HAW HAMBURG

MASTER THESIS
Mario de Jesus da Graca

Leveraging Diffusion-Based Image Generation to Investigate its Impact on Single Cell Detection Accuracy

Faculty of Engineering and Computer Science
Department Computer Science

HOCHSCHULE FÜR ANGEWANDTE
WISSENSCHAFTEN HAMBURG
Hamburg University of Applied Sciences

Mario de Jesus da Graca

Leveraging Diffusion-Based Image Generation to Investigate its Impact on Single Cell Detection Accuracy

Master thesis submitted for examination in Master´s degree
in the study course *Master of Science Informatik*
at the Department Computer Science
at the Faculty of Engineering and Computer Science
at University of Applied Science Hamburg

Supervisor: Prof. Dr. Peer Stelldinger
Supervisor: Prof. Dr. Jörg Dahlkemper

Submitted on: 13th September 2024

Mario de Jesus da Graca

Thema der Arbeit

Nutzung der diffusionsbasierten Bilderzeugung zur Untersuchung ihres Einflusses auf die Erkennungsgenauigkeit einzelner Zellen

Stichworte

Mikrobiologie, Mikroskopie, Deep Learning, Diffusion, Unkonditionierte Diffusion, Bildgenerierung, Objekterkennung, Zellerkennung

Kurzzusammenfassung

Diese Arbeit bewertet den Einfluss von synthetischen Hellfeldmikroskopiebildern auf Objekterkennungsmodelle bei der Einzelzellenerkennung. Eine zuverlässige Erkennung und Analyse einzelner Zellen in Mikroskopiebildern ist eine kritische Aufgabe in vielen biologischen und medizinischen Anwendungen. Deep-Learning Modelle haben großes Potenzial gezeigt, um Zellerkennungsaufgaben zu automatisieren. Jedoch ist ihre Leistung stark von der Verfügbarkeit großer, vielfältiger und genau gelabelter Datensätze abhängig. Ebenso sind hochwertige Mikroskopiebilder teuer und zeitaufwändig zu erwerben, was es schwierig macht, große Datensätze für das Training zu erhalten. Um dieses Problem zu lösen, können generative Modelle verwendet werden, um synthetische Mikroskopiebilder zu erstellen. Durch das Trainieren von unkonditionierten Diffusionsmodellen auf einem Datensatz von echten Hellfeldmikroskopiebildern generiert diese Studie synthetische Bilder, um die Trainingsdaten zu erweitern. Die Erkennungsrate und Genauigkeit von Objekterkennungsmodellen die auf Datensätzen trainiert wurden, die unterschiedliche Anteile von synthetischen Daten enthalten, wird dann auf echten Testdaten bewertet. Die Ergebnisse zeigen, dass diese Diffusionsmodelle in der Lage sind, realistische und hochwertige synthetische Bilder zu generieren, die schwer von echten Bildern zu unterscheiden sind. Aktuelle Objekterkennungsmodelle, die auf diesen synthetischen Daten trainiert wurden, erreichen vergleichbare Ergebnisse zu denen, die nur auf echten Daten trainiert wurden, insbesondere bei niedrigeren IoU-Schwellenwerten. Jedoch wird der Unterschied bei höheren IoU-Schwellenwerten und einem höheren Prozentsatz an synthetischen Daten leicht größer. Die Ergebnisse zeigen, dass synthetische Bilder verwendet werden können, um Objekterkennungsmodelle für Einzelzellenerkennungsaufgaben zu trainieren, was einen vielversprechenden Ansatz darstellt, um die Herausforderungen bei der Datenerfassung und -annotation in der biologischen Forschung zu überwinden.

Mario de Jesus da Graca

Title of Thesis

Leveraging Diffusion-Based Image Generation to Investigate its Impact on Single Cell Detection Accuracy

Keywords

Microbiology, Microscopy, Deep Learning, Diffusion, Unconditional Diffusion, Image Generation, Objet Detection, Cell Detection

Abstract

This thesis evaluates the impact of synthetic brightfield microscopy images on object detection models in single cell detection tasks. Reliably detecting and analyzing individual cells in microscopy images is a critical task in many biological and medical applications. Deep learning models have shown great promise in automating cell detection tasks, but their performance is heavily dependent on the availability of large, diverse, and accurately labeled datasets. But acquiring high-quality microscopy images is expensive and time-consuming, making it challenging to obtain large-scale datasets for training. To address this issue, generative models can be used to create synthetic microscopy images that closely resemble real images. By training unconditional diffusion models on a dataset of real brightfield microscopy images, this study generates synthetic images to augment the training data. The performance of object detection models trained on datasets containing varying proportions of synthetic data is then evaluated on real test data. The results show that these diffusion models are able to generate realistic and high-quality synthetic images that are difficult to distinguish from real images. State-of-the-art object detection models trained on this synthetic data achieve comparable performance to those trained on real data only, particularly at lower IoU thresholds. However, the performance gap slightly widens at higher IoU thresholds and increased percentages of synthetic data. The findings show that synthetic images can be used to train object detection models for single cell detection tasks, offering a promising approach to overcoming the challenges of dataset acquisition and annotation in biological research.

Contents

List of Figures viii

List of Tables x

Abbreviations xi

1 Introduction 1
- 1.1 Problem Statement . . . 1
- 1.2 Objective and Scope . . . 2
- 1.3 Research Questions . . . 3
- 1.4 Thesis Structure . . . 3

2 Related Works 5
- 2.1 Synthetic Brightfield Microscopy . . . 5
- 2.2 AI-Based Microscopy Image Analysis . . . 7

3 Technical Background 10
- 3.1 Diffusion Fundamentals . . . 10
  - 3.1.1 Overview . . . 10
  - 3.1.2 Mathematical Concepts . . . 11
  - 3.1.3 Model Architectures . . . 13
  - 3.1.4 Applications and Popular Models . . . 15
  - 3.1.5 Unconditional Diffusion . . . 16
- 3.2 Object Detection Fundamentals . . . 18
  - 3.2.1 Overview . . . 18
  - 3.2.2 History of Object Detectors . . . 19
  - 3.2.3 Datasets and Performance Metrics . . . 23

4 Motivation 27
- 4.1 Current Challenges and Limitations . . . 27

4.2 Potential Benefits and Impact . . . 29
4.3 Ethical Considerations and Responsible Development . . . 30

5 Methodology: Image Generation 32
5.1 Dataset Acquisition . . . 32
5.1.1 Biological Setup . . . 33
5.1.2 Pre-Processing . . . 34
5.2 Model Architectures . . . 37
5.3 Model Training . . . 39
5.4 Model Evaluation . . . 41
5.4.1 Objective Metrics . . . 41
5.4.2 Subjective Measures . . . 43
5.5 Final Model Decision . . . 43

6 Methodology: Cell Detection 47
6.1 Dataset Acquisition . . . 48
6.1.1 Real Images . . . 49
6.1.2 Generated Images . . . 50
6.2 Model Architectures . . . 51
6.2.1 YOLOv8 . . . 51
6.2.2 YOLOv9 . . . 52
6.2.3 RT-DETR . . . 53
6.3 Model Training . . . 54
6.3.1 Fine-Tuning Approach . . . 54
6.3.2 Training Configuration . . . 55
6.4 Model Evaluation . . . 56

