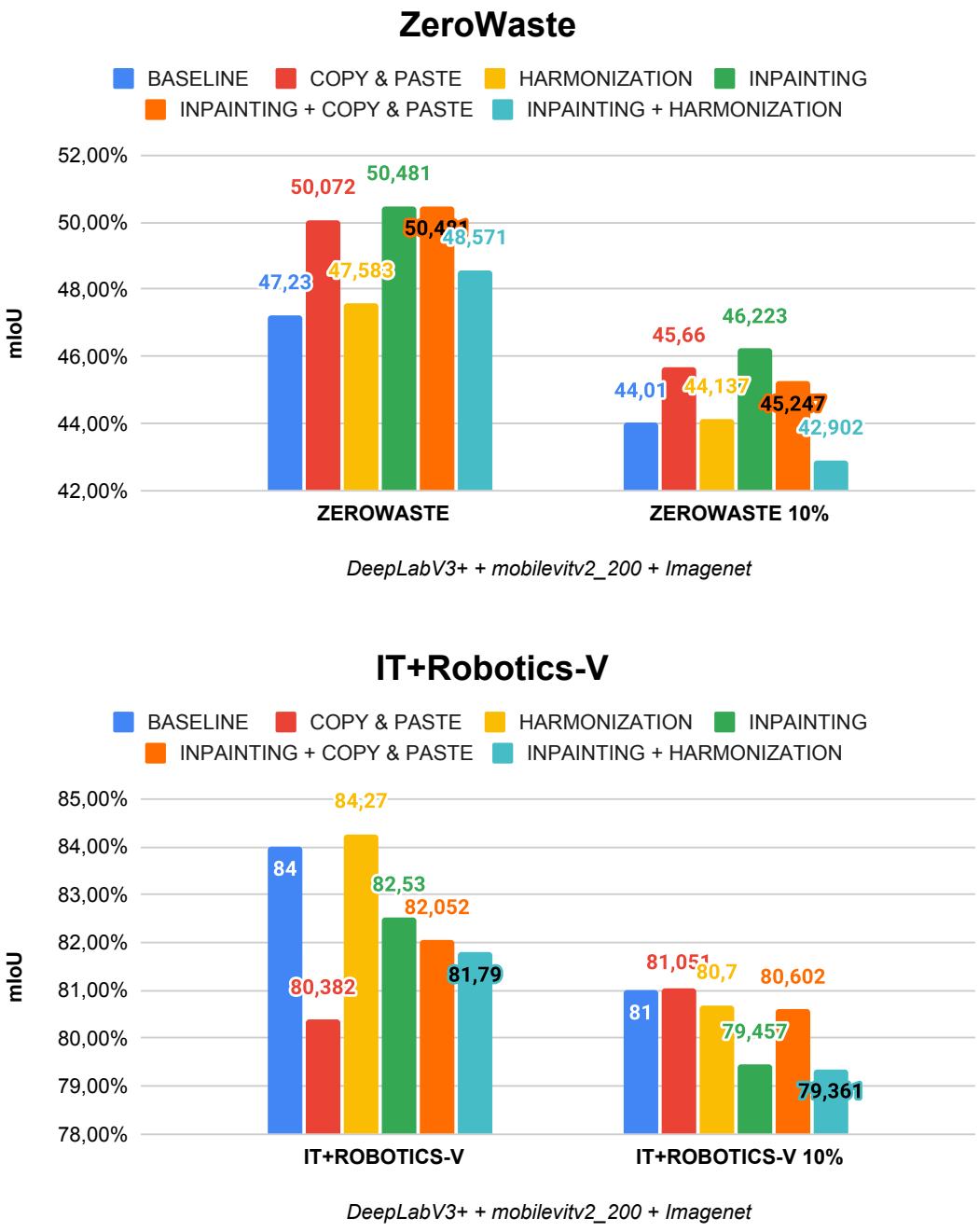

Figure 4.8: A simple graph showing the results of combining synthetic images with both real dataset.

