



UNIVERSIDADE FEDERAL DO RIO GRANDE - FURG
CENTRO DE CIÊNCIAS COMPUTACIONAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL
CURSO DE MESTRADO EM MODELAGEM COMPUTACIONAL

Dissertação de Mestrado

Synthetic Data Generation Tools for Computer Vision using Diffusion Models

José David García Ramos

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal do Rio Grande - FURG, como requisito parcial para a obtenção do grau de Mestre em Modelagem Computacional

Orientador: Prof. Dr. Paulo Lilles Jorge Drews Junior
Co-orientador: Prof. Dr. Felipe Gomes de Oliveira

Rio Grande, 2026

R175s

Ramos, José David García

Synthetic Data Generation Tools for Computer Vision using Diffusion Models /José David García Ramos. -- 2026.

85 f.

Dissertação (Mestrado) – Universidade Federal do Rio Grande – Programa de Pós-Graduação em Modelagem Computacional, 2026.

Orientador: Dr. Paulo Lilles Jorge Drews Junior

Co-orientador: Dr. Felipe Gomes de Oliveira

1. Modelagem Computacional. 2. Geração de dados sintéticos. 3. Modelos de difusão. 4. Aprendizado Zero-Shot. 5. Orientação latente. 6. Segmentação semântica. I. Drews Junior, Paulo Lilles Jorge. II. Oliveira, Felipe Gomes de. III. Título.

CDU 004.414.23

Catalogação na fonte: Bibliotecária Brenda Sequeira CRB10/1979





Dissertação de Mestrado

Synthetic Data Generation Tools for Computer Vision using Diffusion Models

José David García Ramos

Banca examinadora:

Prof. Dr. Thiago L. T. da Silveira

Prof. Dr. Michel Melo da Silva

Prof. Dr. Rodrigo de Bem

ABSTRACT

GARCÍA RAMOS, José David. **Automated Synthetic Data Generation Pipelines for Computer Vision using Diffusion Models**. 2026. 87 f. Dissertação (Mestrado) – Programa de Pós-Graduação em Modelagem Computacional. Universidade Federal do Rio Grande - FURG, Rio Grande.

Data scarcity constitutes a critical bottleneck in deep learning generalization, particularly in specialized domains such as agriculture and underwater exploration. While Diffusion Models offer high-fidelity synthesis, standard approaches often rely on computationally expensive fine-tuning or manual prompt engineering. This thesis addresses these limitations by developing fully automated, training-free generative pipelines. Initially, an efficient image-generation methodology that combines IP-Adapter and SeeCoder was validated on the CottonWeed12Det dataset. This approach achieved a CMMD score of 1.317 and improved the stability of YOLOv8 detection without retraining, outperforming traditional augmentation strategies. Building on these foundations, the research introduced ZEDD-G, a novel zero-shot classification framework that leverages multi-image latent guidance. ZEDD-G demonstrated the ability to boost ResNet-50 accuracy by over 30% across diverse datasets, from natural images to medical X-rays, by directly manipulating latent features without fine-tuning. Finally, to address the challenge of dense prediction, a spatially aware pipeline integrating ControlNet and Vision-Language Models (BLIP) was developed for underwater semantic segmentation. This method demonstrated that detailed semantic conditioning is essential for reducing domain gaps, yielding a 4.93% mIoU improvement on YOLOv11-Seg. Collectively, these contributions establish a robust suite of tools that automate the creation of high-quality, labeled synthetic data, significantly enhancing model training in data-scarce environments while minimizing computational overhead.

Keywords: Synthetic Data Generation, Diffusion Models, Zero-Shot Learning, Latent Guidance, Semantic Segmentation.

ABSTRACT

GARCÍA RAMOS, José David. **Pipelines Automatizados de Geração de Dados Sintéticos para Visão Computacional usando Modelos de Difusão**. 2026. 87 f. Dissertação (Mestrado) – Programa de Pós-Graduação em Modelagem Computacional. Universidade Federal do Rio Grande - FURG, Rio Grande.

A escassez de dados constitui um gargalo crítico na generalização do aprendizado profundo, particularmente em domínios especializados, como agricultura e exploração subaquática. Embora os Modelos de Difusão ofereçam síntese de alta fidelidade, as abordagens padrão frequentemente dependem de ajustes finos (fine-tuning) computacionalmente caros ou de engenharia manual de prompts. Esta dissertação aborda essas limitações desenvolvendo pipelines generativos totalmente automatizados e livres de treinamento. Inicialmente, uma metodologia eficiente de geração de imagens que combina IP-Adapter e SeeCoder foi validada no conjunto de dados CottonWeed12Det. Essa abordagem alcançou uma pontuação CMMD de 1.317 e melhorou a estabilidade da detecção do YOLOv8 sem a necessidade de re-treinamento, superando as estratégias tradicionais de aumento de dados. Com base nessas fundações, a pesquisa introduziu o ZEDD-G, um novo framework de classificação zero-shot que aproveita a orientação latente de múltiplas imagens. O ZEDD-G demonstrou a capacidade de aumentar a precisão da ResNet-50 em mais de 30% em diversos conjuntos de dados, desde imagens naturais até radiografias médicas, manipulando diretamente as características latentes sem ajustes finos. Por fim, para enfrentar o desafio da predição densa, um pipeline espacialmente consciente integrando ControlNet e Modelos de Visão-Linguagem (BLIP) foi desenvolvido para a segmentação semântica subaquática. Esse método demonstrou que o condicionamento semântico detalhado é essencial para reduzir as lacunas de domínio (domain gaps), resultando em uma melhoria de 4.93% no mIoU na YOLOv11-Seg. Coletivamente, essas contribuições estabelecem um conjunto robusto de ferramentas que automatizam a criação de dados sintéticos rotulados de alta qualidade, melhorando significativamente o treinamento de modelos em ambientes com escassez de dados, ao mesmo tempo em que minimizam a sobrecarga computacional.

: Geração de Dados Sintéticos, Modelos de Difusão, Aprendizado Zero-Shot, Orientação Latente, Segmentação Semântica.

LIST OF FIGURES

Figure 1 Visual representation of the forward and reverse diffusion processes. The forward trajectory systematically injects Gaussian noise, while the learned reverse trajectory reconstructs the data distribution. Image reproduced from (Ghojogh and Ghodsi, 2024). 17

Figure 2 Detailed view of the feature injection process through cross-attention layers. The external condition (text or visual prompt) is projected into Key and Value vectors, which interact with the U-Net's internal Query vectors to guide the image synthesis. Image adapted from (Ye et al., 2023). 22

Figure 3 Illustration of the prompt engineering challenge. Minor variations in textual descriptions can lead to significant and unpredictable changes in the generated output, highlighting the need for robust visual prompting methods. 35

Figure 4 Combined use of IP-Adapter and SeeCoder with DINOv2-clustered guide images for enhanced synthetic data generation. 42

Figure 5 CottonWeedYolo dataset original class distribution (Dang et al., 2023) 44

Figure 6 Qualitative Analysis of Image Generation Methods for CottonWeedDet12 Species (*indicates prompt-based methods). 47

Figure 7 mAP values obtained in the first five experiments, analyzing the impact of different training configurations on the performance of the YOLOv8 model for detecting 14 weed classes. 48

Figure 8 Comparison of mAP between Experiments 6 and 7, highlighting the impact of adding synthetic data on the detection accuracy of the 14 weed classes with the scarcity of real data. 49

Figure 9 Overview of the ZEDD-G framework. Image clusters are formed using DINOv2. A primary image is then fed into a SeeCoder to extract visual prompts that guide the modification of latent representations from other images. These are injected into the Diffusion Model (DM) to generate new samples. 52

Figure 10 Visual demonstration of our multi-image latent guidance method. The figure illustrates the impact of our two key hyperparameters: Block Size (BS) and Shuffle Probability (SP). Block size and shuffle probability control the trade-off between blending features from multiple images and preserving the primary image's structure. 54

Figure 11 Qualitative evaluation of synthetic images generated by ZEDD-G across various datasets from the benchmark. Examples shown include the 'tuberculosis' class from Shenzhen TB; 'pyramid,' 'pizza,' and 'jug' from Caltech101; 'Acura ILX' and 'Chevy HHR' from Stanford Cars; and 'bulldog' and 'Persian' from Oxford-IIIT Pets. For each set, the primary guide image is highlighted in pink, with three secondary images providing additional guidance. 57

Figure 12 **Qualitative comparison of synthetic samples.** The figure displays results across three challenging scenarios: (Top) Diver and marine life, (Middle) Close-up marine fauna, and (Bottom) Sea turtle with heavy occlusion. Columns correspond to the generation strategies: v1 (No Prompt), v2 (SeeCoder), v3 (Simple Prompt), and v4 (Complex/BLIP Prompt). 65

Figure 13 Performance trajectory of different augmentation strategies. Version 4 (Complex Prompt) highlights a clear "sweet spot"at the $\times 4$ factor, outperforming simpler prompting methods. 66

Figure 14 Impact of the v4 augmentation strategy on mIoU across different architectures. The red bars indicate the performance after augmentation. 67

Figure 15 Impact of model refinement on per-class performance. The matrix displays the normalized recall for the optimal improved model (x4, v4), with the performance delta relative to the baseline overlaid. Green annotations (+) indicate increased diagonal accuracy, while red annotations (-) denote a reduction in misclassifications. 68

LIST OF TABLES

Table 1 Taxonomy of Control Mechanisms and Optimization Techniques in Diffusion Models. This table contrasts the guidance method with its computational prerequisites. 38

Table 2 Overview of Data Generation Strategies using Diffusion Models. The table highlights the specific application domains and the limitations of current methodologies regarding generalization. 39

Table 3 Quantitative evaluation of synthetic image fidelity using CMMD across baselines and ablation components. 45

Table 4 Comparison of data augmentation methods based on classification accuracy across multiple datasets (ResNet-50 trained from scratch and pre-trained). The benchmark is adapted from DreamDA (Fu et al., 2024); unlike our zero-shot ZEDD-G, all other methods require dataset-specific fine-tuning and prompt engineering. 56

Table 5 Evolution of mIoU (%) on YOLOv11-Seg across varying augmentation factors for different prompting strategies. The v4 strategy achieves optimal performance at $\times 4$. 64

Table 6 Comparative benchmark of segmentation architectures before and after applying the v4 augmentation strategy. YOLOv11 offers the best trade-off between performance gain and training efficiency. 66

LIST OF ABBREVIATIONS AND ACRONYMS

AI Artificial Intelligence

AP Average Precision

BLIP Bootstrapping Language-Image Pre-training

CGAN Conditional Generative Adversarial Network

CLIP Contrastive Language-Image Pre-training

CMMD CLIP-MMD Distance

COW Cyclic One-Way Diffusion

DDIM Denoising Diffusion Implicit Model

DDPM Denoising Diffusion Probabilistic Model

DM Diffusion Model

FID Frechet Inception Distance

GAN Generative Adversarial Network

GDA Generalized Diffusion Adaptation

GFLOP Giga Floating Point Operations

IoU Intersection over Union

KNN k-Nearest Neighbors

LDM Latent Diffusion Model

LLM Large Language Model

mAP mean Average Precision

mIoU mean Intersection over Union

MMD Maximum Mean Discrepancy

ODE Ordinary Differential Equation

OOD Out-Of-Distribution

QKV Query, Key, and Value

SAM Segment Anything Model

SDE Stochastic Differential Equation

SOTA State-of-the-Art

VLB Variational Lower Bound

VLM Vision-Language Model

CONTENTS

1 Introduction **13**

2 Theoretical Framework of Denoising Diffusion Models **16**
2.1 Diffusion-based Learning 16
2.2 Latent Diffusion Models (LDMs) 21
2.3 Conditioning and Feature Injection Mechanisms 21
2.3.1 Cross-Attention and the QKV Formulation 21
2.3.2 Visual Prompting and Decoupled Cross-Attention 22
2.4 Foundation Models for Advanced Control 22
2.4.1 Spatial Control via ControlNet 23
2.4.2 Self-Supervised Clustering with DINOv2 23

3 Synthetic Data Generation: Methods, Applications, and Challenges **24**
3.1 Data Scarcity in Computer Vision Tasks 24
3.1.1 Traditional Data Augmentation 25
3.1.2 Synthetic Data and Graphics Engines 25
3.1.3 Generative Adversarial Networks 26
3.2 Diffusion Models 27
3.2.1 Control Models 28
3.2.2 Adapters for Guidance 28
3.2.3 Attention Guide 29
3.2.4 Latent Space Guide 30
3.3 Evaluation Metrics for Synthetic Data in Generative Models 31
3.4 Application and Challenges in Data Generation 31
3.4.1 Diffusion Models as Data Augmentation 31
3.4.2 Diffusion Models as Synthetic Data Generators 32
3.4.3 Computational Costs 33
3.4.4 Guide Challenge (Prompt Engineering) 34
3.5 Automatic Synthetic Data Generation 35
3.5.1 Automatic Labels and Masks Generation 36
3.5.2 Tools and Frameworks 36
3.6 Evaluation Metrics for Synthetic Data 37
3.7 Comprehensive Summary and Research Gap 38
3.7.1 Trade-offs in Control and Efficiency 38
3.7.2 Fragmentation in Data Generation Strategies 39
3.7.3 Conclusion: The Need for Automated Efficiency 39

4 Training-Free Synthetic Data Strategies for Agricultural Object Detection 40
4.1 Synthetic Images Generation by DreamBooth Model . . . 40
4.1.1 Clustering based on the DINOv2 Algorithm . . . 41
4.1.2 Enhanced Synthetic Data Generation . . . 41
4.2 Experiments . . . 42
4.3 Synthetic data setup . . . 43
4.3.1 Experimental Setup . . . 43
4.3.2 Synthetic data evaluation . . . 45
4.3.3 Evaluation of the Dataset Formation for Detection . . . 46
4.4 Discussion and Limitations . . . 49

5 Automated Zero-Shot Augmentation for Robust Image Classification 51
5.1 ZEDD-G Method . . . 51
5.1.1 Clustering with DINOv2 . . . 51
5.1.2 Visual Feature Extraction with the SeeCoder . . . 52
5.1.3 Multi-Image Latent Guidance . . . 52
5.1.4 Zero-Shot Diffusion-based Data Generation . . . 54
5.1.5 Benchmark . . . 55
5.2 Experimental Results . . . 55
5.3 Discussion . . . 58

6 UW-DataGen: A Semantic-Aware Synthetic Data Pipeline for Underwater Segmentation 59
6.1 Methodology . . . 59
6.1.1 Guidance Evaluation . . . 61
6.1.2 Generation Volume Trade-off . . . 62
6.1.3 Final Pipeline . . . 62
6.2 Results . . . 63
6.2.1 Qualitative Evaluation of Generated Samples . . . 63
6.2.2 Impact of Prompt Conditioning and Data Volume . . . 64
6.2.3 Cross-Architecture Generalization . . . 66
6.2.4 Error Analysis . . . 67
6.3 Discussion . . . 67

7 Conclusions and Future Works 69
7.1 General Conclusions . . . 69
7.2 Future Works . . . 70

Bibliography 72

ANEXOS

A Qualitative Analysis of Generative Failures in Weed Synthesis 82

B Visual Limitations of Multi-Image Latent Guidance in ZEDD-G 84

C Qualitative Analysis of Artifacts in Underwater Segmentation Generation 86

1 INTRODUCTION

Deep learning's progress in computer vision is often constrained by the scarcity and sparsity of large-scale datasets. This is a critical factor that hinders complex models' ability to achieve robust generalization (Aljohani and Alharbe, 2022; Alaa et al., 2022; Feng et al., 2022), particularly in domains where collecting thousands of annotated samples is physically or economically prohibitive, such as agriculture and underwater exploration. Although synthetic data derived from Diffusion Models (DMs) offers high-fidelity solutions to this data scarcity, the practical application of these generative tools remains hindered by excessive training times and computational demands.

This research is therefore driven by the need to address two core challenges: the fundamental lack of diverse data for specialized computer vision tasks, and the absence of automated, accessible frameworks capable of generating such data without incurring prohibitive computational costs.

Current State-of-the-Art (SOTA) methods for synthetic data generation face a severe operational bottleneck. They force researchers to choose between two highly inefficient paths: extensive manual intervention, involving complex prompt engineering and manual pixel-level annotation, or resource-intensive model fine-tuning (Saragih et al., 2024). Standard techniques like DreamBooth (Ruiz et al., 2023), for example, demand high-end GPUs that often require 16GB to 24GB of VRAM and significant storage for every new concept. Furthermore, this fine-tuning process can take several hours per class, making it completely unviable for practical, resource-constrained environments or datasets with dozens of categories.

To overcome this barrier, this thesis proposes a comprehensive suite of automated, diffusion-based tools designed to democratize access to high-quality synthetic data. Instead of relying on a single, monolithic approach, we develop progressive solutions tailored to specific computer vision tasks. For classification and object detection, we introduce highly efficient, training-free frameworks that completely bypass the need to retrain pre-trained DMs, reducing the generation overhead to mere seconds per image on standard consumer hardware. For more complex, dense prediction tasks, such as semantic segmentation, we propose a spatially aware pipeline. This approach leverages efficient

conditioning models (ControlNet) and vision-language networks (BLIP) to automatically generate perfectly paired images and pixel-accurate labels.

Across all these proposed methods, the primary objective is to minimize the consumption of critical resources like processing time, storage, and VRAM, while fully automating the entire pipeline from image synthesis to label generation.

To achieve this, the specific objectives of this thesis are defined as follows:

- To systematically evaluate the trade-off between computationally expensive fine-tuning methods and efficient, training-free architectures in the context of agricultural object detection (specifically, weed detection in crop environments).
- To develop a novel *Zero-Shot* generation framework that leverages multi-image latent guidance to create diverse classification datasets without requiring any model retraining.
- To implement a spatially-aware generation pipeline that integrates ControlNet and Vision-Language Models (VLMs) to automate the creation of precise, pixel-level masks for semantic segmentation in challenging domains, such as underwater imagery.
- To validate the proposed tools by benchmarking their performance against established augmentation techniques on standard datasets, demonstrating that low-resource, automated pipelines can achieve SOTA results while minimizing computational overhead.

This document is organized to reflect the logical evolution of the research, progressing from baseline comparisons to advanced latent manipulation and, finally, to dense prediction tasks. The structure correlates directly with the scientific contributions made during this Master's program:

- **Chapter 2: Theoretical Foundation** lays the groundwork by explaining the fundamental principles of Diffusion-based Learning, including the forward and reverse processes, and key architectures such as Denoising Diffusion Probabilistic Models (DDPMs) and Latent Diffusion Models (LDMs).
- **Chapter 3: State of the Art** reviews the existing literature on data scarcity, positioning DMs against Generative Adversarial Networks (GANs). It critically analyzes current gaps in the field, specifically detailing the high computational costs of fine-tuning and the lack of automated labeling tools for scientific data.
- **Chapter 4: Synthetic Data Pipeline for Cotton Weed Detection** presents the initial investigation into training-free generation. It compares fine-tuning (DreamBooth) against an automated pipeline using IP-Adapter and SeeCoder for agricultural weed detection. This chapter highlights the efficacy of training-free methods

and corresponds to the work presented at the **Brazilian Conference on Robotics (CROS)** (Ramos et al., 2025b).

- **Chapter 5: Automatic Classification Generation (ZEDD-G)** advances the methodology by introducing a novel zero-shot framework. It details the "Multi-Image Latent Guidance" technique, which allows for the generation of diverse classification data without any training. This work benchmarks against SOTA methods like DreamDA and corresponds to the contributions published in the **Conference on Graphics, Patterns and Images (SIBGRAPI)** (Ramos et al., 2025a).

- **Chapter 6: Segmentation Generation** addresses the most complex task: dense prediction. It introduces a pipeline that combines ControlNet with detailed semantic prompting (BLIP) to generate spatially accurate synthetic data for underwater environments. This chapter presents the most advanced iteration of the proposed tools, submitted to the journal **Machine Vision and Applications**.

- **Chapter 7: Conclusions and Future Works** summarizes the general findings, discussing the transition from volume-based augmentation to semantic-based generation, and proposes future directions such as Uncertainty-Driven Active Learning.

2 THEORETICAL FRAMEWORK OF DENOISING DIFFUSION MODELS

This chapter lays the groundwork for understanding the core generative method utilized in this research (DMs). It delves into the fundamental principles of these models, which have demonstrated exceptional ability to synthesize high-dimensional data, particularly images. The chapter explains the two primary phases of their operation: the forward process, in which data are gradually converted into noise, and the reverse process, in which the model learns to denoise and generate new data samples.

2.1 Diffusion-based Learning

DMs have emerged as a promising class of deep generative models (Song et al., 2020), demonstrating remarkable capacity in synthesizing high-dimensional data, especially in the image domain (Song et al., 2023). Inspired by principles of thermodynamics and statistical physics, these models are based on diffusion, a natural process in which particles move from regions of high concentration to areas of low concentration until equilibrium is reached (Dhariwal and Nichol, 2021). Adapting this concept to generative modeling, DMs gradually transforms a complex data distribution into a simple and tractable distribution, typically an isotropic Gaussian, through the sequential addition of noise. The model then learns to reverse this process to generate new samples from pure noise (Song et al., 2020).

The operation of DMs is generally decomposed into two main steps: the *forward process* and the *reverse process*. A visual representation of this dual-phase architecture is shown in Figure 1, which illustrates how the data structure is degraded and subsequently restored.

Forward Process

The forward process is fixed and non-parametric. It consists of gradually degrading an image $x_0 \sim q(x_0)$, coming from the real data distribution, over a sequence of T steps. In each step t, Gaussian noise is added:

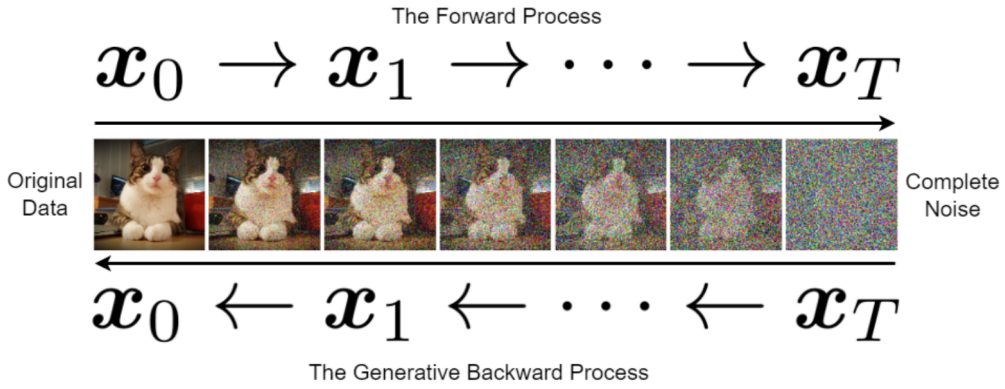


Figure 1: Visual representation of the forward and reverse diffusion processes. The forward trajectory systematically injects Gaussian noise, while the learned reverse trajectory reconstructs the data distribution. Image reproduced from (Ghojogh and Ghodsi, 2024).