7 Results 58
7.1 Survey Results . . . 58
7.2 Visual Comparison of Generated and Real Images . . . 61
7.3 Single Cell Detection Performance . . . 64
7.3.1 Quantitative Metrics . . . 66
7.3.2 Qualitative Assessment . . . 69

8 Discussion 73
8.1 Interpretation of Results . . . 73
8.2 Impact on Biological Research and Applications . . . 76

8.3 Limitations and Future Directions 77

9 Conclusion **79**
9.1 Summary of Findings 79
9.2 Outlook and Recommendations 80

Bibliography **81**

A Survey **111**

B Sample Detections **115**

Glossary **119**

Declaration of Authorship **121**

List of Figures

3.1 The forward (a) and backward (b) diffusion process. 11
3.2 The left image of the figure illustrates the anchor-based approach, where multiple bounding boxes of varying sizes and aspect ratios are generated around the object, represented by different colored rectangles. The right image shows the anchor-free approach, where the object is directly detected without predefined anchor boxes, focusing on the center and scale of the object, indicated by red arrows. This visual comparison highlights the difference in methodology between anchor-based and anchor-free object detectors. (Image source: Zhang et al. [83]) 19

5.1 Resulting images from brightfield and fluorescence imaging using the CELLAVISTA 3.1 RS HE. Typically, the fluorescence image is overlaid on top of the brightfield image and tinted to reflect the emission color of the eGFP protein. 35
5.2 Illustration of various microtiter plate well shapes and a subwell image demonstrating how the well edge can obscure the visibility of cells. 36
5.3 FID metric comparison of the different diffusion architectures throughout the training process. The top 5 lowest FID checkpoints of each model are marked with a star. 44

7.1 Normalized confusion matrix showing the distribution of correct and incorrect classifications for real and generated microscopy images. 58
7.2 Bar chart displaying accuracy for individual AI-generated and real images. 59
7.3 Word cloud depicting common reasons for classifying an image as generated. 60
7.4 Density histogram depicting the contrast levels of generated and real images. 62
7.5 Density histogram illustrating the brightness levels of generated and real images. 63
7.6 mAP@50 values for all models trained on the four datasets. 66
7.7 mAP@75 values for all models trained on the four datasets. 67

7.8 mAP@50:95 values for all models trained on the four datasets. 68
7.9 Overlapping boxes. 69
7.10 Missing box. 70
7.11 Split box. 70

A.1 Survey form – Part 1. 111
A.2 Survey form – Part 2. 112

B.1 Comparison of labels and predictions from the YOLOv8s model that was trained on the *scc_30* dataset. 116
B.2 Comparison of labels and predictions from the YOLOv9e model that was trained on the *scc_10* dataset. 117
B.3 Comparison of labels and predictions from the RT-DETR-l model that was trained on the *scc_10* dataset. 118

List of Tables

5.1 This table presents a detailed comparison of the diffusion model architectures. 39

6.1 Baseline and mixed datasets used for object detection model training. . . 48

7.1 Detailed comparison of the performance metrics for all evaluated models across the different training datasets. 65

A.1 Generated images employed in the survey. 113
A.2 Real images employed in the survey. 114

Abbreviations

AI Artificial Intelligence

AP Average Precision

CHO Chinese Hamster Ovary

CMV Cytomegalovirus

CNN Convolutional Neural Network

COCO Common Objects in Context

DDIM Diffusion Diffusion Inference Model

DDP Distributed Data Parallel

DDPM Diffusion Denoising Probabilistic Model

DETR Detection Transformer

DiT Diffusion Transformer

DPM Deformable Part-based Model

eGFP Enhanced Green Fluorescent Protein

EMA Exponential Moving Average

FID Fréchet Inception Distance

FPN Feature Pyramid Network

GAN Generative Adversarial Network

GPU Graphics Processing Unit

HOG Histogram of Oriented Gradients

ILSVRC ImageNet Large Scale Visual Recognition Challenge

IoU Intersection over Union

IS Inception Score

KID Kernel Inception Distance

mAP Mean Average Precision

MS-COCO Microsoft Common Objects in Context

MS-SSIM Multi-Scale Structural Similarity Index Measure

NMS Non-Maximum Suppression

OID Open Images Dataset

PASCAL Pattern Analysis, Statistical Modelling and Computational Learning

R-CNN Region-based Convolutional Neural Network

RPN Region Proposal Network

RT-DETR Real-Time Detection Transformer

SPPNet Spatial Pyramid Pooling Networks

SSD Single Shot MultiBox Detector

VOC Visual Object Classes

YOLO You Only Look Once

1 Introduction

Cell detection is a fundamental task in many biological and medical applications, serving as a crucial step in understanding cellular processes, diagnosing diseases, and developing new treatments [114]. The ability to accurately identify and locate individual cells within brightfield microscopy images enables researchers to quantify cell populations, analyze cell morphology, and track cellular behavior over time. However, the process of cell detection has traditionally been a labor-intensive and time-consuming task, often requiring manual annotation by trained domain experts.

1.1 Problem Statement

Recent advancements in Artificial Intelligence (AI) and computer vision have shown great promise in automating the cell detection process [42]. For instance, a study by Falk et al. [173] demonstrated the successful application of Artificial Intelligence (AI)-assisted cell detection in analyzing large-scale time-lapse experiments of neural stem cell development. Their deep learning-based approach not only significantly reduced the time required for analysis but also improved the consistency and reproducibility of results compared to manual annotation [173]. This example highlights the potential impact of automated cell detection on accelerating biological research and improving the reliability of experimental outcomes.

Despite the potential benefits of AI-assisted cell detection, several challenges remain in developing robust and accurate automated systems. One of the primary difficulties lies in the inherent variability of cell appearances in microscopy images. Cells can vary wildly in shape, size, and intensity, depending on factors such as cell type, cell cycle stage, and imaging conditions [23]. This variability makes it challenging to develop algorithms that can reliably detect cells across different experimental setups and biological contexts.

Deep learning approaches, particularly Convolutional Neural Networks (CNNs), have emerged as powerful tools for addressing the complexities of cell detection [188]. These models can learn to recognize complex patterns and features directly from raw image data, potentially overcoming the limitations of traditional handcrafted image processing techniques [149]. However, the success of deep learning models is heavily dependent on the availability of large, diverse, and accurately labeled datasets for training [149].

Herein lies the significant bottleneck in the development of automated cell detection systems: the acquisition and labeling of large-scale microscopy datasets. Obtaining a sufficient number of high-quality brightfield microscopy images is both expensive and time-consuming, often requiring specialized equipment and expertise [132]. Moreover, the process of manually annotating these images to create ground truth labels for training is equally demanding, requiring significant time and effort from domain experts.

To address these challenges, researchers have begun exploring the potential of synthetic data generation as a means to augment over even replace real-world datasets [138, 174, 94]. By leveraging generative models, it may be possible to create large volumes of realistic brightfield microscopy images.

1.2 Objective and Scope

The primary objective of this thesis is to evaluate the impact of introducing synthetic brightfield microscopy images into training datasets for object detection models. The investigation faces several key challenges, including the generation of high-quality synthetic microscopy images, the training of various object detection models on datasets containing varying proportions of synthetic data, and the rigorous evaluation of model performance on real-world test data.