4.5 Analysis of the results

In this section, I analyze and summarize the results of the various experiments, just presented in the previous sections, using all the exemplar-driven augmentation pipelines that I have developed. The extensive experiments conducted provide evidence that exemplar-based augmentation compares favorably with other widely used techniques in the field of data augmentation. In particular, the various generative approaches are particularly useful when the number of images available to train a computer vision model is limited, since it seems easier to introduce new knowledge in a context with limited data available. Nevertheless, without wasting any more time, let's get started:

- **ZeroWaste: Balance vs Imbalance.** The results are unequivocal, as can be seen from Figure 4.1, given that for two of the three approaches used to solve this problem, the mean Intersection over Union (mIoU) obtained from the *copy & paste* and *inpainting* approaches, respectively **49,5%** and **52,3%**, is far better than that obtained using imbalance synthetic generated images (**48,7%** and **47,4%**), but also compared to using only real images (**48,8%**). Only in the case of *image harmonization* the results are worse, with a mIoU of **45,4%** using a more balanced sampling versus a mIoU of **47,9%** with the imbalance sampling. One reason could be the limited new knowledge introduced using the *harmonization* approach, since the main goal of this approach is to obtain a more realistic rather than variegated scene;

- **Synthetic vs Real.** The results in this case too are unequivocal, as can be seen from Figure 4.2, since all the approaches used to solve this problem show a much lower mIoU than the semantic segmentation model trained only on real images. This was to be expected since throughout literature there has always been a persistent gap between synthetic and real images. To get a more complete overview, the gap for the ZeroWaste dataset is around **18%** in favor of the model trained only on real images, considering the best results obtained using the *inpainting + copy & paste* approach, while the gap for the IT+Robotics-V dataset is smaller (**9%**) and also in this case the best results for the trained synthetic model are obtained using the *inpainting + copy & paste* approach;

- **Synthetic + Real: Augmentation Rate.** The main purpose of this experiment was to better understand how synthetically generated images can impact the main objective that this thesis attempts to address. As can be seen from Figures 4.3, 4.4 and 4.5, there is no precise and direct correlation between the percentage number of synthetic scenes used and the consequent results obtained. Looking in more detail at all the mIoU, presented in the previous three figures, it appears that using a larger number of synthetic generated images (**1000%** for the IT+Robotics-V dataset and **450%** for the ZeroWaste dataset) is not always optimal and the trade-off with the resources needed to generate new images is not always advantageous;

- **Synthetic + Real: Semantic Segmentation.** Analyzing Figures 4.6, 4.7 and 4.8, which show a brief summary of the combination of synthetic and real images also for the entire dataset, the considerations just made in the previous experiments remain valid. Moreover, it is clear that the approaches used behave very differently depending on the various situations. In summary, it becomes evident that approaches that seek to favor variety over realism appear to produce overall superior results. This is demonstrated by the fact that in the best case, the inpainting approach outperforms the baseline by more than **4%** for the ZeroWaste dataset, while the copy & paste approach outperforms the baseline, in the best case scenario, by more than **2%** for the IT+Robotics-V dataset. This behavior is entirely predictable, since by adding synthetic images to a dataset that contains a large number of real scenes, it simply seems more important have a greater variety of patches at the expense of high photorealism.

5

Conclusions

In this thesis, I explored the functionalities of text-to-image models within the realm of synthetic imaging and their application in computer vision models. Specifically, I investigated the impact of synthetic images on enhancing the performance of semantic segmentation models. This inquiry is particularly pertinent given the notable progress in text-to-image models and the considerable costs associated with obtaining sufficiently large and high-quality datasets. The widespread availability of certain models, such as Stable Diffusion (SD), has fueled the increasing potential of text-to-image models. Their wide availability to both the scientific community and enthusiasts, has spurred notable initiatives aimed at optimizing and broadening their capabilities and applications.
On the other hand, deep learning models require large amounts of data and depend on large annotated datasets. The creation and upkeep of such vast datasets come with significant costs, making them impractical for many researchers. Consequently, the scientific community has prioritized optimizing deep learning architectures rather than developing approaches to mitigate the expenses associated with acquiring and maintaining large datasets.

Considering the results of Chapter 4, in relation to the research question that sought to determine the extent to which images generated by image-to-image models can enhance the performance of computer vision models, we can deduce the following implications:

- There is still a significant gap between synthetic and real images;
- Exemplar-driven generation techniques represent a viable and competitive approach to data augmentation;
- Exemplar-driven augmentation techniques are especially relevant in datasets where data is scarce or expensive to obtain.

A noteworthy consequence is the reduction in the cost of generating extensive and high-quality datasets. As previously mentioned, the focus within the scientific community has been predominantly on refining architectures, rather than enhancing and augmenting datasets [20] and this emphasis stems from the significant cost associated with acquiring datasets, which are necessary for training deep learning architectures [65].