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I), \tag{1}$$

where $\{\beta_t\}_{t=1}^T$ is known as the variance schedule, composed of small positive values. This process forms a Markov chain and, with a sufficiently large number of steps and appropriate choice of β_t, the final distribution x_T approaches a standard Gaussian $\mathcal{N}(0, I)$ (Song et al., 2020).

An important property of the forward process is that we can sample x_t directly from x_0 at any step t:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I), \tag{2}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$. This expression is fundamental for the efficient training of DMs.

Reverse Process

The reverse process is the learned component of the model and aims to reconstruct x_0 from $x_T \sim \mathcal{N}(0, I)$, iteratively removing noise. This process is also modeled as a Markov chain, where each transition $p_\theta(x_{t-1}|x_t)$ is parameterized by a neural network ϵ with weights θ:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \tag{3}$$

DDPMs, introduced by Ho et al. (2020), is an instance of this approach. In these models, the variance schedule $\{\beta_t\}$ is often defined as a linear sequence from $\beta_1 = 10^{-4}$ to $\beta_T = 0{,}02$, with T typically fixed at 1000 steps.

The covariance matrix Σ_θ is often fixed during training, typically set as $\Sigma_\theta(x_t, t) = \sigma_t^2 I$. Two common choices for σ_t^2 are β_t or the posterior variance $\tilde{\beta}_t$, defined as:

$$\tilde{\beta}_t = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t. \tag{4}$$

The prediction of the mean μ_θ is the primary focus of the neural network. Instead of directly predicting this mean, Ho et al. (2020) proposed training the network to estimate the noise ϵ added at each step, i.e., to model:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{5}$$

and use this estimate to reconstruct $\mu_\theta(x_t, t)$:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right). \tag{6}$$

This parameterization proved more stable and effective in practice. The architecture used for ϵ_θ is typically based on U-Net (Ronneberger et al., 2015) and often augmented with attention mechanisms.

The generation of new samples proceeds iteratively: it starts with $x_T \sim \mathcal{N}(0, I)$, and at each step the network is applied to predict the noise and generate x_{t-1} until x_0 is obtained. The gradual and controlled nature of this noise reversal is a key factor in the high quality of samples generated by DMs.

Loss Function

DDPM training involves optimizing a Variational Lower Bound (VLB) of the data log-likelihood. Ho et al. (2020) demonstrated that a simplified version of the loss function, which ignores the weights of the VLB terms, works well in practice and is easier to implement. This simplified loss function is:

$$L_{\text{simple}}(\theta) = \mathbb{E}_{t, x_0, \epsilon} \left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t)\|^2 \right]. \tag{7}$$

In this expression, t is sampled uniformly from $\{1, \ldots, T\}$, x_0 is a sample from the real data distribution $q(x_0)$, and ϵ is sampled from a standard normal distribution $\mathcal{N}(0, I)$. Essentially, the neural network ϵ_θ is trained to predict the noise ϵ from the noisy image $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$ and the time step t. This formulation is intuitive: the model learns to *denoise* the image x_t at each step, estimating the noise that was added.

The sampling algorithm for the DDPM, after training the network ϵ_θ, is demonstrated in Algorithm 1.

Algorithm 1 Diffusion Sampling Algorithm

$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, \ldots, 1$ **do**
    $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
    $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
end for
return $\mathbf{x}_0$

The work by Ho et al. (2020) was fundamental in demonstrating that DMs could achieve image quality comparable to or superior to that of GANs across various *benchmarks*, establishing DDPMs as a powerful and influential approach in generative modeling. The clarity of the formulation and the robustness of the results propelled vast subsequent research in the area.

DDPM Limitations and Advances

Despite the success of DDPMs in generating high-fidelity images, they present limitations that motivated subsequent advances. The main one, as highlighted by Ho et al. (2020) (Ho et al., 2020), is the slow and computationally intensive sampling process: generating a single image typically requires $T = 1000$ steps in the reverse process, evaluating the network ϵ_θ at each step. This disadvantage contrasts with models such as GANs, which generate samples in a single pass through the network.

This slowness is due to the Markovian nature of the reverse process, where each x_{t-1} depends only on x_t. Although this formulation enables stable training, it requires simulating the entire noise-removal trajectory in a step-by-step manner. To mitigate this, recent literature proposes various approaches: optimization of the variance schedule $\{\beta_t\}$ (Sabour et al., 2024), reduced sampling schemes (Lee et al., 2025), and distillation techniques (Song et al., 2023), where a smaller and faster model is trained to imitate a larger and slower model.

Advances also include improvements in sample quality and diversity, with more expressive ϵ_θ networks, conditioning mechanisms (like text or classes), and enhanced loss functions. Furthermore, there has been growing interest in connecting DDPMs to other classes of models, such as *score-based models* (Lim et al., 2023) and *normalising flows* (Zand et al., 2023), interpreting them as discretizations of Stochastic Differential Equations (SDEs) or Ordinary Differential Equations (ODEs). This perspective led to the development of more efficient solvers, allowing accelerated sampling without retraining, as is discussed in the section Denoising Diffusion Implicit Models (DDIMs).

The need for multiple network evaluations during generation remains a practical bottleneck, particularly in real-time image processing applications and in those with computational constraints. This restriction spurred the emergence of faster alternatives, such as DDIMs.

Denoising Diffusion Implicit Models

DDIMs, proposed by Song et al. (2020), generalize DDPMs by accelerating sampling. While DDPMs uses a stochastic, Markovian noise-removal process, DDIMs introduces a non-Markovian, deterministic process that significantly reduces the number of sampling steps without compromising image quality.

The main innovation lies in redefining the reverse process. However, the forward process remains the same as in DDPM (Equation 2).

DDIM derives an alternative update rule for x_{t-1} based on the same network $\epsilon_\theta(x_t, t)$ trained for DDPM. From the noise prediction, x_0 is estimated by:

$$\hat{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t)), \tag{8}$$

and x_{t-1} is calculated by:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{x}_0 + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\epsilon_\theta(x_t, t) + \sigma_t z_t, \tag{9}$$

where $z_t \sim \mathcal{N}(0, I)$ and σ_t controls the stochasticity. When $\sigma_t = 0$, the process becomes fully deterministic, resulting in the following simplified equation:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(x_t, t). \tag{10}$$

This process is no longer Markovian, allowing sampling with a subsequence $\{\tau_1, \ldots, \tau_S\}$ of the original schedule $\{1, \ldots, T\}$, with $S \ll T$ (e.g., 20, 50, or 100 steps), which results in much faster generation. The same network ϵ_θ trained for DDPM can be used directly without retraining.

Furthermore, DDIM includes DDPM as a particular case. When $\sigma_t^2 = \tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$, the DDIM sampling process reduces exactly to that of DDPM. Thus, DDIM offers flexibility for both deterministic (with $\sigma_t = 0$) and stochastic sampling.

Sampling consistency, guaranteed in deterministic DDIM, is helpful for applications requiring reproducibility, such as editing and interpolation. This property is not present in DDPM, which, being stochastic, can generate different images from the same x_T.

In terms of efficiency, DDIMs achieves quality comparable to that of DDPMs with 10-20 times fewer sampling steps. This advantage, combined with the possibility of reusing models trained with DDPM, consolidated DDIM as a significant advance in generative DMs.

It is important to note that, although they share the same network $\epsilon_\theta(x_t, t)$ and loss function based on noise prediction, DDPMs and DDIMs differ substantially in the generation process: while DDPMs follow a stochastic Markov chain, DDIMs adopts a deterministic and more efficient process.

2.2 Latent Diffusion Models (LDMs)

While DDIMs significantly address the speed of the sampling steps, another major computational bottleneck arises from operating directly in the high-dimensional pixel space. Processing images of substantial resolution requires evaluating the U-Net on millions of pixels repeatedly, which severely limits training efficiency and increases memory consumption. To resolve this, Rombach et al. (2022) introduced Latent Diffusion Models (LDMs).

Instead of applying the forward and reverse diffusion processes directly to the image, LDMs operate within a compressed latent space. This architecture employs a pre-trained VAE consisting of an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$. The encoder compresses the input image x into a lower-dimensional latent representation $z = \mathcal{E}(x)$. The entire diffusion process (noise addition and denoising) is then performed on z, making it computationally lightweight. Once the reverse process reconstructs a clean latent representation $\hat{z}$, the decoder maps it back to the pixel space to produce the final image $\hat{x} = \mathcal{D}(\hat{z})$. This paradigm shift is what enables models like Stable Diffusion to generate complex images on consumer-grade hardware.

2.3 Conditioning and Feature Injection Mechanisms

The foundational models described above generate unconditional samples, meaning they produce random images from the learned distribution. However, practical applications in computer vision demand precise control over the output. This control is achieved by conditioning the diffusion model on external signals, such as textual descriptions or reference images.

2.3.1 Cross-Attention and the QKV Formulation

To understand how a pre-trained U-Net can accept and process these external conditioning signals, it is necessary to examine the cross-attention mechanism (Vaswani et al., 2017). In a conditioned LDM, the U-Net architecture is augmented with cross-attention layers. These layers are responsible for fusing the spatial features being denoised with the semantic features provided by the user.

In this architecture, the cross-attention operation requires three distinct matrices: the Query (Q), the Key (K), and the Value (V). The Query is derived directly from the U-Net's internal spatial representations (the latent image being denoised). In contrast, the Key and Value matrices are derived from the external conditioning signal y (such as an embedding from a text encoder like CLIP). The mathematical interaction is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \tag{11}$$

where d is a scaling factor corresponding to the dimensionality of the keys. Conceptually, the Query matrix asks "what visual features are needed at this specific spatial location?", while the Key and Value matrices provide the semantic answers derived from the prompt. This mathematical operation effectively injects the external condition into the generative process without altering the base model's pre-trained weights. The structure of this mechanism is detailed in Figure 2.

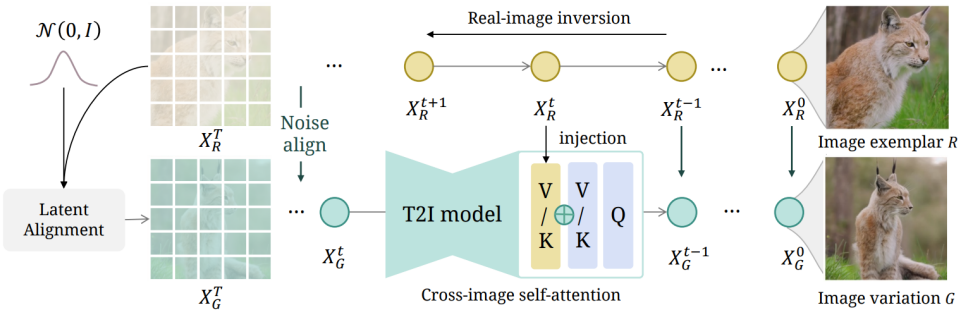


Figure 2: Detailed view of the feature injection process through cross-attention layers. The external condition (text or visual prompt) is projected into Key and Value vectors, which interact with the U-Net's internal Query vectors to guide the image synthesis. Image adapted from (Ye et al., 2023).

2.3.2 Visual Prompting and Decoupled Cross-Attention

Textual prompts often fall short when attempting to convey complex structural details, textures, or specific identities required for tasks like agricultural object detection. This limitation has driven the development of visual prompting techniques, which utilize reference images directly as the conditioning signal y.

A prominent solution in this domain is the IP-Adapter (Ye et al., 2023). Rather than forcing the model to merge text and image features into a single continuous sequence, the IP-Adapter introduces a decoupled cross-attention mechanism. It adds a parallel cross-attention layer specifically dedicated to image features extracted by a visual encoder. This allows the U-Net to process the visual Key and Value matrices independently from the text features. In the context of this thesis, this mechanism is pivotal because it enables the robust transfer of visual styles and structures without the need for computationally heavy fine-tuning methods.

2.4 Foundation Models for Advanced Control

Building upon feature injection, automated synthetic data pipelines require specialized models to handle task-specific constraints, such as spatial boundaries and image curation.

2.4.1 Spatial Control via ControlNet

While cross-attention efficiently handles global semantic guidance, tasks like semantic segmentation demand rigorous pixel-level spatial alignment. Zhang et al. (2023a) introduced ControlNet to address this gap. ControlNet operates by creating a trainable copy of the encoding layers of a pre-trained DM. This auxiliary network learns to process specific spatial conditions, such as segmentation masks, depth maps, or edge boundaries. The output of this sidecar network is then added back into the original, frozen U-Net through zero-initialized convolutions. This strategy prevents the new spatial data from corrupting the base model's generative priors, allowing researchers to generate synthetic images that perfectly align with strict structural labels.

2.4.2 Self-Supervised Clustering with DINOv2

A recurring challenge in visual prompting is the manual selection of reference images. To fully automate a generative pipeline, the system must intelligently identify and cluster visually similar images to use as guides. For this purpose, DINOv2 (Oquab et al., 2023) provides an optimal foundation. Unlike models trained via text-image contrastive learning, DINOv2 relies on self-supervised representation learning. It excels at extracting deep, robust visual features that capture the true structural and textural identity of an object, irrespective of textual descriptions. By utilizing DINOv2 embeddings coupled with clustering algorithms like k-Nearest Neighbors, a pipeline can automatically group diverse but semantically related images, forming the perfect input data for the multi-image latent guidance methods developed later in this research.

3 SYNTHETIC DATA GENERATION: METHODS, APPLICATIONS, AND CHALLENGES

Data scarcity remains a persistent challenge in computer vision, particularly because deep learning models require large-scale datasets for effective generalization. This chapter reviews the evolution of methods, beginning with early approaches, including traditional data augmentation and early forms of synthetic data. It then examines advances in GANs before focusing on current prominent methods, such as DMs.

DMs have significantly advanced image generation, offering high fidelity. Their utility is further enhanced by precise control; this chapter explores techniques for achieving this, including dedicated control networks, lightweight adapters, and methods that leverage attention mechanisms or latent-space manipulation. The practical impact of DMs is evident in two primary roles: as data augmenters that create semantically informed variations and as generators of novel synthetic datasets across diverse applications.

However, the application of DMs also entails challenges, such as computational costs and the complexities of practical prompt engineering, which are discussed herein. The chapter further reviews automated synthetic data generation, with particular attention to the automatic creation of labels. Methodologies for automatic annotation and relevant existing tools and frameworks are also examined. This analysis of DM capabilities and limitations underpins the subsequent research objective: to develop a tool that generates synthetic data to improve the generalization of computer vision models.

3.1 Data Scarcity in Computer Vision Tasks

The significant advances in deep learning technologies for computer vision problems have raised concerns since their inception. The lack of sufficiently large-scale data to enable generalization with larger deep learning models has long hindered progress, whether in specific problems and architectures for object detection (Dang et al., 2023; Sportelli et al., 2023), semantic segmentation (Alomar et al., 2023), or even dense prediction tasks such as depth and pose estimation.

Because data acquisition is inherently complex, time-consuming, and expensive

(Dang et al., 2023), various strategies have been proposed alongside advances in deep learning models to address this challenge. These different strategies are described below.

3.1.1 Traditional Data Augmentation

It comprises techniques that artificially expand the training dataset by applying transformations to the original images, thereby generating synthetic yet realistic samples (Taylor and Nitschke, 2018). The main objective is to improve the model's generalization capability by forcing it to learn the distinctive features of the objects of interest rather than relying on specific visual artifacts or biases present in the training images.

Traditional techniques primarily focus on applying geometric and photometric transformations. Geometric transformations alter the image's spatial structure, thereby training the model to be invariant to changes in position, orientation, and scale. Common examples include flips, especially horizontal flips, which are very effective; rotations to recognize objects regardless of their orientation; translations (shifts); variations in scaling (zooming) to handle objects at different distances; random or centered cropping to focus the model's attention; and more complex deformations like affine (including shear) or perspective transformations, which simulate changes in the camera's viewpoint (Shorten and Khoshgoftaar, 2019).

On the other hand, photometric transformations modify the image's color and illumination properties, thereby making the model robust to variations in capture conditions. This group's brightness adjustments simulate different lighting conditions and changes in contrast, saturation, and color hue (Shorten and Khoshgoftaar, 2019).

Other photometric techniques include the addition of noise, such as Gaussian noise, which can help the model be more robust against image imperfections and prevent overfitting to fine details, and the application of different types of blurring, such as Gaussian blur, to simulate images with lower sharpness or those captured with some motion (Taylor and Nitschke, 2018).

3.1.2 Synthetic Data and Graphics Engines

A data generation approach, distinct from classical data augmentation methods that vary images using standard parameters, is exemplified by the creation of synthetic data. These processes involve more intensive image processing and rendering than traditional data augmentation methods (Zeng, 2024).

For example, in the field of agronomy, Xie et al. (2021) extracts essential items to be recognized by the network using masks, thereby creating new data from the extracted items and using them on different backgrounds and sizes. A similar approach is that of Walawalkar et al. (2020) or Chen et al. (2020), which creates new images from square image patches and combines their databases while preserving the annotations.

Beyond 2D image manipulation, a highly effective approach to synthetic data gene-

ration relies on 3D graphics engines and simulators, such as Unreal Engine or Blender. These simulation environments allow researchers to render highly realistic scenes while automatically extracting pixel-perfect ground truth annotations, such as depth maps, surface normals, and semantic segmentation masks. Recent state-of-the-art models heavily depend on this paradigm. For instance, the Depth Anything model (Yang et al., 2024a) leverages massive amounts of synthetic data rendered from 3D environments to achieve robust monocular depth estimation, proving that simulated datasets can significantly bridge the domain gap. However, building and rendering these 3D environments requires immense manual effort from 3D artists and often lacks the stochastic natural variability found in real-world textures.

Although traditional synthetic data augmentation methods improve model accuracy, the challenge is that the new data generated or augmented remain too similar to the distribution of the original training data (Shipard et al., 2023). This restricted variability limits the model's ability to develop robust, generalizable representations, often resulting in insufficient capture of contextual information and poor feature preservation in unseen scenarios (Islam et al., 2024b).

Consequently, the challenge of generalizing to diverse, real-world data remains unresolved by these approaches alone. Generative image models emerge as a potential solution by synthesizing realistic samples that bridge the gap between limited datasets and real-world variability.

3.1.3 Generative Adversarial Networks

In response to the need for automatic, complex image generation that produces previously unseen samples, various image generation models have emerged (Shipard et al., 2023). Among these, the GAN architecture stood out as particularly prominent for an extended period, becoming a key focus in the field.

The core of GANs lies in an adversarial process between two neural networks: a generator and a discriminator (Mirza and Osindero, 2014). The generator learns to create synthetic data samples, while the discriminator is trained to distinguish between real data and the generator's output. Through this competition, where the generator seeks to fool the discriminator and the discriminator improves its detection capabilities, the generative model is driven to produce increasingly realistic results.

Building on this framework, various GAN architectures have been developed for specific image-generation and translation tasks. These include Conditional Generative Adversarial Networks (CGANs), which allows generation guided by additional information, Pix2Pix for paired image-to-image translation (Mirza and Osindero, 2014), CycleGAN (Zhu et al., 2017) for unpaired domain translation, and StarGAN (Choi et al., 2018). Its successor, StarGAN v2 (Kaneko et al., 2019), is for multi-domain image translation.

These GAN architectures and their variants have been successfully applied in a wide

range of computer vision tasks. Examples include improving general object detection models for natural environments (Sahu et al., 2024) and performing crucial data augmentation in complex domains like medical imaging (Aljohani and Alharbe, 2022).

Despite the success and versatility of GANs, certain limitations persist, including challenges with training stability and variability in image quality. The adversarial nature of GAN training can sometimes hinder their ability to consistently achieve optimal metrics, particularly for complex and diverse data distributions (Dhariwal and Nichol, 2021).

In contrast to some of the challenges faced by GANs, diffusion models have recently emerged as a powerful alternative for image synthesis. A significant development in this area was the demonstration that advanced diffusion models can surpass GANs in generating high-fidelity images across diverse datasets, marking a notable milestone for realistic data synthesis (Wang et al., 2023c).

3.2 Diffusion Models

Research on generative models has evolved significantly with the advent of diffusion models, which can model complex distributions and generate high-quality samples (Cao et al., 2024; Rombach et al., 2022). Inspired by thermodynamics, they operate by gradually destroying the data structure with Gaussian noise and then learning to reverse this process to reconstruct the data (Wang et al., 2024).

DDPMs (Ho et al., 2020) demonstrated high-fidelity image generation through an iterative process of adding and subsequently removing noise. Their work established that diffusion models can achieve image quality comparable to or superior to GANs, demonstrating this through evaluations on quantitative benchmarks.

To accelerate the slow inference of DDPMs, Song et al. (2020) introduced a non-Markovian generation process. This enabled significantly faster sampling (fewer inference steps) without substantial loss of quality, thereby making diffusion models more practical.

A crucial efficiency optimization was introduced by LDMs (Rombach et al., 2022). By applying the diffusion process in a lower-dimensional latent space (learned by an autoencoder) rather than in the pixel space, LDMs significantly reduces computational costs. This facilitated the generation of very high-resolution images and the training on massive datasets, thereby laying the foundation for popular models such as Stable Diffusion.

The need for automatic, complex image generation capable of producing previously unseen samples has driven the emergence of various generative models. Initially, *guided diffusion* (Nichol et al., 2021) addressed the requirement for greater control by utilizing classifiers to steer the sampling process toward desired characteristics. The advent of the Contrastive Language-Image Pre-training (CLIP) model significantly enhanced this control (Radford et al., 2021), which guided diffusion with textual descriptions (prompts),

thereby enabling highly flexible text-to-image generation.

This enhanced control is theoretically supported by the understanding that diffusion models exhibit behavior akin to clustering algorithms (Wang et al., 2024). This inherent clustering property, combined with their iterative nature, provides a theoretical basis for leveraging relevant information during the generation process to guide image synthesis. Achieving precise control over specific requirements in diffusion models is paramount; consequently, a range of advanced techniques has been developed, including model superposition, adapters, and manipulation of attention mechanisms or initial noise.

3.2.1 Control Models

This category employs network architectures designed to incorporate spatial controls into large pre-trained diffusion models (Zhang et al., 2023b). Their mode of operation is based on using two copies of the pre-trained diffusion model, with one of these parallel models serving as an additional control over the other frozen model. This external diffusion model is then used to learn task-conditional controls, such as edge detection, depth estimation, segmentation, or human pose maps (Zhao et al., 2023). The connection between these two models is established via zero convolutions, which prevent the introduction of detrimental noise into the deep features of the frozen base model.

The most characteristic model of this type of control is ControlNet (Zhang et al., 2023b), which is trained on specific tasks to guide image generation (e.g., depth and semantic masks). Uni-ControlNet (Zhao et al., 2023) extends ControlNet's capabilities by providing a unified framework for the simultaneous use of local controls (e.g., edge maps, depth, segmentation masks) and global controls (e.g., CLIP image embeddings) within a single model.