The scope of this research is focused on brightfield microscopy images of single cells, a common and important imaging modality in biological research. While the ultimate goal is not to maximize detection in absolute terms, the study aims to provide a comparative analysis of how synthetic data impacts the performance of different state-of-the-art object detection architectures. This approach will yield insights into the potential benefits and limitations of using synthetic data in the context of cell detection.

1.3 Research Questions

The central research question driving this thesis can be stated as follows:

> **RQ: Does the introduction of synthetic brightfield microscopy images to a training dataset positively impact the performance of object detection models in single cell detection tasks?**

Additionally, the research will explore two related sub-questions:

1. SRQ1: Is it possible to generate synthetic brightfield microscopy images to sufficient quality to make them indistinguishable from real images to human experts?
2. SRQ2: How does the proportion of synthetic data in the training set influence the magnitude and direction of its impact on model performance?

This thesis aims to address these research questions and contribute valuable insights to the fields of computer vision, machine learning, and computational biology. The findings may have significant implications for the development of more efficient and effective cell detection systems, potentially accelerating biological research and improving the accuracy of cellular analysis in various applications.

1.4 Thesis Structure

The structure of this thesis is designed to provide a comprehensive exploration of the research questions and objectives outlined above. Following this introduction, Chapter 2 presents a review of relevant literature, covering topics such as recent advancements in synthetic microscopy image generation and deep learning techniques in microscopy image analysis. Chapter 3 provides essential technical background on diffusion models, different diffusion types and architectures, fundamentals and history of object detection and common metrics used to evaluate object detection models. Chapter 4 goes into more detail on the motivation behind this research, discussing the current challenges and limitations, potential benefits and ethical considerations of using AI, especially synthetic data, in biological research. Chapter 5 outlines the methodology for generating the synthetic brightfield microscopy images used in this study. It contains details on the dataset acquisition and preparation, model architectures trained to generate images, training procedures, model evaluation process, and presents the final model that was selected for

generating the synthetic images used in the subsequent experiments. Chapter 6 builds upon the methodology presented in the previous chapter and describes the process of the cell detection task. It covers the data labeling, state-of-the-art model selection, training configuration, and describes the evaluation process. The results of the experiments are presented in Chapter 7, including a survey of human experts, visual comparisons of generated and real images, and quantitative and qualitative assessments of cell detection performance. Chapter 8 interprets the results, discusses the potential impact of the findings on biological research and applications, and outlines the limitations and future directions of the research. Finally, Chapter 9 summarizes the key findings of the study, provides recommendations for future work, and concludes the thesis.

2 Related Works

Brightfield microscopy is a widely used imaging technique in biological research and medical diagnosis. However, acquiring high-quality brightfield images can be challenging due to various factors such as illumination non-uniformity, focusing issues and other laboratory environment settings. In recent years, there has been growing interest in using artificial intelligence and deep learning approaches to enhance and analyze microscopy images. This chapter provides an overview of relevant prior work in two main areas related to this thesis:

1. Methods for generating synthetic brightfield microscopy images, and
2. AI-based techniques for microscopy image analysis.

The first section focuses on computational approaches that aim to simulate realistic brightfield images, which can aid in training deep learning models when real image data is limited. The second section surveys the application of machine learning, especially deep learning, for various microscopy image analysis tasks such as cell detection, segmentation and classification.

2.1 Synthetic Brightfield Microscopy

One of the early works in this area by Svoboda et al. [33], proposed a technique to generate time-lapse sequences of fully 3D synthetic brightfield microscopy image datasets, including cell shapes, structures and motion. This enabled the creation of ground truth data for evaluating cell segmentation and tracking algorithms. They further extended this work [34] to generate more realistic distributions of cell populations in 3D by controlling cell count and clustering probability.

Several research groups have focused on translating brightfield images into the corresponding fluorescence microscopy images, to reduce the amount of destructive fluorescent that has to be added to cells. This step can greatly reduce the time-consuming and laborious tissue preparation process and further improve high-throughput screening [64]. Christiansen et al. [27] have introduced the approach "In Silico Labeling" (ISL) using deep learning to predict fluorescent labeling from transmitted light images. Building upon this, Lee et al. [63, 64] developed DeepHCS and DeepHCS++ to transform brightfield images into multiple fluorescence channels commonly used in high-content screening. They leveraged multitask learning with adversarial losses to generate realistic virtual stains.

Generative Adversarial Network (GAN) have emerged as a powerful framework for microscopy image generation and translation tasks. Zhang et al. [195] combined a Generative Adversarial Network (GAN) with light microscopy to achieve deep learning-based super-resolution under a large field-of-view. The model can recover a high-resolution accurate image from its single low-resolution measurement. In the same year, Scalbert et al. [111] developed a generic isolated cell image generator by approximating the shape and texture distributions of cell components using GANs. Furthermore, Liu et al. [103] used a GAN to transfer styles across brightfield and fluorescence modalities to augment limited annotated data for cell segmentation. They were able to improve the segmentation accuracy of the two top-ranked Mask Region-based Convolutional Neural Network (R-CNN)-based nuclei segmentation algorithms significantly. And very recently, Liu et al. [101] proposed a two-stage cell image recovery model that is able to reduce or even remove phenotypic feature loss caused by saturation artifacts, enabling a full analysis of the morphological features of cells in a given microscopy image.

More recently, diffusion models have shown promise for image generation in this domain. Cross-Zamirski et al. [28] introduced a class-guided diffusion model to generate cell painting images from brightfield along with class labels. Taking a different approach, Della Maggiora et al. [52] presented a conditional variational diffusion model that learns the noise schedule during training itself, trying to improve a common drawback with diffusion models, and demonstrating applications in super-resolution microscopy and quantitative phase imaging. Lu et al. [105] developed EMDiffuse, a suite of algorithms designed to enhance Electron Microscopy (EM) and Volume Electron Microscopy (vEM). It demonstrates proficiency in several tasks such as reconstruction and generation. Incorporating the physical model of microscopy image formation into the loss function of a Diffusion Denoising Probabilistic Model (DDPM), Li et al. [97] suggest a physics-informed Dif-

fusion Denoising Probabilistic Model (DDPM) (PI-DDPM) to improve reconstructions with reduced artifacts compared to regular DDPMs when trained on synthetic microscopy data.

In summary, the development of deep learning methods for artificially generating and improving brightfield images has made significant progress. Techniques include unconditional generation, cross-modality translation, super-resolution, and style transfer. Although previous research relied heavily on GANs, diffusion models are gaining popularity due to their ability to generate high-quality and diverse samples. However, in particular when training data is limited, further research is needed to improve the reliability and utility of the generated images. The methods discussed here lay the foundations for this research on using unconditional diffusion models to generate realistic brightfield images to improve single cell detection.

2.2 AI-Based Microscopy Image Analysis

In recent years, deep learning approaches have achieved state-of-the-art performance on cell segmentation and detection tasks in microscopy images. Convolutional Neural Networks (CNNs) have proven particularly effective due to their ability to learn hierarchical features directly from raw pixel data. One of the most influential deep learning architectures for biomedical image segmentation is U-Net, originally proposed by Ronneberger et al. [149]. U-Net uses an encoder-decoder structure with skip connections to combine high-resolution features from the contracting path with upsampled outputs from the expanding path. This allows the network to produce precise localization while capturing context. The authors demonstrated U-Net's ability to achieve good segmentation results with very little training data in microscopy applications, while maintaining a very high computation speed. Falk et al. [173] later released a user-friendly ImageJ [158] plugin to make U-Net accessible to non-experts for cell segmentation and detection tasks.