Additionally, as I exclusively employed zero-shot image generation techniques, it's important to note that the outcomes can vary significantly depending on the chosen approach and semantic model. The image inpainting approach appears to be the most promising, as it can introduce new information for the majority of the patches, although it is not free from flaws. Occasionally, the results deviate significantly from the desired outcome, presenting either a blank image or an incompletely inpainted patch. Furthermore, the effectiveness of this approach differs between datasets. Notably, the image inpainting approach proves to be very effective with the ZeroWaste dataset, but less so with the IT+Robotics-V dataset. This discrepancy can be attributed to the nature of the patches within each dataset. Indeed, ZeroWaste includes a large number of patches (over 6000) but lacks diversity, since the scenes from which the patches are extracted are essentially multiple images of the same moving conveyor belt. In contrast, the IT+Robotics-V dataset contains fewer patches (over 600), but shows significant diversity within the specific class (plastic) domain. For this dataset, more traditional approaches that prioritize patch realism over diversity, such as copy & paste and harmonization, appear to be the most promising.

It's also worth noting that copy & paste, remains a valid approach with almost "zero" associated costs.

In conclusion, as evident from the previous chapter, augmenting a small dataset appears to be more effective than augmenting a larger one. Naturally, the introduction of new knowledge appears to be more straightforward when the number and diversity of patches are constrained.
Nevertheless, I observed a notable disparity between synthetic and real images. Consequently, achieving the performance levels of models trained with an ample supply of real images remains unattainable, regardless of the amount of synthetic images employed.

5.1 Future Works

In general, this study is robust and fruitful, providing substantial insights into the research questions. Nevertheless, akin to any analysis of this scale, there are constraints, areas open for enhancement and new research questions that arise during the work. Major limitations have already been identified in the previous sections, including challenges in consistently obtaining high-quality results, enormous computational requirements and the existence of biases. Accordingly, I propose the following research directions to address these issues and broaden the scope of this thesis:

- **Train or fine-tune the employed models.** Even if training or simply fine-tuning these huge models is not feasible with my computational resources, it might be worth trying this route. In fact, this approach might offer potential solutions to the limitations and issues identified in the context of image inpainting;

- **Improving and optimising exemplar-driven generation techniques.** The authors of *Cross-domain Compositing* [23] and *Paint by Example* [64] suggest enhancements to their respective methods, aiming to enhance both the execution times and the representations acquired in customized image-to-image models;

- **Employing superior and more advanced image-to-image models.** One potential strategy is to replicate the experiments using an upgraded version of Stable Diffusion (SD), such as SDXL Turbo or Imagen [51]. In fact, using these models could bring higher quality images and shorter times to generate new scenes. The rationale for utilizing version 1.4 in the current study is its extensive compatibility with the employed libraries;

- **Explore the possibility of creating entirely new scenes from scratch.** Investigating the potential of generating entirely novel scenes, utilizing either text or an image as a prompt, along with the corresponding segmentation mask or bounding box, allowing DMs to fill in the remaining parts. This change in strategy has the potential to fully exploit the capabilities of DMs.

- **Bias reduction.** Mitigating bias is a critical issue in text-to-image models, as extensively discussed in various papers. Large-scale models, in particular, tend to inherit biases and stereotypes from their training data. I advocate for more comprehensive research to tackle this concern, with a specific focus on exploring the impact of exemplar-driven augmentation techniques. These approaches have the potential to transfer biases from image-to-image models to associated models. Due to the social implications at stake, it is imperative to investigate strategies that can minimize such biases and alleviate potential societal challenges that may arise as a consequence.

References

[1] Rameen Abdal et al. *CLIP2StyleGAN: Unsupervised Extraction of StyleGAN Edit Directions*. 2021. arXiv: 2112.05219 [cs.CV].

[2] United States Environmental Protection Agency. *National Overview: Facts and Figures on Materials, Wastes and Recycling*. URL: https://www.epa.gov/facts-and-figures-about-materials-waste-and-recycling/national-overview-facts-and-figures-materials.

[3] J Alammar. *The Illustrated Stable Diffusion*. 2022. URL: https://jalammar.github.io/illustrated-stable-diffusion/.

[4] Alex Andonian et al. *Paint by Word*. 2023. arXiv: 2103.10951 [cs.CV].

[5] Omri Avrahami, Dani Lischinski, and Ohad Fried. "Blended Diffusion for Text-driven Editing of Natural Images". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022. DOI: 10.1109/cvpr52688.2022.01767. URL: http://dx.doi.org/10.1109/CVPR52688.2022.01767.

[6] Shekoofeh Azizi et al. *Synthetic Data from Diffusion Models Improves ImageNet Classification*. 2023. arXiv: 2304.08466 [cs.CV].

[7] Dina Bashkirova et al. *ZeroWaste Dataset: Towards Deformable Object Segmentation in Cluttered Scenes*. 2022. arXiv: 2106.02740 [cs.CV].