Limitations and Relation to this Work: While ControlNet and its variants offer unprecedented spatial precision, their primary limitation is the requirement for task-specific fine-tuning on large datasets of paired images and masks. This introduces a high computational overhead. In this thesis, we address this by leveraging pre-trained ControlNet architectures, as detailed in Chapter 6, and combining them with advanced vision-language models to automate the generation of underwater segmentation datasets without needing to retrain the control module from scratch.

3.2.2 Adapters for Guidance

Adapters provide an efficient means of guiding diffusion models by incorporating additional information without modifying the base model architecture. In simple terms, an adapter is a smaller, lightweight module attached to the base diffusion model to guide the generation process.

The IP-Adapter (Ye et al., 2023) is an example of this, providing a lightweight and practical adapter that enables image prompting in pre-trained text-to-image diffusion mo-

dels. Its key design is a decoupled cross-attention mechanism that separates text and image features, allowing the adapter to achieve performance comparable to, or even better than, thoroughly fine-tuned models such as ControlNet (Zhang et al., 2023b; Ye et al., 2023) with only 22 million parameters.

On the other hand, Generalized Diffusion Adaptation (GDA) (Tsai et al., 2024) is a test-time adaptation method that is robust to various types of Out-Of-Distribution (OOD) data. GDA iteratively guides the diffusion process by applying a model-derived marginal-entropy loss, along with style and content-preservation losses, while accounting for both the model's output behavior and the samples' semantic information.

For image generation via domain or style transfer, PromptStyler (Cho et al., 2023) trains an external classifier based on the required styles in diffusion models' outputs, thereby preserving the semantic content of the same subject across a variety of styles. In this way, the characteristics of the same output subject are maintained across different styles.

Limitations and Relation to this Work: The main limitation of relying solely on visual adapters is that they often capture the global style but can lose strict semantic fidelity or specific object features when transferring domains. To overcome this, Chapter 4 of this research directly employs the IP-Adapter but synergistically combines it with specialized semantic feature extractors, like SeeCoder, to generate highly accurate agricultural datasets in a completely training-free manner.

3.2.3 Attention Guide

Attention mechanisms play a crucial role in controlling diffusion models. Fundamentally, this mechanism operates through three linear projections: Query, Key, and Value (QKV). The attention operation computes the similarity between Query and Key to determine the relevance of Value, thereby enabling precise manipulation of the generative process by selectively incorporating additional conditional information (Vaswani et al., 2017).

Training-Free Structured Diffusion Guidance (Feng et al., 2022) is a technique that enhances models' compositional abilities by incorporating linguistic structures into the diffusion guidance process, manipulating cross-attention layers with specific text tokens to achieve more precise attribute binding and improved image composition without requiring additional training.

Style Injection in Diffusion (Chung et al., 2024) is a style-transfer method that extracts the QKV feature vectors from the self-attention mechanism during DDIM inversion of content and style images. These extracted components are then injected into the stylized image's sampling process to improve its color, texture, and semantics.

RIVAL (Zhang et al., 2023c) is an inference pipeline that generates real-world image variations from a single example, using diffusion models without the need for additional

optimization. This is achieved through cross-attention injection for feature interaction and step-wise distribution normalization to align latent features.

In the context of cross-attention manipulation for text-driven editing, Hertz et al. introduced a training-free method. This approach integrates information from both text and reference images by blending their respective attention maps at each layer, with weights empirically determined (Hertz et al., 2022).

Limitations and Relation to this Work: Despite their elegance, attention-guided methods frequently depend on meticulous prompt engineering or the manual selection of ideal reference images, which limits their scalability for large dataset generation. This thesis builds upon the concept of training-free manipulation by bypassing manual prompt engineering altogether, opting instead for automatic visual encoders to drive the attention layers.

3.2.4 Latent Space Guide

In the image regeneration process via latent diffusion, a dedicated step extracts image features, yielding a latent space to which the denoising module is applied (Song et al., 2020). In this latent space, various works have employed different forms of guidance by varying the arrangement between the condition images and the latent representations of the generated image.

As a first latent-guided model, TALE (Pham et al., 2024) is a training-free framework for image composition across different visual domains. It operates purely in the latent space of diffusion models, using adaptive latent manipulation and energy-guided latent optimization to integrate objects into backgrounds of different styles coherently.

On the other hand, Diffusion in Diffusion (Wang et al., 2023b) explores the properties of diffusion (physics) within diffusion (machine learning), proposing a Cyclic One-Way Diffusion (COW) method to control the diffusion direction in pre-trained models. This technique is helpful for personalization scenarios, as the image's characteristics are incorporated into the scenario via the latent space.

In Cartoondiff (He et al., 2024), the image generation process is divided into semantic and detailed parts. Because their goal is to obtain cartoon images without retraining the network, the authors normalize the resulting latent images in the first phase. In this way, they achieve zero-shot cartoon image generation.

In Universal Guidance for DMs (Bansal et al., 2023), Bansal et al. present an algorithm that enables control of diffusion models via arbitrary guidance modalities, such as segmentation, facial recognition, or object detection, without requiring specific retraining. This is achieved by evaluating the guidance models only on noise-free latent images, using a threshold empirically derived from the conditional information.

Limitations and Relation to this Work: A critical challenge in latent space manipulation is maintaining structural coherence. Excessive or unstructured latent blending

can lead to morphological artifacts, where biological structures may collapse entirely. Chapter 5 directly tackles this limitation by proposing a structured, multi-image latent guidance framework (ZEDD-G) that safely blends features to augment classification datasets without destroying the target object's geometry.

3.3 Evaluation Metrics for Synthetic Data in Generative Models

To properly assess the quality of the datasets produced by the aforementioned generative models, robust quantitative metrics are required. The evaluation of synthetic data has historically relied on the Frechet Inception Distance (FID) (Heusel et al., 2017), which measures the discrepancy between the feature distributions of authentic and generated images.

However, FID has several limitations. It assumes that Inception-v3 embeddings follow a multivariate normal distribution, an assumption that does not hold for the complex outputs of modern generative models (Jayasumana et al., 2024). Additionally, FID requires large sample sizes to produce reliable estimates and fails to consistently reflect incremental improvements or degradations in image quality (de Deijn et al., 2024).

To overcome these limitations in the context of DMs, the CLIP-MMD Distance (CMMD) has been proposed (Jayasumana et al., 2024). This metric combines CLIP embeddings with Maximum Mean Discrepancy (MMD) using a special kernel. This kernel enables robust distribution comparison by measuring similarity via the Euclidean distance between feature vectors in the embedding space.

Unlike FID, CMMD does not rely on assumptions about data distribution; it is computationally efficient and effectively captures incremental changes in image quality. Empirical studies have shown that CMMD aligns more closely with human perception, making it a robust alternative for evaluating modern generative models, and it will be the primary metric utilized in the subsequent chapters of this research to validate our synthetic agricultural and underwater datasets.

3.4 Application and Challenges in Data Generation

As with GAN models before diffusion, DMs are also used in various applications to both augment data more traditionally and generate synthetic data.

3.4.1 Diffusion Models as Data Augmentation

Generative image models, specifically those based on diffusion, enable the creation of entirely new data instances or the modification of existing ones in semantically rich ways.

A direct approach is image-to-image translation between domains. Kwon and Ye (2022) proposes a method for unsupervised image translation. The important aspect here

is how the diffusion model learns to disentangle an image's content and style representations. Then, it can recombine the content of one image with the style of another (or a sampled style), generating a new image that preserves the original semantics while exhibiting a different visual appearance, thereby serving as a powerful data augmentation technique (Kwon and Ye, 2022).

Improving the dataset's intrinsic diversity is another key goal. Shipard et al. (2023) addresses this directly. Instead of simply generating more copies of the training dataset, subtle but significant variations of existing training images can be created. The key insight is that they generate samples that, although close to the original distribution, cover underrepresented regions of the feature space, thereby forcing the model to learn more robust representations (Shipard et al., 2023). A similar case is observed in (Fu et al., 2024), with the caveat that the generation is unguided and filtered empirically. Furthermore, the variations must be applied to a network that has been previously fine-tuned on the data.

Data mixing techniques, such as Cutmix (Walawalkar et al., 2020), have also been reinvented through diffusion models. DiffuseMix (Islam et al., 2024a) proposes an innovative strategy: instead of interpolating images in pixel space (which often produces blurry artifacts), DiffuseMix operates in the diffusion model's latent space. It interpolates the latent representations of two or more images. It then applies diffusion decoding to generate a hybrid sample that more coherently combines the features of its parents.

Similarly, GenMix (Islam et al., 2024b) presents a hybrid approach handy for small datasets. The central idea is to interleave the generation of new samples, meaning that not only are new instances generated, but these instances (and/or the original ones) are additionally interpolated.

Finally, research such as that presented by (Trabucco et al., 2023) systematically explores how diffusion-based transformations can complement or surpass traditional augmentation techniques. These studies often analyze which specific types of generative augmentation (e.g., variations in texture, shape, or context) are most beneficial and how they can be optimally integrated into training pipelines.

3.4.2 Diffusion Models as Synthetic Data Generators

In agronomy and plant pathology, these models are fundamental for generating synthetic images of plant diseases, using techniques such as LDMs and variants of Stable Diffusion to create data for underrepresented or even unseen disease classes, thereby improving the robustness of detection systems (Muhammad et al., 2023; Mori et al., 2024; Rai et al., 2025; Li et al., 2023c; Ranario et al., 2025). Beyond disease detection, they contribute to the creation of augmented datasets for crop monitoring, phenotyping, yield estimation, and semantic segmentation, thereby introducing relevant variations for agricultural analysis (Ranario et al., 2025). They also enable the simulation of diverse agricultural scenarios, such as different soil types or lighting conditions, for more com-

prehensive model training (Heschl et al., 2024a).

In large-scale image classification tasks, such as ImageNet, synthetic data generated by DMs has been shown to improve classifier accuracy and robustness (Azizi et al., 2023). The creation of completely synthetic datasets for model pre-training or evaluation is actively explored, and even zero-shot classification approaches (Sarıyıldız et al., 2023; Li et al., 2023a). A prominent application is mitigating class imbalance by generating examples for rare or underrepresented categories (Samuel et al., 2024).

Remote sensing also benefits significantly, with DMs generating realistic synthetic satellite images for environmental monitoring and urban planning, as well as for simulating the effects of natural disasters or landscape evolution (Sousa et al., 2024; Khanna et al., 2023). These models are adapted for specific tasks, such as image super-resolution, infilling missing data (e.g., due to clouds), and temporal prediction (Khanna et al., 2023). Data augmentation with semantically diverse or controllable synthetic images improves performance on tasks such as object detection and land-cover classification (Tang et al., 2024).

DMs also has up-and-coming applications, specifically in domains where data acquisition is difficult or costly, such as in the medical field. Saragih et al. (2024) explores this potential, demonstrating how DMs can be employed to create realistic synthetic medical images.

Despite these advances, guided generation with DMs remains a significant challenge, as detailed below.

3.4.3 Computational Costs

The primary bottleneck of the aforementioned applications is the immense computational cost associated with adapting the models to new domains. All the articles reviewed in the previous section have a severe limitation: they require fine-tuning for image generation and are therefore unsuitable for rapid deployment on the specific models to be trained. Generally, Dreambooth (Ruiz et al., 2023) is used, a DM training pipeline that accelerates this process. However, it does not address the underlying hardware requirements of at least 8GB of VRAM, sufficient RAM, and substantial disk space. Furthermore, this training must be performed iteratively by validating hyperparameters to obtain the best possible images for a specific problem (Ramos et al., 2025b).

To mitigate these prohibitive costs, several researchers have proposed approximating the models to enable faster, more efficient training. DiffFit (Xie et al., 2023) is a pipeline that fine-tunes a DM. Unlike (Ruiz et al., 2023), it trains on the biases of the general model rather than the entire parameter set, resulting in faster, lighter, and more efficient training.

Similarly, Sauer et al. seeks to distill DMs onto a lighter model, also using a discriminator as a loss function between the student and teacher models, achieving realistic image

generation in fewer steps than the original (Sauer et al., 2024).

An interesting case is also (Kang et al., 2024), which leverages a pre-trained DM as a teacher for a GAN, thereby mitigating the DM's image-quality limitations and yielding a lighter model akin to earlier generative models.

In (Guo et al., 2024), on the other hand, an additional external guide is used. They train a smaller model to guide images to a new domain, focusing on the more well-known problem of domain adaptation. This likewise seeks to guide pictures to a new semantic field, falling within data generation and presenting a possibility when considering retraining the heavier DMs.

Alternative strategies also include optimizing specific components of the model, such as the encoder in LDMs, to achieve faster generation (Li et al., 2023b). Concept elimination is another form of guidance, achieved by limiting the possible outputs. In one of their experiments, Gandikota et al. demonstrates the possibility of guiding generation without requiring additional training (Gandikota et al., 2023).

Despite various developments and proposed methods designed to enhance the efficiency and controllability of DMs, a critical gap remains. These refined approaches frequently yield performance below that of larger-scale models trained extensively on large datasets. Moreover, they frequently yield highly specialized models tailored to narrow tasks or domains, thereby losing some of the broad, general knowledge captured by the extensively pre-trained foundational models. This specialization and loss of expertise ultimately diminish the models' overall capacity and versatility for synthesizing data to meet novel or diverse requirements.

3.4.4 Guide Challenge (Prompt Engineering)

Even if the computational costs of training are bypassed or mitigated, a second major operational bottleneck emerges: the reliance on text prompts to steer the generation. The complexity of obtaining the desired results solely through text is a significant limitation. Although approaches that incorporate additional spatial information exist, such as (Zhao et al., 2023; Zhang et al., 2023b), they fundamentally operate on top of text-to-image architectures. Therefore, they still rely heavily on text-prompt guidance to define the semantic identity of the generated object.

This is evident in the articles presented in Subsection 3.4.2, which all require meticulous prompt engineering to obtain acceptable results. As mentioned in (Hertz et al., 2022), this aspect is so crucial and brittle that the results can vary drastically even with the alteration of a single text token. This unpredictability makes prompt engineering more of an art than a rigorous science (Xu et al., 2024).

As a result, some approaches aim to generate increasingly complex prompts based on the desired final guidance for the images. One example is TF-ICON (Lu et al., 2023), which uses DMs to edit pictures and proposes the creation of an exceptional prompt. The



Figure 3: Illustration of the prompt engineering challenge. Minor variations in textual descriptions can lead to significant and unpredictable changes in the generated output, highlighting the need for robust visual prompting methods.

objective of this is to remove text guidance to avoid interference in the editing altogether.

Another case in which the prompt is fundamental is LLM-grounded Diffusion (Lian et al., 2023a). With this method, scene interpretation is performed using a pre-trained Large Language Model (LLM). This prompt is then introduced to both the DM and its adapter, which, using masks and at each generation step, seeks to ensure that the prompt is fulfilled for image editing.

There is also the problem of domain generalization with this text-based generation guidance. When fixed training is used on a public training set, there will inevitably be more samples from a single domain for the same subject represented by a token (Samuel et al., 2024).

To entirely avoid this problem, Xu et al. proposes Seecoder, a model trained to replace the CLIP component of the diffusion. Seecoder takes a guidance image as input, from which features are extracted and used as the primary guide for diffusion, thereby obviating the need for text or more complex prompt engineering (Xu et al., 2024).

3.5 Automatic Synthetic Data Generation

As demonstrated by research such as the Depth Anything model (Yang et al., 2024a), manually obtained labels are inherently prone to errors. Their research in monocular depth estimation achieved excellent results, attributing this success significantly to the use of synthetic data. Yang et al. highlighted that even with expert annotators, manual labeling introduces inaccuracies. Consequently, the findings underscore the critical value of high-quality, automatically labeled synthetic data as a means to bypass these inherent errors in manual annotation and achieve superior performance in computer vision tasks.

To the best of our knowledge, no unified tools exist that automatically generate synthe-

tic labeled data using DMs without human intervention. The SOTA that uses this generative model performs most of the process manually, including training, empirical guidance, and ultimately image annotation.

Thus, some articles could inform the development of this tool by identifying which key concepts are applicable and whether existing tools provide a suitable basis.

3.5.1 Automatic Labels and Masks Generation

In labeled data generation, there are two approaches. The first, based on ControlNet's ability (Zhang et al., 2023b) to generate images from masks, provides a means to create annotated images by leveraging this guidance. Image variations are made using methods such as FreeMask (Yang et al., 2023) or more complex approaches that combine the ability to perform variations similar to classical data augmentation (filters, copied images) with image generation guided by masks, as in (Xie et al., 2024).

On the other hand, recalling Section 3.2, where it is established that the attention layers of models can be used to guide image generation; in this case, information is extracted from these layers. In (Wang et al., 2023a), this fundamental part of the model is used to automatically extract masks without having to define them beforehand using a heatmap. This method leverages the iterative nature of generation to produce heatmaps of the generated object throughout the process, which are then used to create a mask from these maps.

A similar case is found in DiffuMask (Wu et al., 2023), based on the previous work of Wang et al. (2023a), but with a substantial improvement by introducing a variable that allows obtaining this mask more precisely and without having to perform initial empirical checks.

Furthermore, Nguyen et al. leverages attention maps, refines the prompt with a language model, and employs an auto-training technique to train a semantic segmenter. It also highlights that it generates masks for multiple object types (Nguyen et al., 2023).

3.5.2 Tools and Frameworks

The rapid rise in the popularity of text- and image-generation models, coinciding with heightened general interest in Artificial Intelligence (AI), has led to the emergence of a diverse range of image-generation tools. These tools, offered as both cloud-based and local solutions, have consequently made these underlying architectures more accessible to the general public.

Well-known cases include cloud-based generation methods such as DALL-E (Ramesh et al., 2022) and Stable Diffusion (Rombach et al., 2022), which are implemented in the cloud by their respective proprietary companies.

On the other hand, there are local tools. Their objective is generally to make open-source models available to most interested users.

For example, the graphical tool ComfyUI (Anonymous, 2023) uses the Diffusers library (von Platen et al., 2022), which is optimized for CPU execution, and, with its graphical interface, has become popular in the generation community. Furthermore, this tool enables the creation of dedicated nodes for image generation, with SOTA examples discussed here, such as IP-Adapter (Ye et al., 2023) and Seecoder (Xu et al., 2024), which are implemented for direct use. ComfyUI has the potential to serve as the foundation for a more powerful generative tool for scientific applications, although it has primarily focused on recreational image generation.

Synthetic data has demonstrated effectiveness in improving deep learning model accuracy and generalization, with DMs currently leading in generating realistic images for this purpose and in data augmentation, owing to their versatility and extensive guidance capabilities for image editing. However, much of the research on guided diffusion focuses on recreational image manipulation rather than generating synthetic training data. This process remains slow and resource-intensive due to manual steps like retraining and annotation, despite the increasing accessibility of these generation tools for recreational users. We identified a critical need for an automatic generation tool to overcome this slow pipeline, enabling efficient experimentation with diverse quantities and types of synthetic data without incurring prohibitive computational costs, which can often exceed 10 GPU-days per model.

3.6 Evaluation Metrics for Synthetic Data

To properly assess the utility of the datasets produced by the aforementioned generative frameworks, robust quantitative metrics are essential. The evaluation of synthetic data has historically relied on the FID (Heusel et al., 2017), which measures the discrepancy between the feature distributions of authentic and generated images.

However, FID has several limitations. It assumes that Inception-v3 embeddings follow a multivariate normal distribution, an assumption that does not hold for the complex outputs of modern generative models (Jayasumana et al., 2024). Additionally, FID requires large sample sizes to produce reliable estimates and fails to consistently reflect incremental improvements or degradations in image quality (de Deijn et al., 2024).

To overcome these limitations in the context of DMs, the CMMD has been proposed (Jayasumana et al., 2024). This metric combines CLIP embeddings with MMD using a special kernel. This kernel enables robust distribution comparison by measuring similarity via the Euclidean distance between feature vectors in the embedding space.

Unlike FID, CMMD does not rely on assumptions about data distribution; it is computationally efficient and effectively captures incremental changes in image quality. Empirical studies have shown that CMMD aligns more closely with human perception, making it a robust alternative for evaluating modern generative models. This metric serves as a

foundational tool for validating the synthetic datasets generated in the subsequent chapters of this thesis.

3.7 Comprehensive Summary and Research Gap

To contextualize the proposed research, we present a comprehensive categorization of the reviewed literature. SOTA methods are classified along two fundamental axes: architectural control mechanisms (Table 1) and application strategies for synthetic data generation (Table 2).

3.7.1 Trade-offs in Control and Efficiency

Table 1 organizes the generative methods ranging from computationally intensive training to inference-only guidance. A clear pattern emerges: methods that achieve high fidelity and concept retention typically require expensive fine-tuning (e.g., Dreambooth, ControlNet), while efficient training-free methods (e.g., Attention-based) often lack the robustness needed for automatic, large-scale data generation without manual intervention.

Table 1: Taxonomy of Control Mechanisms and Optimization Techniques in Diffusion Models. This table contrasts the guidance method with its computational prerequisites.

Category	Technique / Architecture	Key Works (Citations)	Computational & Operational Cost
Full Fine-Tuning & Training	Weights Update & Dreambooth	Dreambooth (Ruiz et al., 2023), Textual Inversion (Samuel et al., 2024)	**Very High.** Requires >8GB VRAM per model. Slow iterative training for every new concept.
Structural & Adapter Control	Zero-Convolutions, Parallel Encoders, Lightweight Modules	ControlNet (Zhang et al., 2023b), Uni-ControlNet (Zhao et al., 2023), IP-Adapter (Ye et al., 2023)	**High/Medium.** Reduces parameters but still requires training on paired datasets (images + masks/depth).
Attention & Latent Guidance (Training-Free)	Cross-Attention Injection, Latent Optimization, Energy Functions	Prompt-to-Prompt (Hertz et al., 2022), TALE (Pham et al., 2024), RIVAL (Zhang et al., 2023c), Universal Guidance (Bansal et al., 2023), TF-SDG (Feng et al., 2022), COW (Wang et al., 2023b)	**Medium.** No training required, but inference is complex and slow. Heavily dependent on precise prompt engineering.
Model Optimization & Distillation	Knowledge Distillation, Bias Tuning, Concept Erasing	DiffFit (Xie et al., 2023), Distillation (Sauer et al., 2024; Kang et al., 2024), Erasing Concepts (Gandikota et al., 2023), Faster Diffusion (Li et al., 2023b)	**High (Initial).** Creating the "efficient"model requires a heavy distillation phase. Resulting models often lose zero-shot generality.
Prompt Engineering	Iterative Refinement, LLM-based guidance	LLM-grounded (Lian et al., 2023a), TF-ICON (Lu et al., 2023), PromptStyler (Cho et al., 2023), Seecoder (Xu et al., 2024)	**Low Compute / High Manual Effort.** Shifts the burden from GPU to complex prompt design (or external LLM costs).