Many subsequent works have built upon the U-Net architecture. Caicedo et al. [86] performed an extensive evaluation of U-Net and other deep learning strategies for nucleus segmentation in fluorescence images. They found that U-Net outperformed classical image processing algorithms and could reduce biologically relevant errors by half. Raza et al. [159] offered a multi-scale version of U-Net called Micro-Net for segmenting various objects in microscopy images. By training the network at multiple resolutions, Micro-Net improved performance on datasets with high variability in cell sizes. To further improve

research in brightfield microscopy, Salem et al. [30] present a U-Net architecture, called YeastNet, to conduct semantic segmentation on brightfield microscopy images. The presented model is able to accurately generate segmentation masks, which can further be used for cell labeling and tracking. To evaluate the performance on the HeLa cell line, Ghaznavi et al. [6] proposed a residual attention U-Net and compared it with an attention and simple U-Net. They came to the conclusion that the residual attention architecture achieved the best results regarding the mean Intersection over Union (IoU) and Dice metric. Li et al. [17] trained the state-of-the-art object detection model You Only Look Once (YOLO)X [53] on 2D brightfield microscopy images to detect T-cells with a mean average precision of 96.32%. Taking it a step further in 2024, Ferreira et al. [40] leverage CNN-based models to not only detect cells, but also classify them into different lineages with high accuracy.

Other enhancements to U-Net have also been explored. Ounkomol et al. [26] used a label-free approach to predict 3D fluorescence directly from transmitted-light images, which can be extended to predict immunofluorescence from electron micrograph input. Taking a different approach, Arbelle et al. [13] drew inspiration from GANs to train U-Net in a weakly supervised manner, reducing the need for pixelwise annotations. By adding a modified encoded branch to the standard U-Net architecture, Long et al. [104] proposed a light-weighted variant, called U-Net+. This enables the model to work with low-resource computing hardware, while increasing the average Intersection over Union (IoU) by 1–3% compared to competing methods.

Beyond U-Net, other CNN architectures have also been applied to cell segmentation. 3 years before the U-Net architecture, Akram et al. [153] recommend a CNN-based method that provides high-quality segmentation candidates, which can be used for cell detection, segmentation and tracking. These segmentation candidates provide an efficient way of exploiting the spatial and temporal context to aid in the decision process of ambiguous, overlapping regions. Xie et al. [185] leverage convolutional neural networks to regress a cell spatial density map across the image. This method works especially well, when standard image segmentation fails due to cell clumping or overlap, enabling them to set new state-of-the-art performances for cell counting and detection. Pan et al. [189] designed a novel multi-scale fully convolutional neural network for regression of a density map to robustly detect the nuclei of pathology and microscopy images. To capture the state-of-the-art in 2020, Waithe et al. [38] benchmarked leading object detection networks like Faster Region-based Convolutional Neural Network (R-CNN), YOLOv2, YOLOv3, and RetinaNet for cell detection in fluorescence microscopy. And Mohammed et al. [118]

performed an extensive comparison of U-Net with other state-of-the-art segmentation networks like U-Net++, DeepLabv3+, and a novel architecture called PPU-Net across multiple microscopy modalities, showing, that deeper architectures outperform standard U-Net.

In addition to 2D segmentation, deep learning has been used to analyze 3D volumetric microscopy data. Lugagne et al. [81] developed DeLTA, a 3D cell segmentation and tracking pipeline based on U-Net. Their approach enabled high-throughput analysis of E. coli cells in a microfluidic device without human intervention.

Litjens et al. [54] reviewed over 300 contributions to the field of medical image analysis to give a concise overview of tasks and applications. They end with a summary of the current state-of-the-art, as well as a critical discussion about open challenges and future research. Building on this work, Moen et al. [42] and Durkee et al. [108] provide a comprehensive review of how deep learning is transforming all facets of cellular image analysis, including classification, segmentation, and augmented microscopy. Liu et al. [198] present a survey on applications of deep learning in microscopy image analysis, and also discuss drawbacks of existing deep learning-based methods, highlighting the acquisition of labeled training data.

In addition, Castiglioni et al. [74] also dive into the topic of unbalanced medical datasets, as well as explainable AI to deal with the so-called black box issue [136]. Kopel et al. [129] give a comprehensive review on cell segmentation for label-free imaging contrast microscopy, undermining the need of fluorescence-free microscopy analysis, since the step of cell staining is time-consuming and toxic. The internal structure of the cell is destroyed, rendering the cell unable for later reusability [17].

In summary, deep learning has revolutionized the field of cellular image segmentation and detection. While challenges remain, such as the need for extensive annotated training data, advances in weakly supervised and label-free imaging approaches are promising. As microscopy continues to generate ever-increasing quantities of data, deep learning will be an indispensable tool for extracting biological insights at scale.

3 Technical Background

3.1 Diffusion Fundamentals

Diffusion models are a class of generative models that learn the underlying distribution of the training data and generate new samples that are similar to the learned distribution [69, 164]. These models are inspired by the principles of non-equilibrium thermodynamics [78] and have demonstrated state-of-the-art performance in generating high-quality images. The application of diffusion models extends beyond image generation, as they have been successfully employed in tasks such as audio synthesis [51, 200, 199], voice cloning [170, 176] and lately even video generation [71, 125].

3.1.1 Overview

The core mechanism of diffusion models, firstly presented by Sohl-Dickstein et al. [78] and further developed by Ho et al. [69], involves a two-stage process: the **forward diffusion process** (Fig. 3.1a) and the **reverse diffusion process** (Fig. 3.1b). During the forward diffusion process, the training data is gradually corrupted by adding Gaussian noise at each step. This process aims to transform the complex data distribution into a simple, tractable distribution, typically a standard Gaussian distribution. The reverse diffusion process, on the other hand, learns to recover the original data from the noisy versions by reversing the noising process step by step. By training the model to denoise the corrupted data, it effectively learns the underlying structure and patterns present in the training data.

Once trained, diffusion models can generate new data by sampling random Gaussian noise and passing it through the learned denoising process. The model gradually refines the noise, step by step, to produce a sample that resembles the training data. This generative process allows diffusion models to create novel and diverse samples that capture the essential features of the learned distribution.

Diffusion models have gained significant attention in the research community due to their ability to address the challenges commonly associated with adversarial training in GANs [59]. These challenges include training instability [86], mode collapse [144, 192], and difficulty in achieving convergence [156]. In contrast, diffusion models offer several advantages, such as improved training stability, efficiency, scalability, and the ability to parallelize the training process [36]. These properties make diffusion models an attractive choice for generative modeling tasks, particularly in scenarios where stable and reliable training is crucial.