[8] Bor-Chun Chen and Andrew Kae. "Toward Realistic Image Compositing With Adversarial Learning". In: June 2019, pp. 8407–8416. DOI: 10.1109/CVPR.2019.00861.

[9] Liang-Chieh Chen et al. *Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation*. 2018. arXiv: 1802.02611 [cs.CV].

[10] Mehdi Cherti et al. *Reproducible scaling laws for contrastive language-image learning*. 2022. arXiv: 2212.07143 [cs.LG].

[11] Daniel Cohen-Or et al. "Color harmonization". In: *ACM SIGGRAPH 2006 Papers* (2006). URL: https://api.semanticscholar.org/CorpusID:23178.

[12] Wenyan Cong et al. *DoveNet: Deep Image Harmonization via Domain Verification*. 2020. arXiv: 1911.13239 [cs.CV].

[13] Ekin D. Cubuk et al. *AutoAugment: Learning Augmentation Policies from Data*. 2019. arXiv: 1805.09501 [cs.CV].

[14] Ekin D. Cubuk et al. *RandAugment: Practical automated data augmentation with a reduced search space*. 2019. arXiv: 1909.13719 [cs.CV].

[15] Xiaodong Cun and Chi-Man Pun. "Improving the Harmony of the Composite Image by Spatial-Separated Attention Module". In: *IEEE Transactions on Image Processing* 29 (2020), pp. 4759–4771. ISSN: 1941-0042. DOI: 10.1109/tip.2020.2975979. URL: http://dx.doi.org/10.1109/TIP.2020.2975979.

[16] Prafulla Dhariwal and Alex Nichol. *Diffusion Models Beat GANs on Image Synthesis*. 2021. arXiv: 2105.05233 [cs.LG].

[17] Zhangfu Dong et al. *MNet: Rethinking 2D/3D Networks for Anisotropic Medical Image Segmentation*. 2022. arXiv: 2205.04846 [eess.IV].

[18] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV].

[19] Haoqi Fan et al. *Multiscale Vision Transformers*. 2021. arXiv: 2104.11227 [cs.CV].

[20] Golnaz Ghiasi et al. *Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation*. 2021. arXiv: 2012.07177 [cs.CV].

[21] Ian J. Goodfellow et al. *Generative Adversarial Networks*. 2014. arXiv: 1406.2661 [stat.ML].

[22] Zonghui Guo et al. "Intrinsic Image Harmonization". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2021, pp. 16367–16376.

[23] Roy Hachnochi et al. *Cross-domain Compositing with Pretrained Diffusion Models*. 2023. arXiv: 2302.10167 [cs.CV].

[24] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[25] Kaiming He et al. "Mask R-CNN". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 2980–2988. DOI: 10.1109/ICCV.2017.322.

[26] Amir Hertz et al. *Prompt-to-Prompt Image Editing with Cross Attention Control*. 2022. arXiv: 2208.01626 [cs.CV].

[27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. 2020. arXiv: 2006.11239 [cs.LG].

[28] Tiffany Hsu and Steven Lee Myers. *Can We No Longer Believe Anything We See?* URL: https://www.nytimes.com/2023/04/08/business/media/ai-generated-images.html.

[29] Pavel Iakubovskii. *Segmentation Models Pytorch*. https://github.com/qubvel/segmentation_models.pytorch. 2019.

[30] Tim Salimans Ilya Sutskever and Durk Kingma. *Requests for Research 2.0*. URL: https://openai.com/research/requests-for-research-2.

[31] Jiaya Jia et al. "Drag-and-drop pasting". In: *ACM Trans. Graph.* 25.3 (July 2006), pp. 631–637. ISSN: 0730-0301. DOI: 10.1145/1141911.1141934. URL: https://doi.org/10.1145/1141911.1141934.

[32] Jeremy Jordan. *An overview of semantic image segmentation.* URL: https://www.jeremyjordan.me/semantic-segmentation.

[33] Tero Karras et al. *Analyzing and Improving the Image Quality of StyleGAN*. 2020. arXiv: 1912.04958 [cs.CV].

[34] Tero Karras et al. *Elucidating the Design Space of Diffusion-Based Generative Models*. 2022. arXiv: 2206.00364 [cs.CV].

[35] Bahjat Kawar et al. *Imagic: Text-Based Real Image Editing with Diffusion Models*. 2023. arXiv: 2210.09276 [cs.CV].