3.7.2 Fragmentation in Data Generation Strategies

Table 2 maps how these models are applied to solve data scarcity. While significant progress has been made in specific domains (Agronomy, Remote Sensing) and augmentation techniques (Mixing, Infilling), there is an absence of a unified framework. Existing solutions are highly fragmented and isolated to specific applications. They effectively solve a narrow problem (e.g., removing clouds) but fail to provide a generalizable tool for automatic, labeled data generation across different tasks.

Table 2: Overview of Data Generation Strategies using Diffusion Models. The table highlights the specific application domains and the limitations of current methodologies regarding generalization.

Strategy	Mechanism	Key Works & Applications	Gap / Limitation
Domain-Specific Synthesis	Fine-tuned LDM for rare classes or specific environments	Agronomy: (Muhammad et al., 2023; Mori et al., 2024; Rai et al., 2025; Ranario et al., 2025; Heschl et al., 2024a) Remote Sensing: (Sousa et al., 2024; Khanna et al., 2023; Tang et al., 2024) Medical: (Saragih et al., 2024)	**Lack of Generalization.** Each solution requires a dedicated dataset and training pipeline. Not transferable to new domains without re-implementation.
Advanced Data Augmentation	Latent Space Mixing, Style Transfer, Inpainting	DiffuseMix (Islam et al., 2024a), GenMix (Islam et al., 2024b), DreamDA (Fu et al., 2024), Diff. Translation (Kwon and Ye, 2022), Diversity (Shipard et al., 2023)	**Dependence on Reference.** Mostly varies existing samples rather than generating novel, labeled concepts from scratch.
Class Balancing	Zero-shot generation for long-tail distributions	ImageNet Balancing (Azizi et al., 2023), Fake it till you make it (Sarıyıldız et al., 2023), Your Diffusion (Li et al., 2023a), Rare Categories (Samuel et al., 2024)	**Scale Issue.** Often requires massive pre-trained models or computing clusters to be effective on benchmarks like ImageNet.
Auto-Labeling & Segmentation	Attention Map Extraction, Mask-to-Image, Dataset Generation	DiffuMask (Wu et al., 2023), DatasetDiffusion (Nguyen et al., 2023), DiffusionSecret (Wang et al., 2023a), FreeMask (Yang et al., 2023), MosaicFusion (Xie et al., 2024)	**The Core Gap.** Existing tools are either recreational (ComfyUI), slow, or require manual prompts/masks. No fully automated, low-resource pipeline exists.

3.7.3 Conclusion: The Need for Automated Efficiency

The synthesis of these works reveals that while the components for effective synthetic data generation exist, such as attention control (Section 3.2) and automated mask extraction (Section 3.5), they are currently fragmented.

Current research (Table 1) leans heavily toward computationally expensive training for control, whereas application-level research (Table 2) remains domain-locked. There is a distinct lack of a tool that integrates training-free guidance mechanisms with automated labeling capabilities to efficiently produce diverse synthetic datasets.

4 TRAINING-FREE SYNTHETIC DATA STRATEGIES FOR AGRICULTURAL OBJECT DETECTION

In the preceding analysis of the state of the art, we identified a critical dichotomy: while synthetic data holds immense promise for agriculture, its practical application is often hindered by high computational costs and inconsistent results. The current literature presents conflicting evidence; for instance, studies such as (Deng, 2023) and (Moreno et al., 2023) report divergent outcomes despite employing similar methodologies on comparable datasets. This variability suggests that the effectiveness of synthetic data is not intrinsic but highly dependent on the generation pipeline and the specific integration strategy used.

To address this uncertainty and the lack of a standardized baseline, this chapter proposes a rigorous experimental framework using the (Dang et al., 2023) dataset. We aim to benchmark a novel, training-free generation pipeline against established fine-tuning methods. By systematically evaluating method variability, ranging from class-balancing strategies to the generation architecture itself, we aim to develop a robust protocol for integrating synthetic data into weed-detection tasks, prioritizing both downstream performance and computational efficiency.

4.1 Synthetic Images Generation by DreamBooth Model

Synthetic image generation supports data augmentation strategies. Here, we use DMs that iteratively denoise random noise using learned distributions to produce high-quality images.

For our experiments, we adopt a diffusion-based approach inspired by DreamBooth (Ruiz et al., 2023), which enables the generation of class-specific images using a small set of reference images. DreamBooth fine-tunes a pre-trained DM to learn the distinctive characteristics of each class, thereby producing diverse and accurate synthetic samples. According to (Ruiz et al., 2023), three to eight reference images per class are sufficient to capture key features, making this an efficient method for synthetic data generation.

However, one of the challenges in the synthetic image generation process is preparing

a high-quality reference dataset. A carefully curated, diverse set of images is necessary to ensure that the model generates accurate, varied samples.

4.1.1 Clustering based on the DINOv2 Algorithm

One of the key aspects of building an appropriate reference dataset for training an image generation model is image clustering. Since training a DreamBooth model requires between three and eight reference images per class, efficiently clustering the images into manageable groups is a priority. Typically, this process involves selecting images based on their *similarity* (Moreno et al., 2023).

However, this clustering has two main limitations: subjectivity in determining image similarity and reduced efficiency for large datasets. These limitations can affect the quality and utility of the generated data.

The DINOv2 proposal (Oquab et al., 2023), designed for feature extraction and combined with k-Nearest Neighbors (KNN) for clustering, addresses the aforementioned limitations and clusters the dataset into distinct groups. This approach promises higher-quality group formation, which is relevant to subsequent pipeline stages.

4.1.2 Enhanced Synthetic Data Generation

Based on the SOTA analysis, we propose a methodology for automatic synthetic data generation that addresses key challenges in data augmentation while reducing computational costs. Instead of training a DM from scratch, which is time-consuming and requires significant Video Random Access Memory (VRAM), our approach leverages two complementary methods: IP-Adapter (Ye et al., 2023) and SeeCoder (Xu et al., 2024).

IP-Adapter is a modification of the original DM architecture that enhances the attention calculation. This modification introduces an additional step in which the model compares guide-image embeddings, thereby improving its ability to generate images from reference data rather than textual prompts (Ye et al., 2023). SeeCoder, by contrast, extracts visual features from an image, thereby overcoming the vocabulary limitations of text-to-image models by using images as prompts rather than text (Xu et al., 2024).

The reason for combining these two approaches lies in their complementary strengths. While SeeCoder extracts visual features from images, it still exhibits limitations in image generation, particularly in alignment. Although the visual prompt guides the generation of unseen photos, there is often a gap between the generated output and the original prompt. Images generated with SeeCoder may lack visual realism and deviate significantly from the reference image in content and quality.

On the other hand, the IP-Adapter still faces challenges in prompt engineering, which require significant time and effort to identify the optimal prompt for effective data generation. The trial-and-error process of selecting appropriate words to guide the model can make it challenging to create diverse classes, potentially disrupting synthetic data

generation.

We hypothesize that both models improve the generation process by addressing different limitations. SeeCoder enhances visual feature extraction, enabling more diverse image generation. At the same time, the improved attention mechanism in IP-Adapter helps align the generated images with the desired output, enhancing realism and relevance. Our generation model leverages both and can achieve better results without requiring extensive prompt engineering or additional training.

Figure 4 demonstrates how both IP-Adapter and SeeCoder are used together with guide images to enhance synthetic data generation. The figure illustrates how embeddings from guide images are processed by the modified attention mechanism of IP-Adapter and the visual feature extraction of SeeCoder, yielding more accurate and realistic image generation.

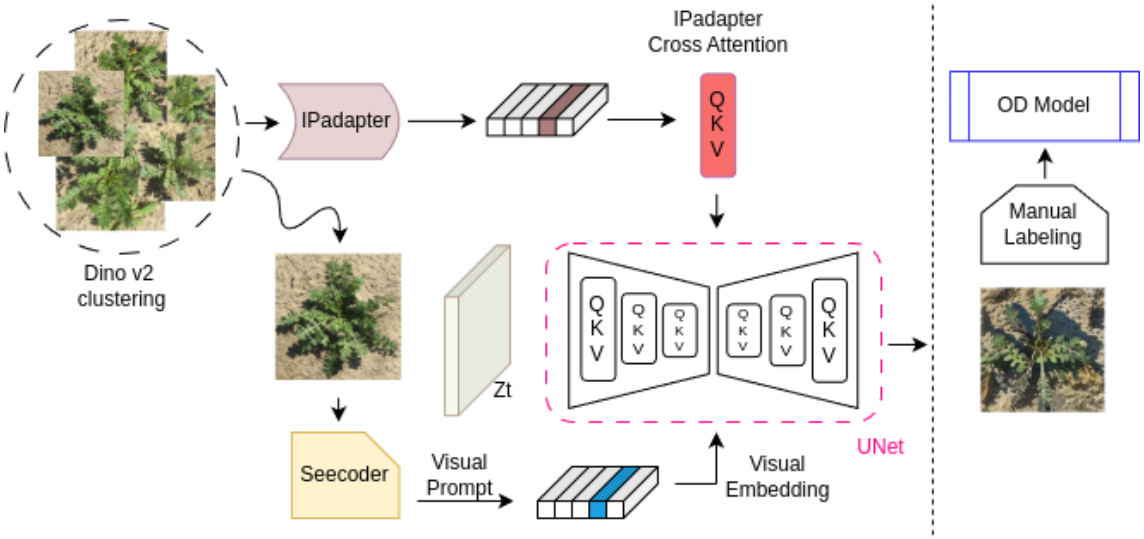


Figure 4: Combined use of IP-Adapter and SeeCoder with DINOv2-clustered guide images for enhanced synthetic data generation.

4.2 Experiments

This chapter details the experimental methodology and results obtained from our research. Building on the theoretical foundations laid in Chapter 2 and the methodological approach described in Section 4.1, we first describe the setup for generating synthetic data, including the specific dataset used, before examining the various experimental configurations and their outcomes.

4.3 Synthetic data setup

Following the methodology outlined in Section 4.1, this section details the setup of our synthetic data generation process, as shown in Figure 4. Our primary focus was to address the challenges of data scarcity and labeling effort, which are especially evident in specialized domains such as agriculture. We detail how the dataset was prepared as a foundation for generating synthetic images and the initial steps taken to apply DMs for data augmentation. This foundational setup allows us to rigorously evaluate the impact of synthetic data on downstream model performance and inform the development of our prompt-free generation pipeline.

4.3.1 Experimental Setup

The experiments were conducted using the YOLOv8 detection network across all configurations to ensure architectural consistency. The models were trained on the CottonWeedYolo dataset (Dang et al., 2023), which consists of 4150 images distributed across 14 weed species: Crabgrass, Eclipta, Goosegrass, Morningglory, Nutsedge, PalmerAmaranth, Prickly Sida, Ragweed, Sicklepod, Spottedspurge, SqurredAnoda, Swinecress, Waterhemp, and Purslane. This dataset was partitioned into 80% training, 10% validation, and 10% testing subsets. The models were trained for 10 epochs with an image size of 512×512, using an automatically determined batch size from the Ultralytics package on an RTX 4060TI GPU.

The inherent class imbalance, as shown in Figure 5, in the original dataset poses a significant challenge for detection models. Seven experimental configurations were designed to address this and systematically evaluate the role of synthetic data in mitigating the imbalance and enhancing detection performance.

- **Experiment 1:** Baseline training using the original imbalanced dataset without modifications.

- **Experiment 2:** Add 300 synthetic images per class using DreamBooth, as recommended in (Xie et al., 2024), to ensure that all classes are augmented equally. This setup is designed to provide a direct comparison with established SOTA methods for synthetic data generation and augmentation.

- **Experiment 3:** Targeted synthetic oversampling to correct the initial class imbalance. We establish a fixed target number of images for every class (e.g., matching the size of the majority class) and generate exactly the required difference for each underrepresented class to reach this threshold. The aim is to measure the impact of a perfectly balanced dataset and evaluate how the network performs when certain classes are heavily dominated by synthetic data.

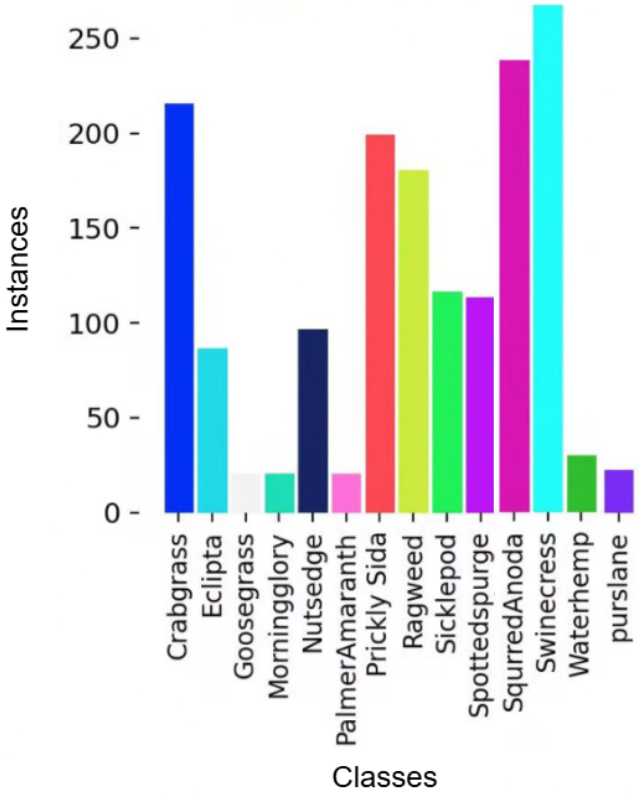


Figure 5: CottonWeedYolo dataset original class distribution (Dang et al., 2023)

- **Experiment 4:** Reduce the number of images in the more represented classes to balance the proportions in the original dataset without adding synthetic data. This experiment seeks to measure the impact of dataset balancing, this time achieved by data reduction, and to determine if significantly overrepresented classes inherently affect model performance.

- **Experiment 5:** Add 300 synthetic images per class generated by the proposed method to balance the dataset. This enables validation of our proposed generation method and Comparison of its effectiveness with the SOTA approach used in Experiment 2.

- **Experiment 6:** Reduce the original dataset to 10%, without any addition of synthetic images, to evaluate the performance on a smaller, real-world dataset. This serves to establish a baseline model performance when faced with a significantly limited amount of training data.

- **Experiment 7:** Use the 10%-reduced dataset, augmented with 300 synthetic images per class generated by the proposed method, to assess the impact of synthetic data on model performance when using a limited real dataset. This configuration is intended to demonstrate whether synthetic data can improve performance relative to the severely limited-dataset baseline (Experiment 6) and to highlight the impact

of our synthetic data more clearly in a data-scarce scenario.

In all experiments, the synthetic images generated were filtered and semi-automatically labeled using AnyLabeling (Nguyen, 2025), which incorporates SOTA object detection and segmentation algorithms. This tool includes a Segment Anything Model (SAM)2-based (Ravi et al., 2024) bounding box tool that, with manual reference points, generates both bounding boxes and segmentation masks—features that could be beneficial for future tasks.

4.3.2 Synthetic data evaluation

The fidelity of the synthetic images was evaluated using the CMMD metric (Jayasumana et al., 2024), which uses CLIP embeddings and the MMD to quantify the difference between real and generated distributions. For this analysis, we generated 300 synthetic images for each class in the dataset. Table 3 presents an ablation study that decomposes the architecture into its core components, SeeCoder and IP-Adapter, comparing them individually and in combination (our proposal, Experiment 5) against the DreamBooth baseline (Experiment 2). In the table, * denotes the prompt "crop *specie*", where *specie* corresponds to the specific weed name.

Table 3: Quantitative evaluation of synthetic image fidelity using CMMD across baselines and ablation components.

Methods	**CMMD (↓)**
Only SeeCoder (Xu et al., 2024)	2.314
Only IP-Adapter* (Ye et al., 2023)	1.647
Dreambooth* (Ruiz et al., 2023)	1.153
Ours	1.317

*Models trained with a "crop *specie*" prompt.

The CMMD metric assesses the distance between the real and synthetic image distributions, with lower values indicating greater similarity. DreamBooth achieved the best performance with a CMMD of 1.153. Our method, which combines the capabilities of IP-Adapter and SeeCoder to generate synthetic images, is closely followed by 1.317. *Unlike DreamBooth, our approach does not require retrained data, prompt engineering, or computationally intensive training.*

Individually, SeeCoder and IP-Adapter obtained CMMDs of 2.314 and 1.647, respectively. By integrating both, our method significantly improves synthetic image generation, enabling high-fidelity outputs with minimal setup and without requiring specialized hardware. With only a 0.164 difference from DreamBooth, our method offers a practical, cost-effective alternative for rapid synthetic image generation.

Regarding computational cost, our method can be executed even on devices with just a CPU or minimal GPU requirements, such as 2 GB of VRAM. In contrast, DreamBooth,

which achieves the second-best quality in synthetic image generation, requires 8 GB of VRAM for training.

Figure 6 presents a qualitative analysis of four species from the CottonWeedDet12 dataset (Dang et al., 2023). The first column shows the original images, while the following rows display synthetic images generated by different methods. We applied a DINOv2-based clustering tool to group visually similar images, forming five clusters when necessary.

However, interpreting these results requires caution. While CMMD provides a statistical measure of distribution similarity and correlates with visual realism, it can be misleading in scientific contexts. A high-fidelity score does not guarantee biological accuracy; generative models may hallucinate plausible textures that lack critical taxonomic details required for species identification, such as vein patterns or specific leaf margins (Moreno et al., 2023). Therefore, visual metrics should not be the sole basis for validation. In this study, we use the downstream object detection metric (mean Average Precision (mAP)) as the primary measure, as it implicitly assesses whether the synthetic data preserves the semantic features required for the neural network to distinguish among weed species.

The qualitative results highlight SeeCoder's inconsistency, as it often introduces artifacts despite capturing general shapes and textures. The IP-Adapter improves visual quality but produces stylized, cartoon-like outputs. DreamBooth performs best, accurately capturing shape, texture, and realism. This success is attributed to the use of similar images during training, which enables the model to learn the distinctive characteristics of each group. Our method achieves high realism, faithfully reproducing key image features while outperforming SeeCoder and IP-Adapter. Though slightly stylized, it remains a strong alternative to DreamBooth in terms of quality and fidelity.

Moreover, it is essential to emphasize that synthetic images are not guaranteed to preserve all taxonomically significant traits of the species, as models may sometimes oversimplify them. Visual examples of these failure cases are provided in Appendix A.

4.3.3 Evaluation of the Dataset Formation for Detection

In the quantitative analysis, we evaluated the results obtained in all seven experiments for each class. In this assessment, results were reported as quantitative metrics of the enhancement process's accuracy. The metric used was mAP, widely used to evaluate the performance of object detection models. The mAP was calculated for each class individually, using an Intersection over Union (IoU) threshold of 0.5, and as the average of the Average Precisions (APs) across all classes.

Figure 7 presents the mAP values obtained in each experiment, comparing the impact of different training configurations on the performance of the YOLOv8 model. Experiment 1 demonstrates lower performance than its counterparts on synthetic data. Experiment 2 achieves the best results across multiple classes, although it exhibits overall model

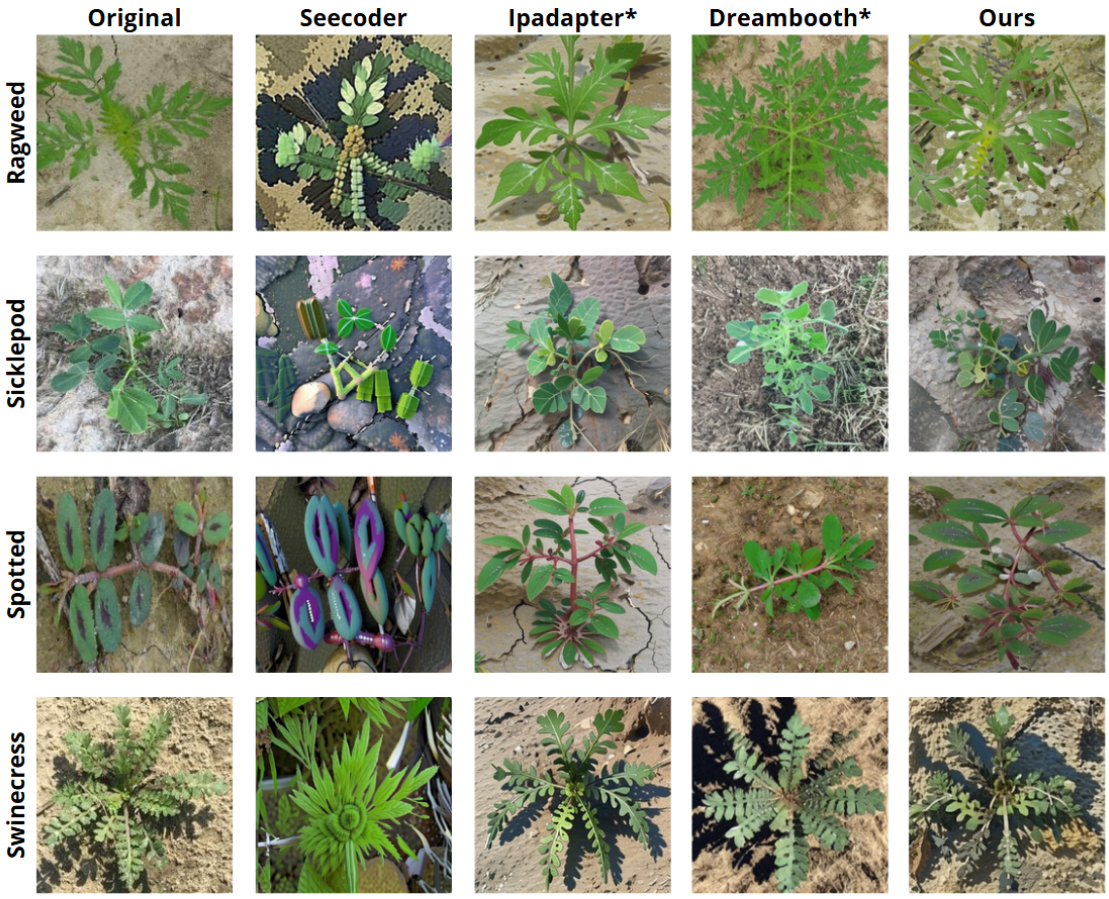


Figure 6: Qualitative Analysis of Image Generation Methods for CottonWeedDet12 Species (*indicates prompt-based methods).

instability, with low precision for some species. However, these shortcomings are minimal. In Experiment 3, despite the addition of more synthetic images, the model's overall performance decreases. This can be explained by the fact that some species have more synthetic images than others, thereby balancing the entire dataset.

On the other hand, Experiment 4 does not include synthetic images and shows significantly lower overall precision. However, when comparing Experiment 3 and Experiment 4, the model trained on synthetic images outperforms the model trained on real images overall. Finally, Experiment 5 stands out as the best overall, demonstrating the highest stability without sacrificing performance for particular species in favor of others, as observed in Experiment 2.