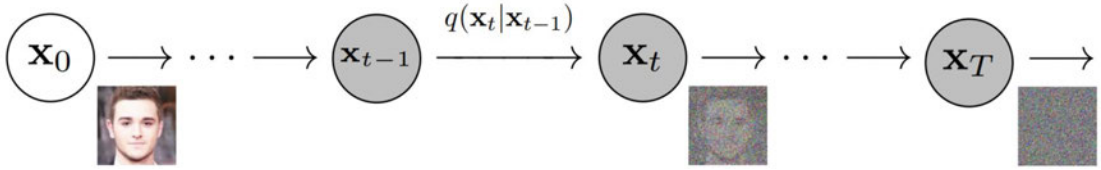
3.1.2 Mathematical Concepts

Forward Diffusion

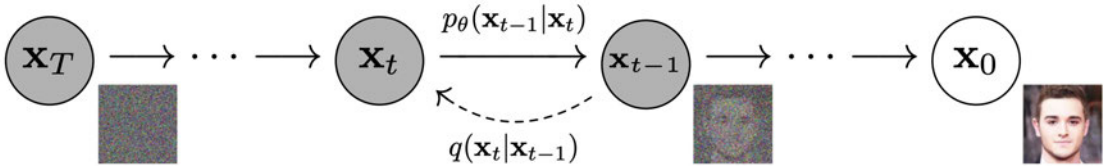
Taking a data point $\mathbf{x}_0 \sim q(\mathbf{x})$ from a real data distribution, a *forward diffusion process* can be defined in which small amounts of Gaussian noise are added to the sample in T number of steps, producing a sequence of noise samples $\mathbf{x}_1, \ldots, \mathbf{x}_T$ [69]. The step sizes are controlled by a variance schedule $\{\beta_t \in (0,1)\}_{t=1}^{T}$.

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) \coloneqq \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \quad q(\mathbf{x}_{1:T}|\mathbf{x}_0) \coloneqq \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}) \qquad \text{(Eq. 1)}$$

With increasing steps t, the data sample $\mathbf{x}_0$ loses more and more distinguishable features until $\mathbf{x}_T$ is indifferentiable from an isotropic Gaussian distribution for $T \to \infty$ [69].



(a) Slowly adding noise to the data sample $\mathbf{x}_0$ over T steps.



(b) Gradually removing noise from the data sample $\mathbf{x}_T$ over T steps.

Fig. 3.1: The forward (a) and backward (b) diffusion process.

$\mathbf{I}$ (Eq. 1) denotes the identity matrix indicating that each dimension in this multi-dimension scenario has the same standard deviation β_t. Note that $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ remains a normal distribution, defined by the mean μ and the variance Σ, where $\mu_t = \sqrt{1-\beta_t}\mathbf{x}_{t-1}$ and $\Sigma_t = \beta_t\mathbf{I}$ [69].

Thus, distinguishing diffusion models from other latent variable models, the transition from the input data $\mathbf{x}_0$ to $\mathbf{x}_T$, the approximate posterior $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$, also called *forward process*, can be defined as [69]:

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}) \tag{Eq. 2}$$

While this approach is mathematically sound, it poses practical challenges. For instance, at timestep $t = 500 < T$, q needs to be applied 500 times to sample $\mathbf{x}_t$. This repetitive application can be computationally expensive.

By defining $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s$, where $\epsilon_0, \ldots, \epsilon_{t-2}, \epsilon_{t-1} \sim \mathcal{N}(0, \mathbf{I})$, the reparameterization [37] trick can be recursively used to prove that [69]:

$$\begin{aligned} \mathbf{x}_t &= \sqrt{1-\beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\epsilon_{t-1} \\ &= \sqrt{\alpha_t}\mathbf{x}_{t-2} + \sqrt{1-\alpha_t}\epsilon_{t-2} \\ &= \cdots \\ &= \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_0 \end{aligned} \tag{Eq. 3}$$

Thus, to produce a sample $\mathbf{x}_t$, following, simplified distribution can be used [69]:

$$\mathbf{x}_t \sim q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{Eq. 4}$$

Since β_t is a hyperparameter, α_t and $\bar{\alpha}_t$ can be precomputed for all timesteps, the diffusion process is able to sample noise at any timestep t and obtain $\mathbf{x}_t$ in one step [69]. Therefore, the latent variable $\mathbf{x}_t$ can be sampled at any arbitrary timestep, without needing to apply Eq. 4 t times [69].

Ho et al. [69] proposed to utilise a linear variance schedule for β_t increasing from $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$. These boundaries were chosen to be relatively small compared to the data that was scaled to $[-1, 1]$, ensuring that the reverse and forward processes have approximately the same functional form while keeping the signal-to-noise ratio at x_T as small as possible [69]. Nichole et al. [121] however, showed that a cosine variance

scheduler works even better, by providing a smoother degradation of the image. Hence, allowing the model to operate longer on less noisy images.

Reverse Diffusion

As $T \to \infty$, the latent variable $\mathbf{x}_T$ approximates an isotropic Gaussian distribution [69]. Therefore, if the reverse distribution $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ can be learned, $\mathbf{x}_T$ can be sampled from $\mathcal{N}(0, \mathbf{I})$ [69]. By running the reverse process, and obtaining a sample from $q(\mathbf{x}_0)$, new, synthetic data points can be generated from the original data distribution [69]. Unfortunately, $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is unknown.

Instead, $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is approximated using a parameterized model p_θ: a neural network. Given that $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ will also be Gaussian for sufficiently small β_t, p_θ is chosen to be Gaussian and the mean and variance are parameterized as follows [69, 78]:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)) \quad \text{(Eq. 5)}$$

To go from $\mathbf{x}_T$ to the data distribution, the reverse formula for all timesteps ($p_\theta(\mathbf{x}_{0:T})$, also called *reverse process*) can be applied [69]:

$$p_\theta(\mathbf{x}_{0:T}) = p_\theta(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \quad \text{(Eq. 6)}$$

3.1.3 Model Architectures

Two common backbone architectures are generally chosen for diffusion models: U-Net and Transformer.

U-Net

The U-Net, as already mentioned in chapter 2.2, is a convolutional neural network architecture that was originally developed for biomedical image segmentation by Ronneberger et al. [149] in 2015. The defining characteristic of the U-Net architecture is its symmetric U-shaped structure, consisting of a contracting path (encoder) and an expansive path (decoder) [149]. The contracting path follows the typical design of a convolutional network, applying a series of convolutions and max pooling operations to capture hierarchical

features. Whereas, the expansive path increases the spatial resolution [149]. Each step in the expansive path consists of upsampling the feature map, followed by a convolution. A key innovation is the use of skip connections, which concatenate features from the contracting path with the upsampled features in the expansive path. This allows the network to combine high-resolution spatial information from the contracting path with the semantic information learned in the expansive path [149].

In diffusion models, the U-Net is used to predict the noise that was added to the image at each timestep of the diffusion process [36]. The skip connections in the U-Net, specifically, allow for the preservation of fine details and spatial information during the denoising process [36]. Several modifications and extensions have been made to the original U-Net architecture for its application in diffusion models. These include the use of residual blocks, attention mechanisms, and the incorporation of timestep embeddings to condition the network on the noise level [69, 164, 122].

Transformer

The Transformer is a deep learning architecture that has revolutionized natural language processing and is now being applied to other domains like computer vision. It was first proposed in the seminal 2017 paper "Attention Is All You Need" by Vaswani et al. [12]. The key innovation of the Transformer is its use of self-attention mechanisms instead of recurrent or convolutional layers to process sequential data. The main components of the Transformer architecture are [12]:

1. Input embedding layer: Converts input tokens into dense vector representations. Learned embeddings capture semantic and syntactic properties of the tokens.

2. Positional encoding: Injects information about the relative or absolute position of tokens in the sequence, since the Transformer contains no recurrence. Positional encodings allow the model to make use of the order of the sequence.

3. Multi-head self-attention: Allows the model to jointly attend to information from different representation subspaces at different positions. Multiple attention heads learn different relationships between tokens in the sequence.