[36] Silpa Kaza et al. *What a Waste 2.0: A Global Snapshot of Solid Waste Management to 2050*. The World Bank, 2018. DOI: 10.1596/978-1-4648-1329-0. eprint: https://elibrary.worldbank.org/doi/pdf/10.1596/978-1-4648-1329-0. URL: https://elibrary.worldbank.org/doi/abs/10.1596/978-1-4648-1329-0.

[37] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. *DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation*. 2022. arXiv: 2110.02711 [cs.CV].

[38] Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes*. 2022. arXiv: 1312.6114 [stat.ML].

[39] Xihui Liu et al. *More Control for Free! Image Synthesis with Semantic Diffusion Guidance*. 2022. arXiv: 2112.05744 [cs.CV].

[40] Ze Liu et al. *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*. 2021. arXiv: 2103.14030 [cs.CV].

[41] Andreas Lugmayr et al. *RePaint: Inpainting using Denoising Diffusion Probabilistic Models*. 2022. arXiv: 2201.09865 [cs.CV].

[42] Alex Nichol and Prafulla Dhariwal. *Improved Denoising Diffusion Probabilistic Models*. 2021. arXiv: 2102.09672 [cs.LG].

[43] Alex Nichol et al. *GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models*. 2022. arXiv: 2112.10741 [cs.CV].

[44] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV].

[45] Aditya Ramesh et al. *Hierarchical Text-Conditional Image Generation with CLIP Latents*. 2022. arXiv: 2204.06125 [cs.CV].

[46] Aditya Ramesh et al. *Zero-Shot Text-to-Image Generation*. 2021. arXiv: 2102.12092 [cs.CV].

[47] Danilo Jimenez Rezende and Shakir Mohamed. *Variational Inference with Normalizing Flows*. 2016. arXiv: 1505.05770 [stat.ML].

[48] Robin Rombach et al. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2022. arXiv: 2112.10752 [cs.CV].

[49] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV].

[50] Nataniel Ruiz et al. *DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation*. 2023. arXiv: 2208.12242 [cs.CV].

[51] Chitwan Saharia et al. *Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding*. 2022. arXiv: 2205.11487 [cs.CV].

[52] Mert Bulent Sariyildiz et al. *Fake it till you make it: Learning transferable representations from synthetic ImageNet clones*. 2023. arXiv: 2212.08420 [cs.CV].

[53] Jascha Sohl-Dickstein et al. *Deep Unsupervised Learning using Nonequilibrium Thermodynamics*. 2015. arXiv: 1503.03585 [cs.LG].

[54] Jiaming Song, Chenlin Meng, and Stefano Ermon. *Denoising Diffusion Implicit Models*. 2022. arXiv: 2010.02502 [cs.LG].

[55] Kalyan Sunkavalli et al. "Multi-scale image harmonization". In: *ACM SIGGRAPH 2010 papers* (2010). URL: https://api.semanticscholar.org/CorpusID:16985983.

[56] Michael Tao, Micah Johnson, and Sylvain Paris. "Error-Tolerant Image Compositing". In: vol. 103. Aug. 2010, pp. 31–44. ISBN: 978-3-642-15548-2. DOI: 10.1007/978-3-642-15549-9_3.

[57] Brandon Trabucco et al. *Effective Data Augmentation With Diffusion Models*. 2023. arXiv: 2302.07944 [cs.CV].

[58] Yi-Hsuan Tsai et al. *Deep Image Harmonization*. 2017. arXiv: 1703.00069 [cs.CV].

[59] Lilian Weng. "What are diffusion models?" In: *lilianweng.github.io* (July 2021). URL: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/.

[60] Wikipedia contributors. *Stable Diffusion — Wikipedia, The Free Encyclopedia*. [Online; accessed 30-January-2024]. 2024. URL: https://en.wikipedia.org/w/index.php?title=Stable_Diffusion&oldid=1199962016.

[61] Weihao Xia et al. *TediGAN: Text-Guided Diverse Face Image Generation and Manipulation*. 2021. arXiv: 2012.03308 [cs.CV].

[62] Shaoan Xie et al. *SmartBrush: Text and Shape Guided Object Inpainting with Diffusion Model*. 2022. arXiv: 2212.05034 [cs.CV].

[63] Ben Xue et al. *DCCF: Deep Comprehensible Color Filter Learning Framework for High-Resolution Image Harmonization*. 2022. arXiv: 2207.04788 [cs.CV].

[64] Binxin Yang et al. *Paint by Example: Exemplar-based Image Editing with Diffusion Models*. 2022. arXiv: 2211.13227 [cs.CV].

[65] Suorong Yang et al. *Image Data Augmentation for Deep Learning: A Survey*. 2023. arXiv: 2204.08610 [cs.CV].