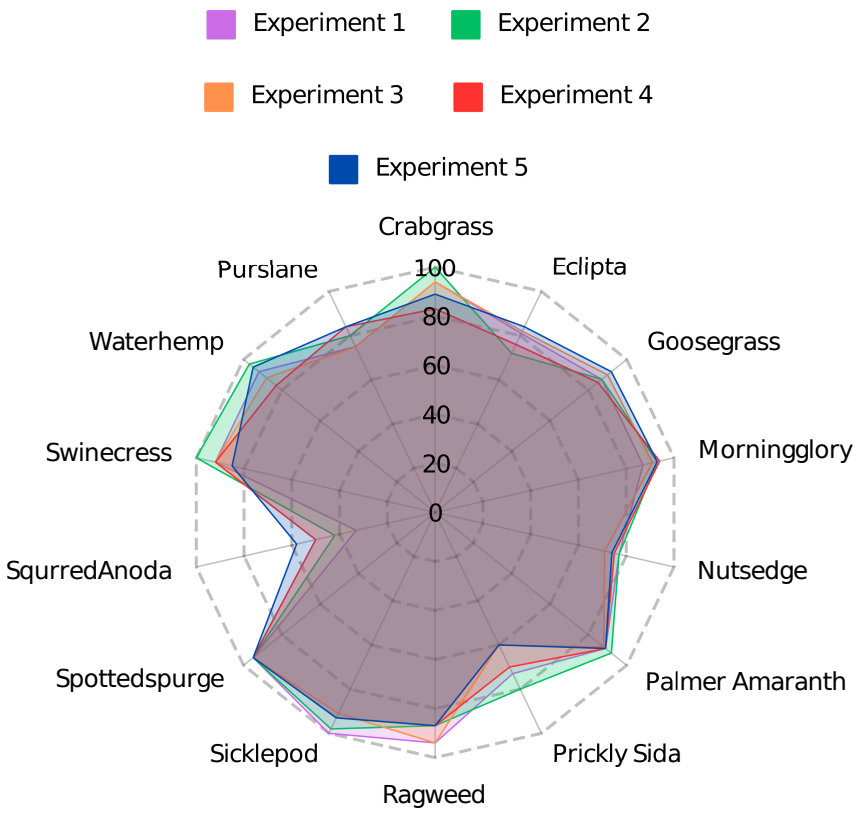


Figure 7: mAP values obtained in the first five experiments, analyzing the impact of different training configurations on the performance of the YOLOv8 model for detecting 14 weed classes.

Comparing the results of Experiments 6 and 7, as shown in Figure 8, it is observed that Experiment 7, which has more data (even if artificially generated), improves the model's overall precision. In Experiment 6, the dataset is minimal, resulting in poor performance and indicating that the model failed to understand the data correctly. On the other hand, with synthetic images, the model demonstrates some learning. However, asserting that the detection model generalizes well is still impossible, as the dataset remains quite restricted.

Thus, it can be concluded that synthetic data is valid but does not replace the need

for real data, which should be as diverse as possible. The combination of artificial and real data is essential for developing robust networks, making it crucial to prioritize quality over quantity in real data collection.

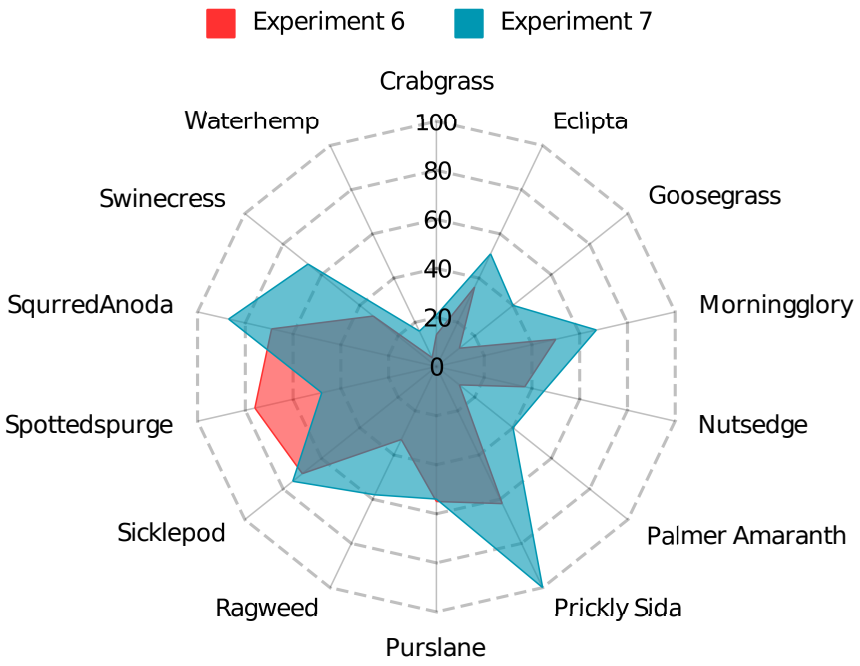


Figure 8: Comparison of mAP between Experiments 6 and 7, highlighting the impact of adding synthetic data on the detection accuracy of the 14 weed classes with the scarcity of real data.

4.4 Discussion and Limitations

This study confirms that the proposed training-free pipeline, which leverages SeeCoder and IP-Adapter, enables the generation of effective synthetic training data. The results demonstrate performance metrics comparable to computationally intensive methods, such as DreamBooth, with an CMMD of 1.317, compared with 1.153. By integrating 300 synthetic images per class in Experiment 5, we achieved the highest stability across weed species. This proves that the method is a viable and robust alternative for data augmentation, particularly suitable for resource-constrained environments where extensive model fine-tuning is not feasible.

However, despite these advantages, several limitations prevent this method from being definitive. Although the training phase is eliminated, the inference process presents a new bottleneck. The generation time of our pipeline remains higher at 8.95 s per image compared to the 3.37 s of the optimized DreamBooth generation. This latency poses a challenge for scaling the method to massive datasets. Furthermore, as observed in the qualitative analysis, the reliance on full image synthesis introduces the risk of morphological hallucinations. The model may generate plausible-looking weeds that technically satisfy the

prompt but lack specific taxonomic details required for expert-level identification.

These trade-offs highlight a crucial insight regarding the architectural components used. The experiments demonstrated that the visual encoders successfully captured the semantic essence of the weed species without prior training. The inefficiency lies primarily in the subsequent pixel reconstruction phase, which incurs high computational costs and introduces potential visual artifacts. This suggests that the extracted visual embeddings have significant representational value in their own right. Consequently, research directions might benefit from focusing on the utility of these rich semantic representations, potentially by leveraging this guidance capability without the computational overhead of generating complete pixel arrays.

5 AUTOMATED ZERO-SHOT AUGMENTATION FOR ROBUST IMAGE CLASSIFICATION

Transitioning toward the objective of maximizing automation in image generation, a pressing need arises for systems capable of automatically producing labeled images. As established in previous chapters, manual curation and prompting are bottlenecks; therefore, it is crucial to demonstrate that fully automated synthetic samples can effectively guide the training of classification models without human intervention.

A critical challenge in this landscape of data augmentation techniques is whether zero-shot synthetic data generation can enhance classification performance without requiring model fine-tuning or textual prompts. While recent work has demonstrated the potential of DMs for data augmentation, most approaches still rely on computationally expensive training or complex prompt engineering, thereby limiting their practical applicability. To address this gap, we propose ZEDD-G, a novel framework that establishes a fully automated pipeline for this task. Our approach leverages multi-image guidance within the latent space of a pre-trained DM, combining direct feature manipulation with SeeCoder-based semantic extraction to generate diverse, high-fidelity training data, as illustrated in Fig. 9.

5.1 ZEDD-G Method

5.1.1 Clustering with DINOv2

A foundational requirement for many data generation techniques, such as DreamBooth (Ruiz et al., 2023), is a curated set of visually similar images to guide the process. A significant limitation of this approach is its reliance on manual curation. This subjective and time-consuming task can lead to the omission of relevant images due to the lack of a comprehensive overview of the dataset.

To address this limitation and enable a fully automated workflow, our method begins with the automatic clustering of images. For this task, we employ DINOv2 (Oquab et al., 2023) following the methodology described in the previous chapter.

As demonstrated there and in our work (Ramos et al., 2025b), this serves as a robust substitute for manual selection. The output of this initial step is a set of well-defined sub-

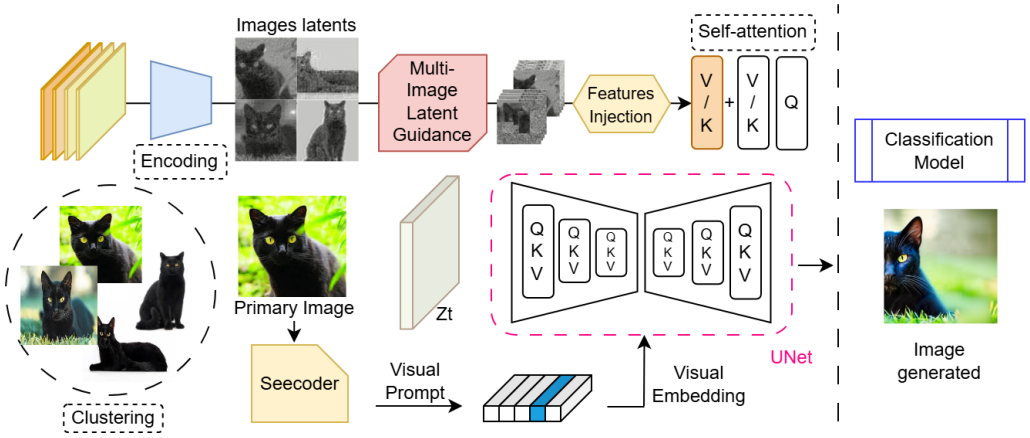


Figure 9: Overview of the ZEDD-G framework. Image clusters are formed using DINOv2. A primary image is then fed into a SeeCoder to extract visual prompts that guide the modification of latent representations from other images. These are injected into the DM to generate new samples.

groups, each containing visually coherent images. These subgroups serve as the basis for subsequent multi-image generation, constituting the first stage of our automated pipeline.

5.1.2 Visual Feature Extraction with the SeeCoder

The SeeCoder (Xu et al., 2024) is a central component of the proposed multi-image latent guidance approach, enabling the extraction and encoding of visual features directly from a set of reference images. By operating without textual prompts, the SeeCoder overcomes the vocabulary limitations typical of text-to-image models, providing a richer semantic representation that is more closely aligned with the visual content.

Its main advantage lies in its ability to integrate complex visual information from multiple images simultaneously, thereby eliminating the need for manual engineering. This process is often time-consuming and subjective. Through a detailed analysis of texture, shape, color, and spatial structure, the SeeCoder generates a unified latent embedding that cohesively synthesizes the visual diversity present in the input set.

This embedding acts as a conditioning vector in the diffusion process, guiding the generation of synthetic images that reflect the learned features from the multiple references. However, using the SeeCoder in isolation can lead to visual misalignment or loss of fidelity to the original images (Ramos et al., 2025b).

To mitigate these limitations, we propose combining SeeCoder with additional latent-space alignment mechanisms, inspired by approaches such as RIVAL (Zhang et al., 2023c).

5.1.3 Multi-Image Latent Guidance

DMs operate sequentially, with modern DDIM-based samplers typically employing around 50 iterative steps. This inherent structure enables guiding the generation process,

steering the trajectory from an initial random noise distribution to a target image distribution.

A straightforward approach to this that avoids external networks, such as adapters or additional encoders, is latent alignment. RIVAL (Zhang et al., 2023c), for instance, performs feature inversion on a single guide image to extract its latent representation. These latent features are then randomized and injected into the U-Net denoiser's self-attention layers during generation. However, this method has two key limitations: its reliance on an LLM for guidance and its inability to use multiple guide images. We propose a novel method to address these shortcomings, hypothesizing that leveraging multiple guide images is particularly beneficial for targeted data augmentation in classification problems.

Our proposed process is as follows. First, features are extracted from a set of guide images (we use four in our experiments for simplicity) as described in Section 5.1.1. One is designated as the primary guide. To enhance generative robustness, this primary image is also used as an input to the SeeCoder, which provides an additional guidance signal, the functioning of which is explained in detail in Section 5.1.2. The latent features of the primary guide image, extracted from the DM's encoder, serve as the base representation and are partitioned into blocks. The size of these blocks constitutes the first hyperparameter of our method. We observed that larger block sizes better preserve the image's high-level structural information, which is crucial for datasets in which composition is essential (e.g., medical images). This structured partitioning is a key improvement over the absolute randomization used in prior work. Each block is then randomly shuffled (i.e., permuted) according to a probability parameterized by our second hyperparameter, which ranges from 0 to 1.

For the non-primary guide images, we repurposed these same hyperparameters to control the contribution of their corresponding latent features. Specifically, using larger block sizes with a low shuffle factor promotes the blending of spatial characteristics from the secondary images with the primary one. Conversely, employing smaller block sizes and a high shuffle factor ensures that the features of the primary image are predominantly preserved. The resulting blend of features is then concatenated with the primary image's original latent representation before injection. This is a crucial step that anchors the final output to the primary guide while introducing targeted variations from the others. A detailed demonstration of the method and its parameters is provided in Figure 10.

Therefore, by fusing latent representations from multiple images, our approach generates significantly more diverse data than single-image techniques such as RIVAL (Zhang et al., 2023c), whose variations are fundamentally constrained to the characteristics of a single source.

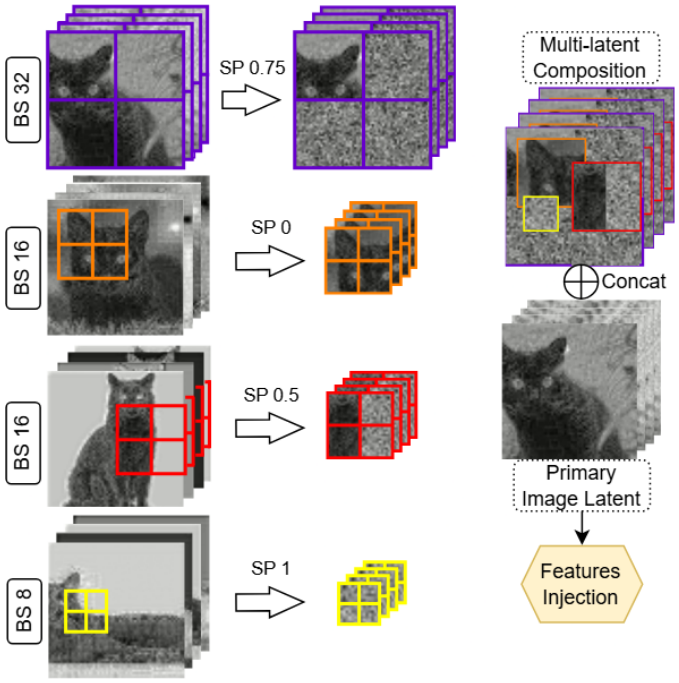


Figure 10: Visual demonstration of our multi-image latent guidance method. The figure illustrates the impact of our two key hyperparameters: Block Size (BS) and Shuffle Probability (SP). Block size and shuffle probability control the trade-off between blending features from multiple images and preserving the primary image's structure.

5.1.4 Zero-Shot Diffusion-based Data Generation

With all the aforementioned components implemented, we have established a complete pipeline for zero-shot synthetic data generation. This pipeline, named ZEDD-G, requires no prior training and can be applied directly to a dataset to augment it using diffusion methods. The overall workflow of ZEDD-G is depicted in Figure 9.

The primary image selection strategy varies with the desired augmentation scale. For substantial augmentation (e.g., a 20x increase by generating 80 images from a subgroup of 4), each of the four images serves as the principal 20 times. For smaller-scale augmentations, however, a single principal image is randomly selected from the subgroup to ensure variability.

The SeeCoder then processes the selected primary image to extract visual prompts. Simultaneously, the latent representations of the remaining guide images are modified.

Finally, these modified latents are injected into the cross-image self-attention layers of the pretrained DM, guiding the denoising process with the new features. This method enables the generation of diverse synthetic data derived from the original dataset, all without requiring any model retraining.

5.1.5 Benchmark

To evaluate our method, we adopt the benchmark proposed by Fu et al. (Fu et al., 2024). This framework evaluates performance using a ResNet-50 classifier on a diverse set of datasets, including Caltech101, the fine-grained Stanford Cars and Oxford-IIIT Pets datasets, and the Shenzhen TB medical imaging collection. The results are compared against other diffusion-based augmentation methods detailed in the benchmark. A key distinction is that these competing methods require dataset-specific fine-tuning, whereas our approach is in accordance with a zero-shot manner.

In accordance with the benchmark protocol, we expanded the datasets by the prescribed scaling factors: Oxford-IIIT Pets by 30, Stanford Cars and Caltech101 by 20, and Shenzhen TB by 10. For our generation method, we consistently used four guide images per class to generate synthetic data.

To maximize diversity, the hyperparameters of the variation process are typically sampled randomly for each generation. Block sizes are selected as multiples of 4, up to a maximum of 16, and a random-shuffle probability is applied to the feature composition. The sole exception to this rule is the generation of medical images. While larger block sizes from secondary images typically increase variability, we observed that medical images are susceptible to such variations and can lose realism. Therefore, to ensure plausible results on the Shenzhen TB dataset, we used a fixed block-size structure of $32, 8, 8, 8$, which was empirically determined to be the optimal configuration for this domain.

5.2 Experimental Results

For the benchmark evaluation, we performed data augmentation according to the protocols prescribed for each dataset. All experiments were conducted on an NVIDIA RTX 4070 Ti GPU. Our implementation is publicly available.[1].

Our experimental procedure and the results for all competing diffusion-based methods were adapted from the DreamDA benchmark (Fu et al., 2024). *It is noteworthy that, unlike our process, all these alternative approaches require computationally expensive retraining on each specific data subset.*

The results of this comparative analysis are presented in Table 5.2. We use classification accuracy as the primary evaluation metric, as it directly measures the practical utility of the generated data. This metric assesses the performance of a downstream classifier trained on the augmented dataset (real and synthetic images), providing a clear indication of the quality and effectiveness of the artificial data. To ensure a fair and unbiased evaluation, the test set was held out and remained completely separate throughout all experiments.

The quantitative results demonstrate that our method achieves highly competitive per-

[1] https://github.com/Gardiy/ZEDDG

Table 4: Comparison of data augmentation methods based on classification accuracy across multiple datasets (ResNet-50 trained from scratch and pre-trained). The benchmark is adapted from DreamDA (Fu et al., 2024); unlike our zero-shot ZEDD-G, all other methods require dataset-specific fine-tuning and prompt engineering.

ResNet-50 Stage	Method	Natural Images			Medical Images	Fine-tuning required?
		Caltech101	Cars	Pets	Shenzhen TB	
Pretrained	No data Augmentation	89.1	89.2	86.9	88.3	–
	CutMix	88.9	89.1	87.1	89.1	✓
	Random Erase	89.5	89.2	87.2	89.2	✓
	RandAugment	89.5	89.3	87.4	90.0	✓
	Stable Diffusion	87.2	92.8	87.0	90.4	✓
	SDEdit	88.6	92.5	87.8	90.7	✓
	DI	87.8	92.8	88.0	89.5	✓
	BoT	88.4	92.5	88.2	89.2	✓
	GIF-SD	86.4	93.1	87.4	88.8	✓
	DreamDA	93.1	94.5	93.2	91.1	✓
	ZEDD-G (Ours)	89.9 (+0.8)	92.3 (+3.1)	91.1 (+4.2)	90.4 (+2.1)	✗
Scratch	No data Augmentation	43.9	46.6	43.8	72.7	–
	CutMix	55.8	60.1	58.8	74.2	✓
	Random Erase	55.3	63.3	55.6	74.1	✓
	RandAugment	55.5	63.8	55.4	75.5	✓
	Stable Diffusion	76.6	81.8	76.0	79.9	✓
	SDEdit	76.2	81.3	75.7	79.7	✓
	DI	73.1	81.2	75.1	79.3	✓
	BoT	77.9	82.9	78.7	80.0	✓
	GIF-SD	65.1	75.1	65.9	77.3	✓
	DreamDA	85.3	87.1	86.5	83.5	✓
	ZEDD-G (Ours)	80.3 (+36.4)	85.1 (+38.5)	80.9 (+37.1)	81.2 (+8.5)	✗

formance, surpassing most benchmarked data generation techniques. Specifically, our method boosts accuracy by an average of over 30 percentage points (pp) for a ResNet-50 trained from scratch (with gains as high as +38.5 pp on Cars) and around 2.6 pp for a pre-trained model (peaking at +4.2 pp on Pets). This presents a visual comparison of results, showing superior performance. However, DreamDA requires training a DM on every fifth image in the dataset, a process that demands significant computational resources and often spans multiple GPUs (Zhang et al., 2023c). In stark contrast, our approach offers a fast and efficient pipeline that proceeds directly from data clustering to generation, *eliminating the need for model retraining or prompt engineering*.

To assess the qualitative performance of our method, Fig. 11 presents a visual comparison of results across the different datasets. To offer a transparent view of the method's performance, the figure intentionally includes some samples with artifacts. As shown in the sample images, our method generally produced high-quality results for most datasets. The most notable exception is the chest X-ray dataset, which exhibited the lowest visual fidelity and the highest artifact incidence. However, minor imperfections were also present, albeit to a lesser extent, in other classes. Specifically, these generated images exhibit artifacts, such as anatomically impossible bone structures and subtle domain shifts (see Appendix B for a detailed gallery of these artifacts and failure cases). This reduced visual quality correlates with the quantitative results, which also yielded the smallest gains

in accuracy, underscoring its importance as a challenging test case and reaffirming that further validation is essential for any practical application in this domain.

In contrast, the generation for other datasets proved highly effective. The visual results confirm that our multi-image latent alignment method is optimal for creating meaningful variations. For instance, we successfully generated images of the same car from different viewpoints and composed scenes containing multiple pets. We transferred stylistic features from the Caltech dataset to categories such as 'jug'. Although the figure displays a limited sample due to space constraints, the selected images are intended to be representative of the diversity observed across most classes.