4. Feed-forward networks: Consist of two linear transformations with a ReLU activation in between. These are applied to each position separately and identically to transform the attended representations.

5. Layer normalization and residual connections: Used after each sub-layer to stabilize training. Residual connections help propagate the signal and layer normalization keeps the scale of representations consistent across layers.

The Transformer also follows an encoder-decoder structure, with the encoder mapping the input sequence to a sequence of continuous representations, and the decoder generating an output sequence one element at a time while attending to the encoder representations [12].

Only recently, the Transformer architecture has been adapted for use in diffusion models for image generation. Peebles et al. [186] proposed the Diffusion Transformer (DiT), which replaces the typical U-Net backbone of diffusion models with a Transformer that operates on latent image patches. The Diffusion Transformer (DiT) architecture makes a few key modifications to the standard Transformer [186]:

- It uses an adaptive layer normalization mechanism to inject conditional inputs like the diffusion timestep and class label into each Transformer block.
- The input "patchify" layer linearly embeds image patches into a sequence of tokens that are fed into the Transformer stack.
- Deeper and wider DiT variants with more Transformer blocks and attention heads tend to improve performance, but also increase computation time and cost.

The largest DiT-XL/2 model outperformed prior diffusion models on the class-conditional ImageNet generation benchmarks while being relatively compute-efficient [186].

3.1.4 Applications and Popular Models

Diffusion models have found applications in a wide range of areas, leveraging their ability to model complex data distributions and generate realistic samples. Some notable applications include:

- Image Generation: Diffusion models excel at generating high-resolution, diverse, and realistic images. Models like DALL-E 2 [126], Imagen [155], and Stable Diffusion [147, 128] have showcased impressive capabilities in text-to-image synthesis, enabling users to generate images based on textual descriptions.

- Image Inpainting and Outpainting: Diffusion models can be used for image inpainting [9, 22], where missing regions of an image are filled in, and outpainting [160], where the image is extended beyond its original boundaries. This has applications in image editing, restoration, and creative design.
- Super-Resolution: Diffusion models can enhance the resolution and quality of low-resolution images, enabling the generation of high-resolution versions while preserving important details and structures [21, 193, 204].
- Video Generation: Diffusion models have been extended to generate videos by modeling the temporal dynamics and generating frames sequentially. This has potential applications in video synthesis, animation, and special effects [71, 125, 70].
- Audio Generation: Diffusion models can be applied to audio data, enabling the generation of realistic speech, music, and sound effects. This has implications for text-to-speech systems, audio synthesis, and creative applications [51, 148].

Several diffusion models have gained significant attention and popularity due to their impressive performance and wide-ranging capabilities. Stable Diffusion, developed by Stability AI, is an open-source text-to-image diffusion model that has gained widespread adoption. It offers high-quality image generation capabilities and supports various styles and domains. With version one [147], two [147] and three [128], Stable Diffusion has been a leader in advancing the art and technology of image generation using diffusion models. Developed by OpenAI, DALL-E 2 [126] & 3 [127] are powerful text-to-image diffusion models that can generate highly realistic and diverse images from textual descriptions. They have showcased remarkable abilities in capturing complex scenes and concepts. Imagen [155] was developed by Google and is another state-of-the-art text-to-image diffusion model that produces high-quality images with fine-grained control over the generated content. Lastly, Midjourney [116] is a popular diffusion-based model that focuses on artistic and creative image generation. It allows users to explore and generate images with unique styles and aesthetics

3.1.5 Unconditional Diffusion

While the previous sections have provided an overview of diffusion models and their applications, it is important to distinguish between conditional and unconditional diffusion

models, as this study focuses specifically on unconditional diffusion-based image generation. Unconditional diffusion models generate samples from a learned data distribution without any additional input or conditioning information. In contrast, conditional diffusion models incorporate extra information to guide the generation process, such as class labels, text prompts, or other forms of conditioning.

The key difference lies in the input and training process. Unconditional models are trained solely on a dataset of images, learning to generate samples that match the overall distribution of the training data. Whereas, conditional models receive additional inputs that allow more controlled generation, such as generating images based on a text description or belonging to a specific class. For unconditional image generation, the diffusion model learns to denoise randomly sampled noise tensors into coherent images resembling the training distribution, without any guiding information. This allows for open-ended image synthesis but provides less control over the specific content generated. Some key advantages of this approach include simplicity, flexibility and unsupervised learning. Unconditional diffusion models do not require paired data or labels, simplifying data collection and training. They can generate diverse samples from the learned distribution without constraints, while also potentially discovering underlying patterns and structures in the data distribution without supervision. However, unconditional diffusion models lack control over the generated content, compared to their conditional counterparts. There is no direct way to specify desired attributes or content in the generated images. This characteristic intrinsically links a single unconditional diffusion model to a specific dataset representing a particular biological setup. When dealing with multiple cell lines, experimental conditions, or imaging setups, it becomes necessary to train separate models for each scenario. This represents a limitation of unconditional diffusion models, as it necessitates more extensive and multiple training processes. In contrast, conditional models offer the advantage of being trainable on a single, diverse dataset, where each image is labeled with its corresponding condition. This allows for greater flexibility and efficiency in handling varied experimental setups within a single model framework.

In the context of this study, unconditional diffusion models present an interesting avenue for generating diverse microscopy images. By learning the underlying distribution of cellular images without explicit conditioning, these models may capture nuanced features and variations that could impact detection algorithms. The use of unconditional models also aligns with the often label-scarce nature of microscopy datasets, where paired condition-image data may be limited.

3.2 Object Detection Fundamentals

Object detection is a fundamental task in computer vision that involves identifying and localizing objects of interest within an image or video frame [197]. The goal is to develop computational models that can accurately determine what objects are present and where they are located in the visual input. This information is crucial for a wide range of applications, from self-driving cars and video surveillance to medical imaging and robotics [177].

3.2.1 Overview

The importance of object detection lies in its ability to provide a foundation for higher-level computer vision tasks. By accurately detecting and localizing objects, subsequent processes such as object tracking, instance segmentation, and scene understanding can be performed more effectively [201, 15]. Object detection enables computer vision systems to interpret and interact with their environment, making it a critical component in the development of intelligent and autonomous systems [177].

Object detection techniques have evolved significantly in the last two decades by going through two main historical periods [197]: traditional object detection (pre-2014) and deep learning-based object detection (post-2014). The early approaches relied on template matching, sliding window methods and many more classical image processing techniques. These methods were limited in their ability to handle variations in object appearance, scale, and occlusion. The introduction of AlexNet in 2012 [3], originally designed for image recognition and classification, started the revolution of object detection to rely heavily on deep learning, especially the CNN-based model architecture. It showed that CNNs could automatically learn powerful visual features, reducing the need for hand-crafted features in object detection pipelines [3].

In general, there are the two main approaches for detecting objects in images using deep learning: **One-stage object detectors** and **two-stage object detectors**. The key difference is that two-stage detectors like Faster R-CNN [161] first generate region proposals using a Region Proposal Network (RPN), and then classify and refine the coordinates of these proposals in a second stage. In contrast, one-stage detectors like YOLOX [143] and Single Shot MultiBox Detector (SSD) [102] directly predict object classes and locations in a single forward pass of the network, skipping the region proposal

step. This makes one-stage detectors faster and simpler, but sometimes less accurate than two-stage methods.

Within these two approaches, models can further be categorized as anchor-based or anchor-free (Fig. 3.2). Anchor-based detectors like Faster R-CNN [161] and RetinaNet [100] use preset anchor boxes of different scales and aspect ratios, and predict offsets relative to these anchors. Anchor-free detectors like CornerNet [93] and some versions of YOLO [53] eliminate anchor boxes, and instead directly predict keypoints like object centers or corners. The advantage is they avoid the complexities of tuning anchor hyperparameters. Overall, while two-stage anchor-based detectors dominated for many years, one-stage anchor-free models have rapidly caught up in accuracy while being faster and more flexible.

Fig. 3.2: The left image of the figure illustrates the anchor-based approach, where multiple bounding boxes of varying sizes and aspect ratios are generated around the object, represented by different colored rectangles. The right image shows the anchor-free approach, where the object is directly detected without predefined anchor boxes, focusing on the center and scale of the object, indicated by red arrows. This visual comparison highlights the difference in methodology between anchor-based and anchor-free object detectors. (Image source: Zhang et al. [83])

3.2.2 History of Object Detectors

The early era of object detection was characterized by handcrafted features and ingenious designs to overcome the limitations of computing power and image representation. In 2001, Viola and Jones [178, 179] achieved a breakthrough with their real-time face detection algorithm. Running on modest hardware, their detector was remarkably faster

than contemporary methods while maintaining comparable accuracy. The Viola-Jones detector introduced three key innovations: *Integral image representation*, *Feature selection* and *Detection cascades*. These techniques allowed for efficient computation and rapid rejection of non-face regions, enabling real-time performance. Dalal and Triggs [29] proposed the Histogram of Oriented Gradients (HOG) descriptor in 2005, which represented a significant improvement over previous feature extraction methods. Histogram of Oriented Gradients (HOG) balanced feature invariance and nonlinearity by computing gradients on a dense grid of uniformly spaced cells with overlapping local contrast normalization. While versatile, HOG was primarily motivated by pedestrian detection. Deformable Part-based Model (DPM), developed by Felzenszwalb et al. [47, 48, 49], dominated object detection challenges from 2008 to 2010. It extended the HOG detector using a "divide and conquer" philosophy, decomposing objects into parts for more robust detection. Deformable Part-based Model (DPM) introduced several influential concepts: *Mixture models*, *Hard negative mining*, *Bounding box regression* and *Context priming*. Although surpassed in accuracy by modern methods, many of DPM 's insights, like *Bounding box regression* continue to influence contemporary object detectors. Going into more detail on these methods is beyond the scope of this study, but they laid the foundation for modern object detection techniques.

The rise of deep learning, particularly CNNs, revolutionized object detection [92]. This era can be broadly categorized into two-stage and one-stage detectors. Girshick et al. [151, 58] introduced Regions with CNN features (R-CNN) in 2014, marking the beginning of the deep learning era in object detection. R-CNN used selective search to generate object proposals, then applied a CNN to extract features from each proposal. Linear SVMs were used for final classification. While R-CNN significantly improved detection accuracy, it suffered from slow inference due to redundant computations. He et al. [65] proposed Spatial Pyramid Pooling Networks (SPPNet) in 2014 to address R-CNN's speed limitations. Spatial Pyramid Pooling Networks (SPPNet) introduced a Spatial Pyramid Pooling (SPP) layer, allowing CNNs to generate fixed-length representations regardless of input image size. This innovation enabled to compute the features of an entire image just once, dramatically improving detection speed. Building on his previous work, Girshick [150] proposed Fast R-CNN in 2015. This detector enabled simultaneous training of a detector and bounding box regressor under the same network configuration. Fast R-CNN significantly improved both accuracy and speed compared to its predecessors. Ren et al. [161] introduced Faster R-CNN shortly after Fast R-CNN [150], achieving near real-time detection speeds. The key innovation was the Region Proposal Network

(RPN), which enabled efficient generation of object proposals within the neural network. Faster R-CNN represented a major step towards end-to-end trainable detection systems. Lin et al. [99] proposed Feature Pyramid Network (FPN) in 2017 to address the challenge of detecting objects at various scales. Feature Pyramid Network (FPN) introduced a top-down architecture with lateral connections, building high-level semantic feature maps at all scales. This approach has significantly improved detection of objects across a wide range of sizes.

You Only Look Once (YOLO), proposed by Joseph et al. [143] in 2015, was the first one-stage detector of the deep learning era. YOLO framed detection as a regression problem, applying a single neural network to the full image to predict bounding boxes and class probabilities simultaneously. While sacrificing some localization accuracy, especially for small objects, YOLO achieved unseen detection speeds. Since then, many more versions of YOLO have been developed and published [141, 142, 18, 53, 183, 184, 182]. Liu et al. [102] introduced the Single Shot MultiBox Detector (SSD) model in 2015, incorporating multi-reference and multi-resolution detection techniques. SSD improved upon YOLOv1's accuracy, particularly for small objects, while maintaining high detection speeds. A key innovation was detecting objects at different scales on different layers of the network. Lin et al. [100] proposed RetinaNet in 2017 to address the accuracy gap between one-stage and two-stage detectors. They identified extreme foreground-background class imbalance as a key issue and introduced the focal loss to address this problem. RetinaNet achieved comparable accuracy to two-stage detectors while maintaining the speed advantages of one-stage approaches.

Most of the above-mentioned deep learning-based object detection models rely heavily on non-maximum suppression as a post-processing step. Non-Maximum Suppression (NMS) is a crucial technique used to eliminate redundant bounding boxes and select the most relevant detections [76]. The primary purpose of Non-Maximum Suppression (NMS) is to refine the output of object detectors by removing duplicate detections of the same object, thereby improving the overall accuracy and reducing false positives [120]. NMS has its roots in early computer vision techniques. It was initially proposed as a method for edge detection in images [24]. As object detection algorithms evolved, NMS was adapted to handle bounding box refinement. The concept gained significant attention with the rise of sliding window detectors and has since become an integral part of modern object detection pipelines [29]. The NMS algorithm [5] typically follows the steps shown in Algorithm 1.

Algorithm 1 Non-Maximum Suppression (NMS)

Input:
  List of Bounding Boxes B
  List of Corresponding Scores S
  $\text{IoU}_{\text{threshold}}$
Output: List of Remaining Boxes D
$D \leftarrow \emptyset$
Sort B in descending order of scores S
while $B \neq \emptyset$ **do**
  $b_{\max} \leftarrow$ first box in B ▷ Get box with highest score
  $D \leftarrow D \cup \{b_{\max}\}$ ▷ Add it to the final list
  $B \leftarrow B \setminus \{b_{\max}\}$ ▷ Remove it from the initial list
  for all $b \in B$ **do**
    if $\text{IoU}(b_{\max}, b) > \text{IoU}_{\text{threshold}}$ **then**
      $B \leftarrow B \setminus \{b\}$ ▷ Eliminate boxes with high enough IoU
    end if
  end for
end while
return D

This process ensures that only the most confident and least overlapping detections are retained, effectively suppressing less confident, redundant predictions [102]. In two-stage detectors like R-CNN and its variants (Fast R-CNN, Faster R-CNN), NMS is typically applied twice: First after the RPN to refine the proposed regions and after the final classification and bounding box regression to eliminate duplicate detections [161]. For one-stage detectors like YOLO and SSD, NMS is applied once at the end of the network: After the network generates a set of bounding boxes and class probabilities in a single forward pass [143]. In both cases, NMS plays a critical role in reducing the number of detections and improving the overall accuracy of the object detection system [100]. But it suffers from several limitations that researchers are trying to address:

- Inability to detect nearby objects: Traditional NMS suppresses all bounding boxes that have significant overlap with the highest-scoring detection. This can lead to missed detections when objects are close together or partially occluded [139].
- Reliance on arbitrary thresholds: NMS typically uses a fixed IoU threshold to determine which boxes to suppress. This arbitrary threshold may not be optimal for all object classes and scenarios [190, 2].