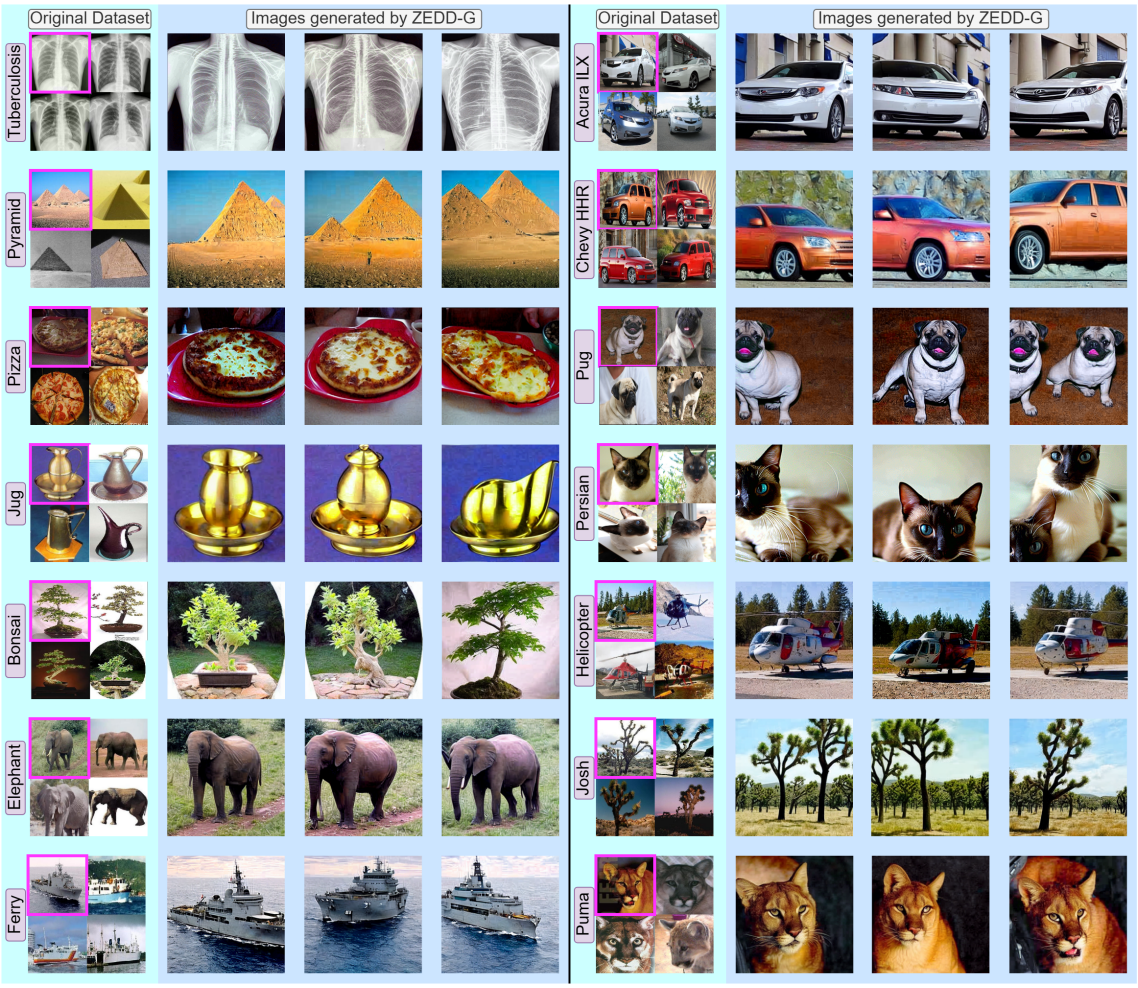


Figure 11: Qualitative evaluation of synthetic images generated by ZEDD-G across various datasets from the benchmark. Examples shown include the 'tuberculosis' class from Shenzhen TB; 'pyramid,' 'pizza,' and 'jug' from Caltech101; 'Acura ILX' and 'Chevy HHR' from Stanford Cars; and 'bulldog' and 'Persian' from Oxford-IIIT Pets. For each set, the primary guide image is highlighted in pink, with three secondary images providing additional guidance.

5.3 Discussion

It is essential to acknowledge that the latent manipulation process, while effective for feature enrichment, introduces randomness that can compromise strict visual adherence to the primary structure. As observed in the qualitative analysis, this can result in images that appear visually distorted or contain artifacts. However, for classification tasks, these high-level semantic variations prove beneficial as they force the model to learn more robust features. While this approach is sufficient for classification, tasks requiring precise spatial alignment, such as segmentation mask generation, would necessitate a more robust guidance mechanism. Nevertheless, ZEDD-G establishes a valuable precedent, demonstrating that a training-free tool works effectively across diverse domains.

In summary, we propose and validate a novel method for data augmentation that enhances classification performance across diverse real-world datasets *without requiring DM training*. Our results demonstrate that our approach achieves performance comparable to SOTA techniques, which are significantly more complex and computationally expensive to train. The primary advantage of our method lies in its fully automated, plug-and-play pipeline, which seamlessly handles everything from selecting similar images to generating synthetic data. The experiments presented were intentionally designed to evaluate the raw output of this core method, which, on its own, proves effective for data augmentation.

For future work, we have identified several promising directions. Since our method operates by editing existing images, it is inherently compatible with other synthesis techniques. We plan to integrate practical features from benchmarked methods, such as applying stylistic image filters, to enhance our approach further while maintaining its training-free nature. Additionally, we will explore post-processing methods to refine visual quality and assess their impact on classification performance. A promising approach is to implement a quality-assessment module to automatically filter out generated samples that contain the previously described artifacts.

Furthermore, we aim to expand the model's capabilities by implementing new, controllable generation rules, such as simulating varied lighting conditions or introducing specific artifacts on objects, to improve the resulting metrics further. Finally, we intend to conduct a more rigorous evaluation focused on model generalization, assessing the impact of our synthetic data on a classifier's performance on entirely new datasets, external to those used for augmentation and training. This will provide a deeper understanding of the utility of our data in real-world scenarios.

6 UW-DATAGEN: A SEMANTIC-AWARE SYNTHETIC DATA PIPELINE FOR UNDERWATER SEGMENTATION

In the previous chapter, we demonstrated that zero-shot generative pipelines, such as ZEDD-G, can effectively augment datasets for image classification by leveraging latent space guidance and visual feature extraction. However, while classification requires only global semantic coherence, the challenge escalates significantly when addressing *dense prediction tasks*. Semantic segmentation requires not only the presence of an object but also precise pixel-level alignment between the visual signal and its corresponding label mask.

Current SOTA models, including the SAM and its derivatives (Montalvo et al., 2025; Li et al., 2025; Lian et al., 2024), have shown impressive generalization capabilities. Yet, the development of robust segmenters for specialized domains remains heavily constrained by the scarcity of pixel-accurate annotations. This limitation is particularly acute in the underwater environment, where optical degradation, caused by light absorption and scattering, compromises both visual fidelity and the feasibility of manual annotation (Lian et al., 2023b).

While our previous latent-guidance approach excelled at generating diverse scenarios, it lacks the strict spatial control required to produce perfectly paired images and segmentation masks. To bridge this gap, this chapter proposes a shift from purely latent manipulation to *spatially conditioned generation*. We hypothesize that integrating ControlNet architectures with advanced semantic prompting (BLIP) enables the synthesis of highly realistic, labeled underwater imagery that adheres to complex geometries. This methodology aims to assess whether synthetic data can overcome the *domain gap* in dense prediction tasks, thereby enabling a generative model to serve as an automatic annotator for the most challenging environments.

6.1 Methodology

This section details the methodology for generating synthetic data within the challenging domain of semantic segmentation. This field fundamentally requires precise,

pixel-level annotations to enable supervised training. While various SOTA models have been developed to optimize key performance metrics—specifically mean Intersection over Union (mIoU), which evaluates the spatial overlap between the prediction and the ground truth, and Pixel Accuracy, which measures the global rate of correctly classified pixels (Manakitsa et al., 2024)—they universally suffer from data scarcity, as previously discussed.

It is widely accepted that augmenting datasets with relevant, high-quality data improves model precision, enhances evaluation metrics, and contributes to more robust and generalizable architectures (Azizi et al., 2023).

In our specific application domain, **underwater imagery**, this challenge is even more pronounced. The underwater environment is characterized by extreme data scarcity and inherent visual difficulties (e.g., poor visibility, color cast, and light scattering). Therefore, we propose a method for generating synthetic data of sufficient quality to improve the performance of underwater semantic segmentation models, including those that are not initially specialized for this environment.

To this end, we identify **ControlNet** as a prime candidate. Its primary advantage lies in its highly controllable generation process. By training a DM conditioned on segmentation masks, we can generate synthetic images that are inherently paired with their precise ground-truth labels. This process provides an automatic annotation pipeline for the newly created data, which can then be used to train and improve subsequent segmentation models.

However, based on our previous studies on synthetic data generation, practical experience indicates that several factors must be carefully evaluated. These factors include the generative model's training conditions, post-generation filtering techniques, and the optimal volume (or ratio) of synthetic images to introduce during training.

Our methodology includes a thorough evaluation of prompting strategies. As demonstrated in Chapter 5, the quality and nature of the text prompt significantly influence the quality of the generated image, which in turn impacts its utility for model training. Therefore, we conduct a study comparing several guidance methods: visual guidance (dataset v1), empty prompts (dataset v2), simple prompts (dataset v3), and complex prompts generated automatically (dataset v4).

To find the optimal trade-off between generation speed (inference time) and the quality of the synthetic data, we conduct a second set of experiments. This study evaluates the impact of the number of diffusion **inference steps** (generation cycles). While more steps generally produce higher-fidelity images, they also increase the computational cost of data generation. This evaluation aims to identify the most efficient balance that yields data capable of improving the segmentation model.

The culmination of this process is a fully automated synthetic image generation pipeline. This system integrates the optimal guidance strategy, automatic label pairing, and

pre-defined generation parameters. The pipeline is designed to run end-to-end, requiring minimal intervention. Its goal is to produce a dataset that improves the performance of any segmentation architecture in the underwater semantic segmentation domain.

For the evaluations, explicitly focusing on conditioning guidance and the generation trade-off, we employed the SOTA YOLOv11-n-seg (Khanam and Hussain, 2024) as the primary benchmark. To ensure a rigorous comparative analysis across all experimental configurations, a standardized training protocol was strictly enforced. This protocol used identical hyperparameters, including a batch size of 32 and a fixed maximum epoch limit. Furthermore, an early-stopping mechanism was systematically implemented to mitigate overfitting, ensuring that all models were evaluated at their respective optimal convergence points.

To further validate the robustness and universality of the proposed pipeline, the optimal augmentation strategy was subsequently benchmarked across a diverse set of segmentation architectures. These included TopFormer-B (Zhang et al., 2022), a representative of lightweight transformer-based models; DDRNet-23 (Yang et al., 2024b), a representative of efficient dual-resolution networks; and Lawin MiT-B3 (Yan et al., 2022), a high-performance transformer architecture. All secondary models were subjected to the same rigorous training conditions to guarantee a fair assessment of the synthetic data's impact across varying model capacities.

6.1.1 Guidance Evaluation

As introduced in the preceding section, we evaluate different types of generation guidance. ControlNet's architecture augments standard text-to-image DMs. It requires a text prompt (processed via cross-attention), in addition to the primary visual **condition** (in our case, the segmentation mask), which is injected to control the generation process. Because the underlying model is still a text-to-image framework, the text prompt remains a critical component for output quality, presenting a significant optimization challenge (Ye et al., 2023).

As part of our comparison, we first evaluate the use of SeeCoder. Given the successful application of this visual guidance method in Chapter 5 for object detection and classification tasks, it is a logical extension to adapt this approach for semantic segmentation. This experiment tests SeeCoder's ability to provide practical guidance for generating images with pixel-dense annotations, which are significantly more complex than the bounding boxes or class labels used in detection or classification.

Conversely, recognizing that the underlying DM is fundamentally text-oriented, we also evaluate the impact of automatically generated text prompts. For this task, we employ the Bootstrapping Language-Image Pre-training (BLIP) model (Li et al., 2022). BLIP is a vision-language model highly effective at generating rich, descriptive captions from images. We leverage this capability to automate the creation of complex prompts. In our

implementation, we guide this captioning process: we provide BLIP with an initial cue based on the ground-truth mask (e.g., *"A photo containing fish and coral"*) and task it with elaborating this into a detailed, descriptive prompt that emphasizes these elements based on the source image.

To complete the comparison, we also test two other conditions: 1) **Simple prompts**, which consist only of a list of class labels present in the mask (e.g., *label1, label2, ...*), and 2) **Empty prompts**, which serve as an unconditional baseline control.

6.1.2 Generation Volume Trade-off

Several SOTA studies have focused on massive data augmentation, expanding datasets by factors of up to 30x (Shipard et al., 2023; Nguyen et al., 2023; Heschl et al., 2024b; Ramos et al., 2025a). This strategy is only viable if the model performance scales with the data volume. However, this often leads to diminishing returns: model precision may not increase proportionally to the exponential growth in the volume of synthetic data. Therefore, a dedicated study is necessary to analyze our pipeline. We aim to determine an optimal augmentation factor (i.e., the ratio of synthetic-to-real data) for semantic segmentation that provides a meaningful performance benefit without incurring excessive computational overhead.

6.1.3 Final Pipeline

Building upon the empirical findings of the conditioning strategy and augmentation factor analysis, we formalize the final architecture of our synthetic data generation pipeline. The system operates as a dual-conditional framework that bridges the semantic gap in underwater imagery.

The pipeline consists of two synchronized modules. The first is the **Semantic Conditioning Module**, which processes pixel-level segmentation masks to enforce structural constraints, ensuring that the generated geometry aligns precisely with the ground-truth labels. The second is the **Textual Conditioning Module**, driven by the BLIP architecture. As demonstrated in our experiments, standard prompting is insufficient for complex scenes; therefore, we integrate BLIP to extract dense, context-aware descriptions from reference images. These detailed prompts act as a semantic guide, controlling texture synthesis and lighting conditions to minimize artifacts.

Functionally, the system combines these spatial (mask) and semantic (text) inputs to drive the diffusion-based generative process. The resulting pipeline is architected as a modular, open-source tool. It is designed to be plug-and-play, allowing researchers to seamlessly inject synthetic batches into the training loop of any standard segmentation network (e.g., YOLO, DDRNet) without modifying the downstream model architecture.

6.2 Results

In this section, we present a comprehensive evaluation of the synthetic data generation framework implemented in this study. Our analysis is structured into three key components: first, we qualitatively assess the visual fidelity of samples generated under different prompting strategies (v1–v4). Next, we quantitatively evaluate the impact of the volume of synthetic data on YOLOv11 segmentation performance to identify the optimal augmentation ratio. Finally, building upon the analysis of prompting strategies and synthetic data volume, we validate the method's generalization capabilities by benchmarking against SOTA architectures and conducting a pixel-level error analysis using confusion matrices.

6.2.1 Qualitative Evaluation of Generated Samples

We conducted a visual inspection of the synthetic samples generated by the four strategies (v1–v4) across three distinct underwater scenarios to assess semantic consistency and textural fidelity. Figure 12 presents a comparative overview of these generations.

Scenario 1: Diver and Marine Life. The first set features a complex scene with a diver and a fish in the background, characterized by low contrast and difficult object delineation. In the samples generated by v1 (No Prompt) and v2 (SeeCoder), the background details are lost mainly. In contrast, v2 recovers the fish's structural outline; both methods fail to reconstruct the diver, rendering the human category unrecognizable. The v3 (Simple Prompt) strategy introduces noticeable visual artifacts; although it preserves the underwater lighting atmosphere, the generated objects suffer from shape distortion and incoherent colorization. Conversely, v4 (Complex/BLIP Prompt) successfully reconstructs the scene with high semantic fidelity. The diver is clearly depicted exploring the reef, directly reflecting the input description: *"The image features a diver in the water, exploring a coral reef with a snorkel attached... The diver is observed swimming along the edge of the reef. There is also a large amount of seaweed surrounding the diver..."*. Due to space efficiency, the complete list of prompts used for all strategies is available in the supplementary GitHub repository[1].

Scenario 2: Close-up Marine Fauna. In the second set, focusing on a close-up view of fish, distinct performance gaps are observed. v2 (SeeCoder) demonstrates a clear advantage over v1 by generating more anatomically correct fish structures, highlighting the benefit of visual encoding over unconditional generation. However, v3 exhibits significant failures, producing chromatic aberrations and hallucinating species inconsistent with the surrounding environment. The v4 strategy generates a coherent scene with correct species representation, although a slight loss of high-frequency detail is observed in the seabed texture compared to the absolute reference.

Scenario 3: Occlusion Handling (Sea Turtle). The final set depicts a sea turtle sur-

[1] https://github.com/Gardiy/Semantic-Aware-UG

rounded by coral, presenting a challenging occlusion scenario. This case proved difficult for all methods; none of the strategies fully resolved the occlusion boundaries between the turtle's shell and the dense coral structure, indicating a limitation in current diffusion-based in-painting for heavily occluded objects. This remains an open challenge for future study. Nevertheless, in terms of overall generation fidelity and background preservation, v4 significantly outperforms the other variants, maintaining the highest level of scene coherence.

These qualitative observations suggest that the inclusion of detailed textual conditioning (v4) is critical for preserving semantic integrity in complex underwater scenes. However, it is essential to note that the generation process remains susceptible to specific artifacts, particularly in edge cases involving extreme occlusion or turbidity; a detailed visual analysis of these failure modes is provided in Appendix C. To verify whether the observed visual superiority translates into measurable performance gains in downstream tasks, we proceed to the quantitative evaluation in the following subsection.

6.2.2 Impact of Prompt Conditioning and Data Volume

We evaluated the performance of the YOLOv11-Seg model under the four distinct prompting strategies (v1–v4), incrementally increasing the synthetic data ratio from $\times 0$ (baseline) to $\times 8$. Table 5 summarizes the mIoU evolution for each version.

As illustrated in Figure 13, the strategies exhibit divergent behaviors as the volume of synthetic data increases. The baseline strategy (v1: No Prompt) and v2 (SeeCoder) show early saturation, with performance plateauing and subsequently degrading beyond the $\times 3$ factor. In contrast, the **v4 (Complex/BLIP Prompt)** strategy demonstrates superior robustness. It follows a steep upward trajectory, achieving a peak mIoU of **47.85%** at the $\times 4$ augmentation factor, representing a net improvement of $+4.93\%$ over the baseline.

It is observed that beyond the $\times 4$ threshold, performance begins to exhibit diminishing returns across all strategies. This inflection point indicates a saturation limit at which introducing additional synthetic samples yields noise or redundancy rather than discriminative features.

Table 5: Evolution of mIoU (%) on YOLOv11-Seg across varying augmentation factors for different prompting strategies. The v4 strategy achieves optimal performance at $\times 4$.

Strategy Version	**Base (0)**	**x1**	**x2**	**x3**	**x4**	**x5**	**x6**	**x7**	**x8**
v1 (No prompt)	42.92	43.10	43.30	43.45	43.20	42.90	42.50	42.00	41.50
v2 (SeeCoder)	42.92	43.15	43.35	43.50	43.25	43.00	42.60	42.10	41.60
v3 (Simple prompt)	42.92	43.60	44.40	45.10	45.35	44.90	44.20	43.60	42.90
v4 (Complex prompt)	42.92	44.50	46.10	47.20	47.85	47.10	46.40	45.60	44.90

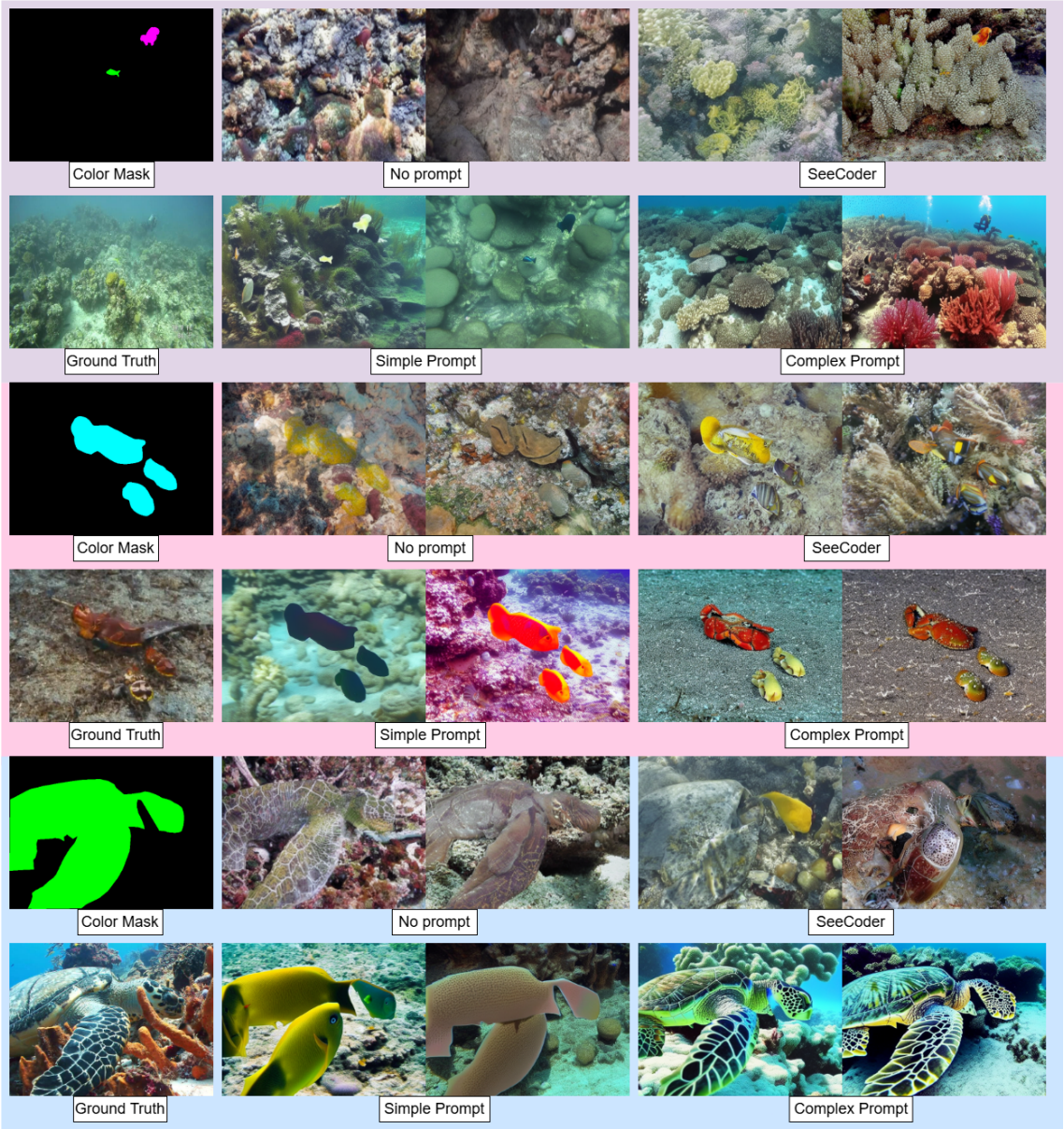


Figure 12: **Qualitative comparison of synthetic samples.** The figure displays results across three challenging scenarios: (Top) Diver and marine life, (Middle) Close-up marine fauna, and (Bottom) Sea turtle with heavy occlusion. Columns correspond to the generation strategies: v1 (No Prompt), v2 (SeeCoder), v3 (Simple Prompt), and v4 (Complex/BLIP Prompt).

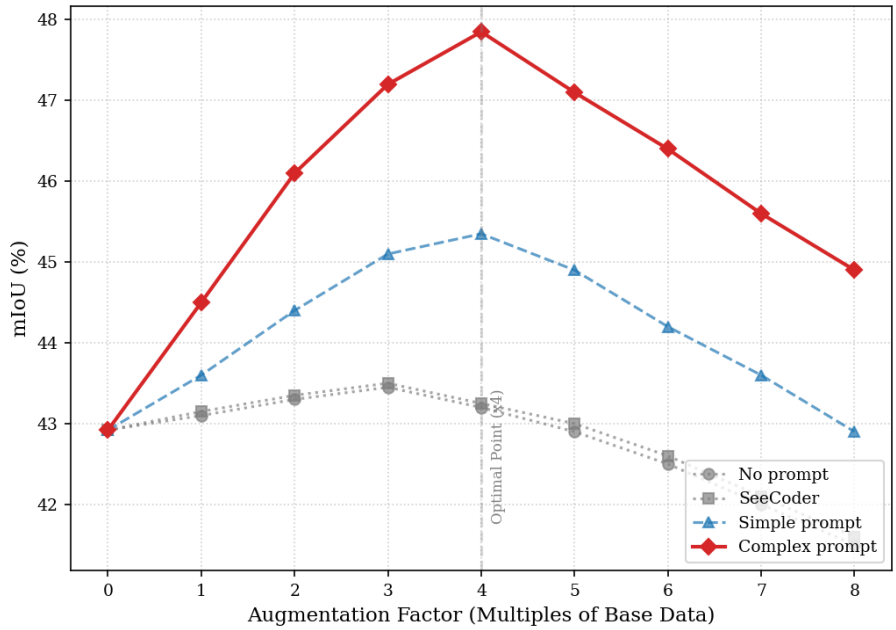


Figure 13: Performance trajectory of different augmentation strategies. Version 4 (Complex Prompt) highlights a clear "sweet spot"at the $\times 4$ factor, outperforming simpler prompting methods.

6.2.3 Cross-Architecture Generalization

To validate the generalizability of the proposed method, we applied the optimal augmentation strategy (v4 at $\times 4$ factor) to three additional architectures with varying computational complexities: TopFormer-B (lightweight), DDRNet-23 (medium), and Lawin MiT-B3 (heavy). The comparative results are presented in Table 6 and visualized in Figure 14.

The results indicate that while synthetic augmentation benefits all tested architectures, the magnitude of improvement is non-uniform. YOLOv11-Seg exhibited the highest relative gain ($+4.93\%$), suggesting a high plasticity in learning from synthetic variations. Conversely, TopFormer-B showed limited receptivity, with a marginal improvement of $+1.15\%$, despite requiring significantly longer training time (4.8 hours) than YOLO (2.3 hours). Although Lawin MiT-B3 achieved the highest absolute accuracy (53.60%), its computational cost (62 Giga Floating Point Operationss (GFLOPs), 6.9 hours) is substantially higher, highlighting the efficiency of the augmented YOLOv11 approach.

Table 6: Comparative benchmark of segmentation architectures before and after applying the v4 augmentation strategy. YOLOv11 offers the best trade-off between performance gain and training efficiency.

Model Architecture	Complexity	Training Time (h)	Base mIoU (%)	Augmented mIoU (%)	Net Improvement
YOLOv11-n-Seg (Khanam and Hussain, 2024)	Medium (6.5 GFLOP)	2.3	42.92	47.85	+4.93
TopFormer-B (Zhang et al., 2022)	Low (1.8 GFLOP)	4.8	41.50	42.65	+1.15
DDRNet-23 (Yang et al., 2024b)	Medium (5 GFLOP)	7.5	48.63	52.80	+4.17
Lawin MiT-B3 (Yan et al., 2022)	High (62 GFLOP)	13.9	50.10	53.60	+3.50

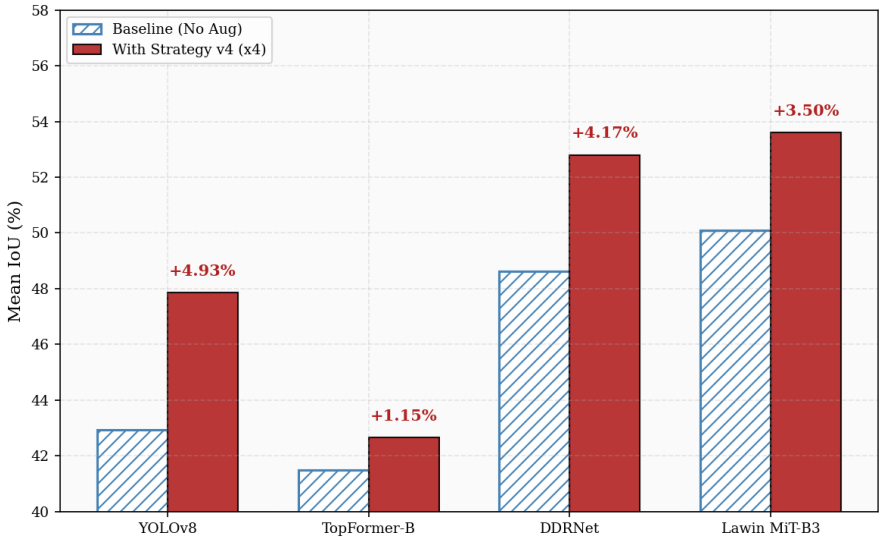


Figure 14: Impact of the v4 augmentation strategy on mIoU across different architectures. The red bars indicate the performance after augmentation.

6.2.4 Error Analysis

To further investigate the source of the performance gains, we analyzed the pixel-level classification errors using normalized confusion matrices (Figure 15). Compared with the baseline YOLOv11, the augmented model (v4) generally strengthens the diagonal elements, corresponding to an increase in the actual positive rate.

A granular inspection reveals a significant decrease in False Negatives for structurally complex classes, such as *Plants* and *Ruins*. In the baseline model, these categories were frequently misclassified as background due to texture ambiguity and low contrast. The v4 augmentation strategy effectively mitigated these errors, enabling the model to recover previously missed objects. This improved robustness in differentiating challenging classes from the background validates the semantic fidelity of the synthetic data. It is the primary driver of the increase in overall mIoU from 42.92% to 47.85%.

6.3 Discussion

The experimental results presented in this chapter provide compelling evidence that synthetic data generation is a viable strategy for overcoming data scarcity in semantic segmentation, provided that the generation pipeline is semantically aware. The comparative analysis of prompting strategies reveals a crucial insight: structural guidance (masks) alone is insufficient for complex domains. The significant performance gap between the unconditional baseline (v1) and the BLIP-enhanced strategy (v4) confirms that detailed textual conditioning is required to ”fill in”the texture and lighting details plausibly within the constraints of the segmentation mask.

Empirically, the identification of a saturation threshold at the $\times 4$ augmentation factor

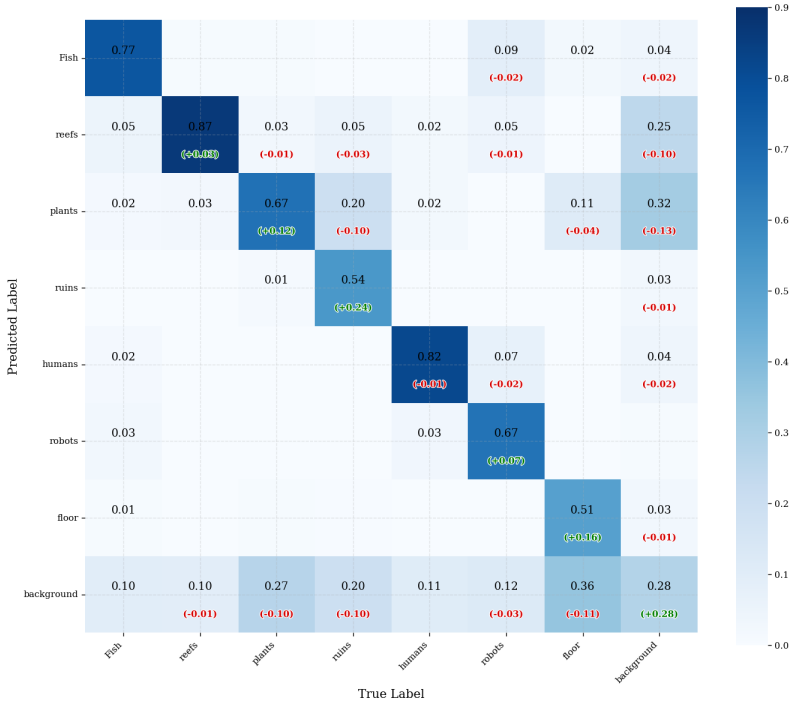


Figure 15: Impact of model refinement on per-class performance. The matrix displays the normalized recall for the optimal improved model (x4, v4), with the performance delta relative to the baseline overlaid. Green annotations (+) indicate increased diagonal accuracy, while red annotations (-) denote a reduction in misclassifications.

highlights a fundamental trade-off. While quadrupling the dataset yielded a substantial $+4.93\%$ mIoU improvement in YOLOv11, pushing beyond this ratio resulted in diminishing returns. This phenomenon suggests that while synthetic data effectively fills gaps in the feature space, it eventually reaches a ”diversity ceiling,”at which the model begins to overfit to the generative artifacts rather than learning new discriminative features. Furthermore, the qualitative analysis identified persistent challenges in handling severe occlusions (e.g., the sea turtle scenario), indicating that current diffusion inpainting still struggles with complex depth layering.

Despite these limitations, the efficiency gains are substantial. The proposed pipeline enabled a standard real-time architecture (YOLOv11) to approach the accuracy of significantly larger models, such as Lawin MiT-B3, while incurring a fraction of the computational cost of training. This finding is particularly relevant to the deployment of resource-constrained underwater vehicles (ROVs).

7 CONCLUSIONS AND FUTURE WORKS

This thesis addressed the pervasive challenge of data scarcity in computer vision, particularly within specialized domains such as precision agriculture and underwater exploration. Through the development and validation of novel generative pipelines, we demonstrated that synthetic data, when generated with semantic awareness and proper guidance, can effectively bridge the gap between limited real-world datasets and the data-hungry nature of deep learning models. The research trajectory evolved from addressing the computational bottlenecks of fine-tuning to establishing fully automated, zero-shot generation frameworks for both classification and dense prediction tasks.

7.1 General Conclusions

The investigation began by examining the inconsistency in the effectiveness of synthetic data in agricultural settings, as detailed in Chapter 4. Our experiments demonstrated that while fine-tuning methods such as DreamBooth generate high-fidelity images (CMMD = 1.153), they impose a significant computational burden. However, the introduction of our proposed pipeline, which combines IP-Adapter and SeeCoder, demonstrates that training-free architectures can achieve comparable stability and realism (CMMD of 1.317), as shown in the ablation study in Table 3. This finding was fundamental because it established that the visual features extracted by modern encoders are sufficiently robust to guide generation without requiring expensive per-class model retraining, thereby addressing the computational gap identified in the SOTA analysis.

Building on the potential of training-free guidance, Chapter 5 introduced the ZEDD-G framework, shifting the focus to fully automated classification. The results presented in Table 5.2 validated the hypothesis that latent space manipulation using multi-image guidance is a powerful tool for data augmentation. ZEDD-G achieved accuracy gains of over 30% for networks trained from scratch on diverse datasets, from natural images to medical X-rays. This confirmed that the semantic richness of pre-trained DMs can be effectively harnessed in a zero-shot manner, eliminating the need for prompt engineering and manual curation through the integration of DINOv2 clustering. The qualitative results

in Figure 11 further illustrate that, even without explicit training, the model can capture and recombine complex visual features to generate discriminative training samples.

Finally, the research addressed the most challenging task of dense prediction in Chapter 6. The transition to the underwater domain highlighted that the domain gap is fundamentally a semantic gap. As evidenced by the qualitative comparison in Figure 12, structural guidance via ControlNet alone is insufficient for complex scenes. The integration of VLMs (BLIP) to generate detailed prompts proved essential for preserving scene coherence, resulting in a 4.93% mIoU improvement for the YOLOv11 architecture (Table 5). However, a critical finding from the volume analysis in Figure 13 was the identification of a clear saturation point at the $\times 4$ augmentation factor. This indicates that increasing the volume of synthetic data without regard to its distribution yields diminishing returns, as the synthetic distribution eventually ceases to provide new information and may induce overfitting to generative artifacts.

In summary, this thesis has demonstrated that the utility of synthetic data is not merely a function of volume but of semantic fidelity and alignment. By moving from manual fine-tuning to automated latent guidance and, finally, to semantically conditioned spatial generation, we have established a set of robust, architecture-agnostic tools that improve model performance in data-scarce environments.

7.2 Future Works

The limitations identified throughout this study, particularly the inference latency discussed in Chapter 4 and the data saturation phenomenon observed in Chapter 6, open new avenues for research. The most promising direction is to move from static augmentation to dynamic, intelligent generation.

The saturation effect highlights a significant inefficiency in current approaches, where generating uniform batches of synthetic data eventually leads to redundancy. To overcome this, future work will focus on implementing a closed-loop Active Generative Learning system. Instead of randomly augmenting all classes, the generative model should be guided by the downstream classifier's uncertainty. By analyzing the entropy of model predictions on a validation set, it is possible to identify edge cases in which the model fails, such as occluded objects or specific lighting conditions. The pipeline would then dynamically generate synthetic samples targeting only these high-uncertainty regions. We hypothesize that this targeted injection of synthetic data will overcome the saturation ceiling observed in our experiments, enabling continuous performance improvements with a smaller, more information-dense synthetic dataset.

Furthermore, while this thesis focused on frame-by-frame processing, the primary application domains of agricultural robotics and underwater ROVs operate in continuous time. A critical limitation of current diffusion-based augmentation is the lack of temporal

consistency. Future research must explore Video DMs or optical-flow constraints within the latent space to generate coherent video sequences. This advancement would enable training of tracking algorithms and action recognition models, which are essential for next-generation autonomous robotic systems.

Finally, to fully realize the potential of these systems in field robotics, future work should investigate Knowledge Distillation techniques to compress these generative pipelines. The goal is to enable on-device generation, allowing a robot to encounter a novel object, generate synthetic variations locally to update its internal model, and adapt to new environments in real-time without relying on cloud connectivity or heavy GPU infrastructure.

BIBLIOGRAPHY

Alaa, A., Van Breugel, B., Saveliev, E. S., and Van Der Schaar, M. (2022). How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning*, pages 290–306. PMLR.

Aljohani, A. and Alharbe, N. (2022). Generating synthetic images for healthcare with novel deep pix2pix gan. *Electronics*, 11(21):3470.

Alomar, K., Aysel, H. I., and Cai, X. (2023). Data augmentation in classification and segmentation: A survey and new strategies. *Journal of Imaging*, 9(2):46.

Anonymous, C. (2023). Comfyui: Open source node-based application for generative ai. https://github.com/comfyanonymous/ComfyUI.

Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., and Fleet, D. J. (2023). Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*.

Bansal, A., Chu, H.-M., Schwarzschild, A., Sengupta, S., Goldblum, M., Geiping, J., and Goldstein, T. (2023). Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 843–852.

Cao, H., Tan, C., Gao, Z., Xu, Y., Chen, G., Heng, P., and Li, S. Z. (2024). A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):2814–2830.

Chen, P., Liu, S., Zhao, H., Wang, X., and Jia, J. (2020). Gridmask data augmentation. *arXiv preprint arXiv:2001.04086*.

Cho, J., Nam, G., Kim, S., Yang, H., and Kwak, S. (2023). Promptstyler: Prompt-driven style generation for source-free domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15702–15712.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proce-*

edings of the IEEE conference on computer vision and pattern recognition, pages 8789–8797.

Chung, J., Hyun, S., and Heo, J.-P. (2024). Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8795–8805.

Dang, F., Chen, D., Lu, Y., and Li, Z. (2023). Yoloweeds: A novel benchmark of yolo object detectors for multi-class weed detection in cotton production systems. *Computers and Electronics in Agriculture*, 205:107655.

de Deijn, R., Batra, A., Koch, B., Mansoor, N., and Makkena, H. (2024). Reviewing fid and sid metrics on generative adversarial networks. *arXiv preprint arXiv:2402.03654*.

Deng, B. (2023). Stable diffusion for data augmentation in coco and weed datasets. *arXiv preprint arXiv:2312.03996*.

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

Feng, W., He, X., Fu, T.-J., Jampani, V., Akula, A., Narayana, P., Basu, S., Wang, X. E., and Wang, W. Y. (2022). Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*.

Fu, Y., Chen, C., Qiao, Y., and Yu, Y. (2024). Dreamda: Generative data augmentation with diffusion models. *arXiv preprint arXiv:2403.12803*.

Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., and Bau, D. (2023). Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2426–2436.

Ghojogh, B. and Ghodsi, A. (2024). Diffusion models: Tutorial and survey.

Guo, J., Zhao, J., Du, C., Wang, Y., Ge, C., Ni, Z., Song, S., Shi, H., and Huang, G. (2024). Everything to the synthetic: Diffusion-driven test-time adaptation via synthetic-domain alignment. *arXiv preprint arXiv:2406.04295*.

He, F., Li, G., Si, L., Yan, L., Hou, S., Dong, H., and Li, F. (2024). Cartoondiff: Training-free cartoon image generation with diffusion transformer models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3825–3829. IEEE.

Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-Or, D. (2022). Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Heschl, A., Murillo, M., Najafian, K., and Maleki, F. (2024a). Synthset: Generative diffusion model for semantic segmentation in precision agriculture. *arXiv preprint arXiv:2411.03505*.

Heschl, A., Murillo, M., Najafian, K., and Maleki, F. (2024b). Synthset: generative diffusion model for semantic segmentation in precision agriculture. In *European Conference on Computer Vision*, pages 267–283. Springer.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Islam, K., Zaheer, M. Z., Mahmood, A., and Nandakumar, K. (2024a). Diffusemix: Label-preserving data augmentation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27621–27630.

Islam, K., Zaheer, M. Z., Mahmood, A., Nandakumar, K., and Akhtar, N. (2024b). Genmix: effective data augmentation with generative diffusion model image editing. *arXiv preprint arXiv:2412.02366*.

Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., and Kumar, S. (2024). Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9307–9315.

Kaneko, T., Kameoka, H., Tanaka, K., and Hojo, N. (2019). Stargan-vc2: Rethinking conditional methods for stargan-based voice conversion. *arXiv preprint arXiv:1907.12279*.

Kang, M., Zhang, R., Barnes, C., Paris, S., Kwak, S., Park, J., Shechtman, E., Zhu, J.-Y., and Park, T. (2024). Distilling diffusion models into conditional gans. In *European Conference on Computer Vision*, pages 428–447. Springer.

Khanam, R. and Hussain, M. (2024). Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*.

Khanna, S., Liu, P., Zhou, L., Meng, C., Rombach, R., Burke, M., Lobell, D., and Ermon, S. (2023). Diffusionsat: A generative foundation model for satellite imagery. *arXiv preprint arXiv:2312.03606*.

Kwon, G. and Ye, J. C. (2022). Diffusion-based image translation using disentangled style and content representation. *arXiv preprint arXiv:2209.15264*.

Lee, H., Lee, H., Gye, S., and Kim, J. (2025). Beta sampling is all you need: Efficient image generation strategy for diffusion models using stepwise spectral analysis. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4215–4224. IEEE.

Li, A. C., Prabhudesai, M., Duggal, S., Brown, E., and Pathak, D. (2023a). Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2206–2217.

Li, H., Lian, S., Li, Z., Cong, R., and Kwong, S. (2025). Uwsam: Segment anything model guided underwater instance segmentation and a large-scale benchmark dataset. *arXiv preprint arXiv:2505.15581*.

Li, J., Li, D., Xiong, C., and Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Li, S., Hu, T., Khan, F. S., Li, L., Yang, S., Wang, Y., Cheng, M.-M., and Yang, J. (2023b). Faster diffusion: Rethinking the role of unet encoder in diffusion models. *CoRR*.

Li, Y., Guo, J., Qiu, H., Chen, F., and Zhang, J. (2023c). Denoising diffusion probabilistic models and transfer learning for citrus disease diagnosis. *Frontiers in plant science*, 14:1267810.

Lian, L., Li, B., Yala, A., and Darrell, T. (2023a). Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*.

Lian, S., Li, H., Cong, R., Li, S., Zhang, W., and Kwong, S. (2023b). Watermask: Instance segmentation for underwater imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1305–1315.

Lian, S., Zhang, Z., Li, H., Li, W., Yang, L. T., Kwong, S., and Cong, R. (2024). Diving into underwater: Segment anything model guided underwater salient instance segmentation and a large-scale dataset. *arXiv preprint arXiv:2406.06039*.

Lim, J. H., Kovachki, N. B., Baptista, R., Beckham, C., Azizzadenesheli, K., Kossaifi, J., Voleti, V., Song, J., Kreis, K., Kautz, J., et al. (2023). Score-based diffusion models in function space. *arXiv preprint arXiv:2302.07400*.

Lu, S., Liu, Y., and Kong, A. W.-K. (2023). Tf-icon: Diffusion-based training-free cross-domain image composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2294–2305.

Manakitsa, N., Maraslidis, G. S., Moysis, L., and Fragulis, G. F. (2024). A review of machine learning and deep learning for object detection, semantic segmentation, and human action recognition in machine and robotic vision. *Technologies*, 12(2):15.

Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Montalvo, J., García-Martín, Á., Carballeira, P., and SanMiguel, J. C. (2025). Unsupervised class generation to expand semantic segmentation datasets. *Journal of Imaging*, 11(6):172.

Moreno, H., Gómez, A., Altares-López, S., Ribeiro, A., and Andújar, D. (2023). Analysis of stable diffusion-derived fake weeds performance for training convolutional neural networks. *Computers and Electronics in Agriculture*, 214:108324.

Mori, N., Naito, H., and Hosoi, F. (2024). Application of a latent diffusion model to plant disease detection by generating unseen class images. *AgriEngineering*, 6(4).

Muhammad, A., Salman, Z., Lee, K., and Han, D. (2023). Harnessing the power of diffusion models for plant disease image augmentation. *Frontiers in Plant Science*, 14:1280496.

Nguyen, Q., Vu, T., Tran, A., and Nguyen, K. (2023). Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. *Advances in Neural Information Processing Systems*, 36:76872–76892.

Nguyen, V. A. (2025). Anylabeling - effortless data labeling with ai support. Available at: https://github.com/vietanhdev/anylabeling. Accessed: Jan. 19, 2025. Licensed under GPL-3.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. (2021). Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. (2023). Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Pham, K. T., Chen, J., and Chen, Q. (2024). Tale: Training-free cross-domain image composition via adaptive latent manipulation and energy-guided optimization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3160–3169.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from

natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Rai, N., Schumann, A. W., and Boyd, N. (2025). Phytosynth: Leveraging multi-modal generative models for crop disease data generation with novel benchmarking and prompt engineering approach. *arXiv preprint arXiv:2505.01823*.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

Ramos, J. D. G., Schein, T. T., Brião, S. L., Oliveira, F. G., and Drews, P. L. J. (2025a). Zedd-g: Zero-shot diffusion-based data generation for improving classification tasks. *2025 38th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 1–6.

Ramos, J. D. G., Schein, T. T., Gomes, L. S., Brião, S. L., Kappel, K. S., Borges, P. V. K., Oliveira, F. G., Da S. Guerra, R., and Drews, P. L. J. (2025b). Impact of synthetic data from diffusion models on weed detection performance. In *2025 Brazilian Conference on Robotics (CROS)*, volume 1, pages 1–6.

Ranario, E., Lundqvist, L., Yun, H., Bailey, B. N., and Earles, J. M. (2025). Agile: A diffusion-based attention-guided image and label translation for efficient cross-domain plant trait identification. *arXiv preprint arXiv:2503.22019*.

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al. (2024). Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. (2023). Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510.

Sabour, A., Fidler, S., and Kreis, K. (2024). Align your steps: Optimizing sampling schedules in diffusion models. *arXiv preprint arXiv:2404.14507*.

Sahu, Y. K., Mukhopadhyay, A., Kumar, G., Kumar, A., and Biswas, P. (2024). A comparative study on image translation gan models to improve object detection accuracy in low-resource domains. In *2024 International Conference on Vehicular Technology and Transportation Systems (ICVTTS)*, volume 1, pages 1–6. IEEE.

Samuel, D., Ben-Ari, R., Raviv, S., Darshan, N., and Chechik, G. (2024). Generating images of rare concepts using pre-trained diffusion models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5):4695–4703.

Saragih, D. G., Hibi, A., and Tyrrell, P. N. (2024). Using diffusion models to generate synthetic labeled data for medical image segmentation. *International journal of computer assisted radiology and surgery*, 19(8):1615–1625.

Sarıyıldız, M. B., Alahari, K., Larlus, D., and Kalantidis, Y. (2023). Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8011–8021.

Sauer, A., Lorenz, D., Blattmann, A., and Rombach, R. (2024). Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103. Springer.

Shipard, J., Wiliem, A., Thanh, K. N., Xiang, W., and Fookes, C. (2023). Diversity is definitely needed: Improving model-agnostic zero-shot classification via stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 769–778.

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

Song, J., Meng, C., and Ermon, S. (2020). Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, Y., Dhariwal, P., Chen, M., and Sutskever, I. (2023). Consistency models. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 32211–32252. PMLR.

Sousa, T., Ries, B., and Guelfi, N. (2024). Data augmentation in earth observation: A diffusion model approach. *arXiv preprint arXiv:2406.06218*.

Sportelli, M., Apolo-Apolo, O. E., Fontanelli, M., Frasconi, C., Raffaelli, M., Peruzzi, A., and Perez-Ruiz, M. (2023). Evaluation of yolo object detectors for weed detection in different turfgrass scenarios. *Applied Sciences*, 13(14):8502.

Tang, D., Cao, X., Wu, X., Li, J., Yao, J., Bai, X., Jiang, D., Li, Y., and Meng, D. (2024). Aerogen: enhancing remote sensing object detection with diffusion-driven data generation. *arXiv preprint arXiv:2411.15497*.

Taylor, L. and Nitschke, G. (2018). Improving deep learning with generic data augmentation. In *2018 IEEE symposium series on computational intelligence (SSCI)*, pages 1542–1547. IEEE.

Trabucco, B., Doherty, K., Gurinas, M., and Salakhutdinov, R. (2023). Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*.

Tsai, Y.-Y., Chen, F.-C., Chen, A. Y., Yang, J., Su, C.-C., Sun, M., and Kuo, C.-H. (2024). Gda: Generalized diffusion for robust test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23242–23251.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Nair, D., Paul, S., Berman, W., Xu, Y., Liu, S., and Wolf, T. (2022). Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers.

Walawalkar, D., Shen, Z., Liu, Z., and Savvides, M. (2020). Attentive cutmix: An enhanced data augmentation approach for deep learning based image classification. *arXiv preprint arXiv:2003.13048*.

Wang, J., Li, X., Zhang, J., Xu, Q., Zhou, Q., Yu, Q., Sheng, L., and Xu, D. (2023a). Diffusion model is secretly a training-free open vocabulary semantic segmenter.

Wang, P., Zhang, H., Zhang, Z., Chen, S., Ma, Y., and Qu, Q. (2024). Diffusion models learn low-dimensional distributions via subspace clustering. *arXiv preprint arXiv:2409.02426*.

Wang, R., Yang, Y., Qian, Z., Zhu, Y., and Wu, Y. (2023b). Diffusion in diffusion: Cyclic one-way diffusion for text-vision-conditioned generation. *arXiv preprint arXiv:2306.08247*.

Wang, Z., Pang, T., Du, C., Lin, M., Liu, W., and Yan, S. (2023c). Better diffusion models further improve adversarial training. In *International conference on machine learning*, pages 36246–36263. PMLR.

Wu, W., Zhao, Y., Shou, M. Z., Zhou, H., and Shen, C. (2023). Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. *Proc. Int. Conf. Computer Vision (ICCV 2023)*.

Xie, E., Yao, L., Shi, H., Liu, Z., Zhou, D., Liu, Z., Li, J., and Li, Z. (2023). Difffit: Unlocking transferability of large diffusion models via simple parameter-efficient fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4230–4239.

Xie, J., Li, W., Li, X., Liu, Z., Ong, Y. S., and Loy, C. C. (2024). Mosaicfusion: Diffusion models as data augmenters for large vocabulary instance segmentation. *International Journal of Computer Vision*, pages 1–20.

Xie, S., Hu, C., Bagavathiannan, M., and Song, D. (2021). Toward robotic weed control: detection of nutsedge weed in bermudagrass turf using inaccurate and insufficient training data. *IEEE Robotics and Automation Letters*, 6(4):7365–7372.

Xu, X., Guo, J., Wang, Z., Huang, G., Essa, I., and Shi, H. (2024). Prompt-free diffusion: Taking"text"out of text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8682–8692.

Yan, H., Zhang, C., and Wu, M. (2022). Lawin transformer: Improving semantic segmentation transformer with multi-scale representations via large window attention. *arXiv preprint arXiv:2201.01615*.

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., and Zhao, H. (2024a). Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911.

Yang, L., Xu, X., Kang, B., Shi, Y., and Zhao, H. (2023). Freemask: Synthetic images with dense annotations make stronger segmentation models. In *NeurIPS*.

Yang, Z., Bi, F., Hou, X., Zhou, D., and Wang, Y. (2024b). Ddrnet: Dual-domain refinement network for remote sensing image semantic segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Ye, H., Zhang, J., Liu, S., Han, X., and Yang, W. (2023). Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Zand, M., Etemad, A., and Greenspan, M. (2023). Diffusion models with deterministic normalizing flow priors. *arXiv preprint arXiv:2309.01274*.

Zeng, W. (2024). Image data augmentation techniques based on deep learning: A survey. *Mathematical Biosciences and Engineering*, 21(6):6190–6224.

Zhang, L., Rao, A., and Agrawala, M. (2023a). Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847.

Zhang, L., Rao, A., and Agrawala, M. (2023b). Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847.

Zhang, W., Huang, Z., Luo, G., Chen, T., Wang, X., Liu, W., Yu, G., and Shen, C. (2022). Topformer: Token pyramid transformer for mobile semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12083–12093.

Zhang, Y., Xing, J., Lo, E., and Jia, J. (2023c). Real-world image variation by aligning diffusion inversion chain. *Advances in Neural Information Processing Systems*, 36:30641–30661.

Zhao, S., Chen, D., Chen, Y.-C., Bao, J., Hao, S., Yuan, L., and Wong, K.-Y. K. (2023). Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:11127–11150.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.

A QUALITATIVE ANALYSIS OF GENERATIVE FAILURES IN WEED SYNTHESIS

This appendix provides a critical visual examination of the limitations encountered during the synthetic generation process for the CottonWeedYolo (Dang et al., 2023) dataset. While quantitative metrics such as CMMD suggested high distributional similarity, a granular inspection reveals specific generative artifacts that may not be captured by global statistics but are critical for agricultural precision.

Figure 16 illustrates six representative classes where these failure modes are most evident. To facilitate comparison, the original reference images used to guide the generation are displayed alongside the synthetic outputs. We identified three primary categories of generation errors:

1. **Background Pattern Repetition:** In several samples, the DM struggles to synthesize naturalistic soil variability. Instead, it produces repetitive, tile-like textures in the background. This suggests that the model occasionally overfits to high-frequency texture details in the reference images rather than learning the terrain's stochastic nature.

2. **Morphological Inconsistency:** A divergence is observed between the generated leaf structures and the biological reality of the species. While the model successfully generates "plant-like" features, it often fails to preserve specific taxonomic traits such as leaf venation, margin serration, or phyllotaxy (leaf arrangement). The synthetic leaves appear visually distinct from the Reference, creating a "generic weed" that may confuse a classifier trained on fine-grained features.

3. **Semantic Domain Shift (The 'Ragweed' Case):** A notable hallucination occurs specifically within the *Ragweed* class. As shown in the figure, the generated leaves exhibit a morphological structure that deviates from that of the target species, bearing a strong visual resemblance to *Cannabis sativa* or other distinct flora. This likely indicates a bias in the pre-trained foundation model (Stable Diffusion), which may have a stronger internal representation of common internet imagery (e.g., can-

nabis) than of specific agricultural weeds, leading the model to “drift” toward the more familiar concept during the denoising process.

It is essential to acknowledge that the evaluation of these synthetic images was conducted from a computer vision perspective. No formal botanical validation was performed during this study due to resource constraints.

However, given the observed morphological discrepancies, we identify this as a crucial step for future research. To deploy these generative models in real-world agronomy, it is essential to establish a validation loop involving a panel of expert agronomists. Such a team would be tasked with assessing the generated samples not only for visual realism but also for taxonomic accuracy, ensuring that the synthetic data preserves the distinguishing biological features necessary for reliable species identification.

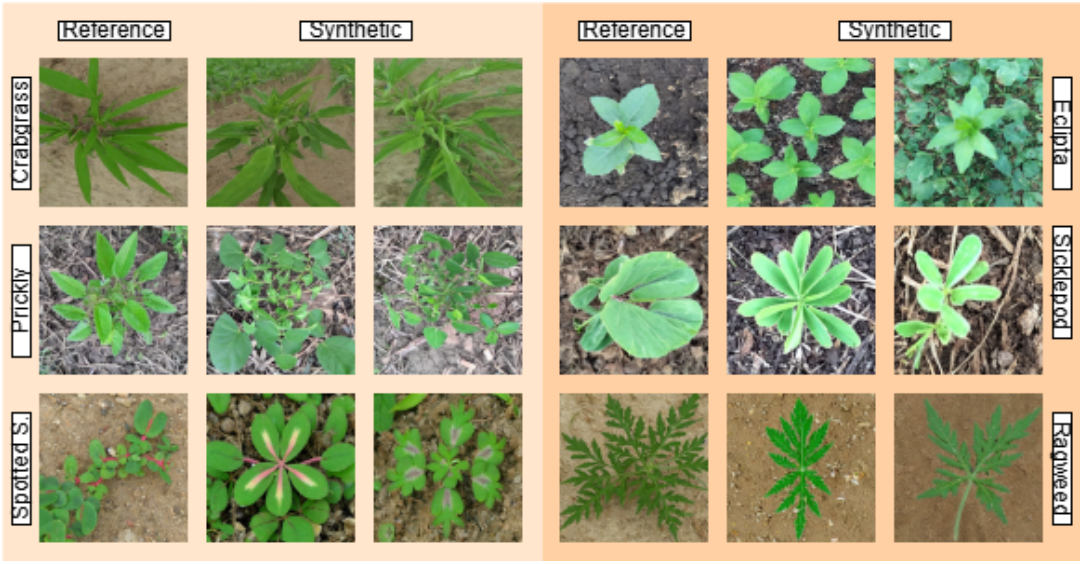


Figure 16: Visual analysis of failure cases across six classes of the CottonWeedYolo dataset (Dang et al., 2023). The figure contrasts the **Reference** images (top of each block) with the **Synthetic** outputs (bottom), highlighting artifacts such as background repetition, loss of leaf morphology, and semantic drift in the ‘Ragweed’ class.

B VISUAL LIMITATIONS OF MULTI-IMAGE LATENT GUIDANCE IN ZEDD-G

This appendix presents a detailed qualitative analysis of the failure modes of the ZEDD-G zero-shot generation framework. As discussed in Chapter 5, our method manipulates the latent space by shuffling and blending feature blocks from multiple reference images without fine-tuning the model. While this strategy is effective for transferring high-level semantic features for classification, the stochastic nature of the latent mixing can occasionally disrupt the spatial coherence required for photorealism.

Figure 17 presents a grid of samples generated from the Caltech-101 dataset. The grid is structured as a 3×6 matrix, displaying pairs of generated images for nine distinct classes. The organization proceeds from left to right: **Row 1** contains the classes *Ant, Chandelier, and Chair*; **Row 2** displays *Emu, Dolphin, and Motorbike*; and **Row 3** shows *Octopus, Helicopter, and Llama*. A close examination reveals distinct artifact patterns that correlate strongly with the object's structural complexity.

The most severe artifacts are observed in the biological categories (*Ant, Emu, Dolphin, Octopus, Llama*). Because biological forms adhere to strict anatomical rules regarding symmetry and limb arrangement, the random permutation of latent blocks often results in catastrophic morphological failures. For instance, the *Ant* and *Octopus* samples exhibit disjointed limbs and textural noise that defies biological structure. Similarly, the *Emu*, *Dolphin*, and *Llama* classes suffer from severe distortions, where the guidance from the SeeCoder seemingly conflicts with the latent shuffling, producing entities that possess the correct texture (fur, skin) but lack a coherent skeletal structure. These "nightmarish" artifacts are visually conspicuous and represent the significant limitation of unsupervised latent guidance in organic domains.

In contrast, rigid, artificial objects (*Chair, Motorbike, Helicopter*) tend to retain better global coherence, though they are not free from defects. The artifacts here manifest primarily as geometric impossibilities rather than textural chaos. The *Chair* and *Motorbike* samples display unrealistic curves or impossible perspectives. At the same time, pixelated streaks and vertical noise are visible in classes like the *Helicopter*, likely resulting from the block-based shuffling mechanism disrupting the continuity of straight lines in

the latent space.

The most extreme failure case is observed in the *Chandelier* class (Row 1, middle pair). Here, the generation fails to produce any recognizable object, yielding an abstract, noisy composition. This "semantic collapse" suggests that for visually complex objects with intricate high-frequency details (e.g., crystal ornaments), the latent blending process may destroy the structural integrity of features, rendering the guidance ineffective.

It is important to note that despite these visual defects, the quantitative results in Chapter 5 demonstrate that these images still contribute positively to classification accuracy. This suggests that the generated samples, while visually flawed, effectively capture discriminative features (e.g., color histograms, texture patterns) that assist the classifier. However, to deploy such a system in production environments that require visual intelligibility, this analysis confirms the urgent need for a smart filtering stage. As proposed in the Future Works (Chapter 7), integrating an uncertainty-based Active Learning loop would allow the system to automatically identify and discard these catastrophic failure cases, ensuring that only semantically and structurally coherent samples are added to the training set.

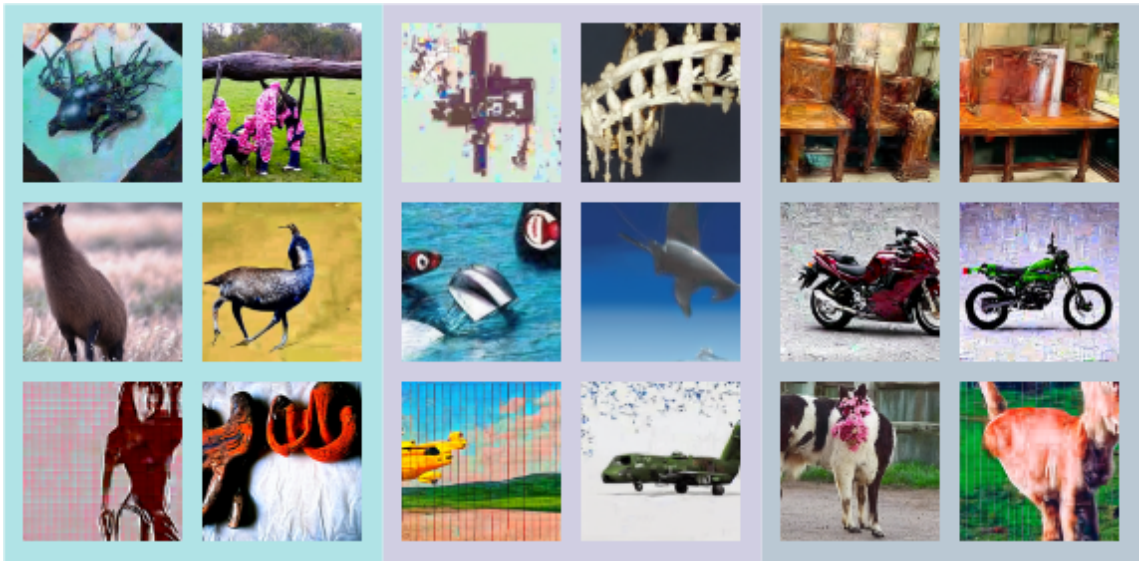


Figure 17: Gallery of failure cases in ZEDD-G using Caltech-101 classes. The grid is arranged in 3 rows and 6 columns, with two samples per class. **Row 1:** Ant, Chandelier, Chair. **Row 2:** Emu, Dolphin, Motorbike. **Row 3:** Octopus, Helicopter, Llama. The visual analysis highlights a clear degradation in biological realism compared to rigid objects, with extreme semantic collapse observed in the Chandelier class.

C QUALITATIVE ANALYSIS OF ARTIFACTS IN UNDERWATER SEGMENTATION GENERATION

This appendix provides a critical examination of the visual anomalies and failure cases encountered during the synthetic generation of underwater imagery for semantic segmentation. While the quantitative metrics presented in Chapter 6 demonstrate a net improvement in model performance, a visual inspection reveals artifacts characteristic of diffusion-based synthesis in this domain. Figure 18 illustrates a representative set of these generated samples, highlighting the discrepancy between spatial guidance and textural realism.

A prominent artifact observed across multiple samples is a spectral anomaly characterized by an over-stimulation of the red channel. In underwater environments, red light is the first wavelength to be absorbed, so natural deep-water imagery typically lacks strong reddish tones. However, as seen in several generated samples, the model frequently hallucinates vibrant red hues or applies a reddish color cast to the entire scene. This suggests a domain bias in the pre-trained DM, which likely prioritizes the general statistical distribution of terrestrial images over the specific physical constraints of the underwater optical model, despite the prompt conditioning.

Regarding structural integrity, a distinct decoupling between shape and content is evident. The ControlNet guidance successfully enforces the correct global geometry; the object silhouettes and the overall scene layout adhere strictly to the input masks. However, the internal texture generation often fails to match this spatial precision. We observe instances of semantic hollowness, particularly in the backgrounds, where the model generates sparse, featureless seabeds that lack the complex, high-frequency textures (e.g., sand ripples or coral debris) typical of real ocean floors. Additionally, there are instances of morphological hallucinations in which rock formations or marine structures appear architecturally artificial, resembling artificial ruins or impossible geometric shapes rather than natural geological features.

Furthermore, despite using the advanced BLIP-based prompting strategy (v4) to contextualize the scene, prompt injection sometimes introduces artificial patterning. In an attempt to satisfy the underwater keyword, the model occasionally overlays exaggera-

ted caustic patterns or bubbles that do not align with the scene's lighting, resulting in a stylized, *painted* look rather than a photorealistic one. These observations reinforce the conclusion that prompt engineering alone cannot fully resolve domain shifts. Consequently, this qualitative evidence strongly supports the proposal for an Uncertainty-Driven Active Learning loop, as discussed in the Future Works, to automatically detect and filter out these spectrally and texturally inconsistent samples before they enter the training pipeline.

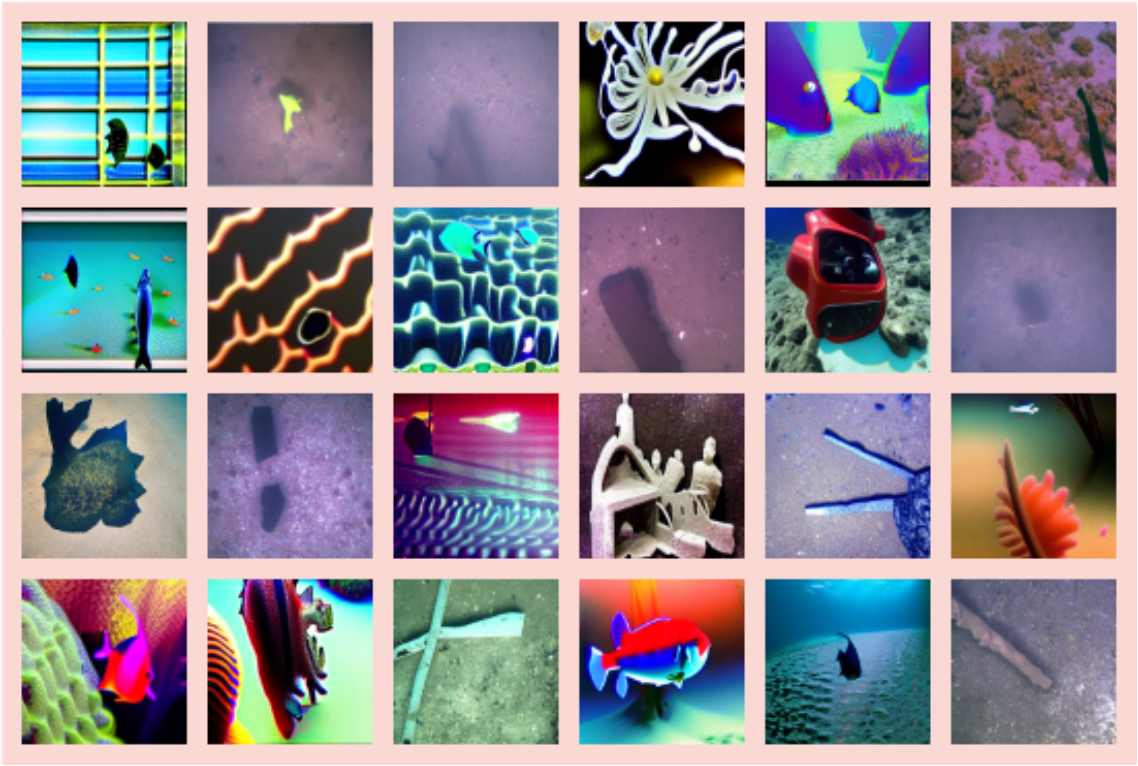


Figure 18: Visual analysis of generative artifacts in the underwater domain. The samples demonstrate correct spatial alignment (guided by ControlNet) but exhibit characteristic failures, including unnatural red color casts, sparse background textures, and artificial morphological patterns on marine objects.