- Discarding potentially useful information: By completely eliminating overlapping boxes, NMS discards information that could be valuable for improving detection accuracy [120].
- Sensitivity to localization errors: Small errors in bounding box coordinates can lead to incorrect suppression decisions, especially with high IoU thresholds [85].
- Computational inefficiency: The iterative nature of NMS can be computationally expensive, particularly for large numbers of detections [10].

The success of Transformer architectures in natural language processing has recently influenced object detection. In 2020, Carion et al. [25] proposed Detection Transformer (DETR) (**DE**tection **TR**ansformer), framing object detection as a set prediction problem using Transformers. This approach eliminated the need for many hand-designed components like anchor generation and most importantly, non-maximum suppression. Zhu et al. [202] further improved upon this concept with Deformable DETR, addressing issues of slow convergence and limited performance on small objects. In 2023, Zhao et al. [106] proposed Real-Time Detection Transformer (RT-DETR), by building on top of DETR [25]. This approach has enabled transformer-based detection models to be on par when it comes to detection speed of non-transformer models, while maintaining the accuracy and detection rate of previous DETR models. These Transformer-based approaches represent a new paradigm in object detection, achieving state-of-the-art performance on benchmark datasets [106].

3.2.3 Datasets and Performance Metrics

The creation of large-scale, diverse datasets has been instrumental in pushing the boundaries of object detection capabilities. Several key datasets have emerged as benchmarks for the computer vision community, each contributing to the field's progression in unique ways.

The Pattern Analysis, Statistical Modelling and Computational Learning (PASCAL) Visual Object Classes (VOC) Challenges, running from 2005 to 2012, were pivotal in the early development of object detection algorithms [112, 113]. Two versions of this dataset are particularly noteworthy: **VOC07** and **VOC12**. VOC07 consists of 5,000 training images with roughly 12,000 annotated objects, whereas VOC12 expands this dataset with an additional 6,000 training images and 15,000 annotated objects, resulting

in 11,000 images and 27,000 objects. Both datasets feature annotations for 20 common object classes, such as “person”, “cat” and “bicycle”, providing a foundation for detecting everyday objects [112, 113]. ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [124], organized annually from 2010 to 2017, significantly advanced the state of the art in generic object detection. The detection challenge within ImageNet Large Scale Visual Recognition Challenge (ILSVRC) utilized ImageNet images and expanded the scope to 200 object classes. This dataset marked a substantial increase in scale, with the number of images and object instances surpassing that of Pattern Analysis, Statistical Modelling and Computational Learning (PASCAL) Visual Object Classes (VOC) by two orders of magnitude. Introduced in 2014, Microsoft Common Objects in Context (MS-COCO) quickly became one of the most challenging and widely used object detection datasets [175]. Key features of Microsoft Common Objects in Context (MS-COCO) include:

- A focus on 80 object categories, with fewer classes than ILSVRC but more instances per class.
- The inclusion of per-instance segmentation annotations, in addition to bounding boxes, to aid in precise localization.
- A higher proportion of small objects (area $< 1\%$ of the image) and more densely located objects.

The MS-COCO-17 version, for example, contains 164,000 images with 897,000 annotated objects. The dataset’s emphasis on small and densely packed objects has made it particularly valuable for developing robust detection algorithms. Launched in 2018, the Open Images Dataset (OID) challenge represents the next leap in dataset scale and complexity [7]. The Open Images Dataset (OID) dataset includes 1.91 million images with 15.44 million annotated bounding boxes across 600 object categories for standard object detection, and a visual relationship detection task, focusing on identifying paired objects in specific relations. The unprecedented scale of the OID has further pushed the boundaries of object detection research.

The evolution of evaluation metrics in object detection reflects the changing priorities and capabilities of detection algorithms over time. In the early days of object detection research, particularly in pedestrian detection, metrics such as “miss rate vs. false positives per window (FPPW)” were common [29]. However, this per-window measurement was found to be flawed and not indicative of full image performance [133]. The introduction

of the Caltech pedestrian detection benchmark in 2009 [133, 134] shifted the focus to false positives per image (FPPI), providing a more realistic assessment of detector performance in full images. Currently, the most widely used evaluation metric for object detection is Average Precision (AP), first introduced in VOC2007 [112]. Average Precision (AP) is calculated as the average detection precision across different recall levels and is typically computed for each object category separately. The Mean Average Precision (mAP) across all categories serves as an overall performance metric. To assess localization accuracy, the IoU between predicted and ground truth bounding boxes is used. Traditionally, a threshold of 0.5 IoU has been used to determine whether an object is correctly detected:

$$\text{mAP@50} = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{Eq. 7}$$

where:

- N is the number of classes.
- AP_i is the Average Precision for class i.

The Average Precision (AP) for a single class is computed as:

$$AP = \sum_{k=1}^{n} (P(k) \cdot \Delta R(k)) \tag{Eq. 8}$$

where:

- n is the number of detections.
- $P(k)$ is the precision at cutoff k.
- $\Delta R(k)$ is the change in recall at cutoff k.

With the introduction of the MS-COCO dataset [175], a more strict evaluation protocol was established. The MS-COCO AP is averaged over multiple IoU thresholds ranging from 0.5 to 0.95, encouraging more precise object localization (Eq. 9). This metric has become particularly relevant for applications requiring high localization accuracy, such as robotic manipulation tasks or medical image analysis.

$$\text{mAP@[50:95]} = \frac{1}{M} \sum_{j=1}^{M} \left(\frac{1}{N} \sum_{i=1}^{N} AP_i(IoU_j) \right) \tag{Eq. 9}$$

where:

- M is the number of IoU thresholds (10, corresponding to 50%, 55%, ..., 95%).
- IoU_j is the j-th IoU threshold.
- N is the number of classes.
- $AP_i(IoU_j)$ is the Average Precision for class i at IoU threshold IoU_j.

The development of comprehensive datasets and rigorous evaluation metrics has been crucial in advancing the field of object detection. From the early days of PASCAL VOC to the current state-of-the-art MS-COCO and OID, researchers have continually raised the bar for detection algorithms. Similarly, the evolution of evaluation metrics from simple per-window measurements to sophisticated multi-threshold AP calculations reflects the increasing demands placed on modern object detectors. As the field continues to progress, it is likely that even more challenging datasets and nuanced evaluation metrics will emerge, further driving innovation in object detection techniques.

4 Motivation

Single-cell analysis has emerged as a powerful approach for understanding cellular heterogeneity and dynamics in biological systems. By examining individual cells rather than bulk populations, researchers can uncover critical insights into developmental processes, disease mechanisms, and therapeutic responses [44, 40]. However, the detection and segmentation of individual cells from microscopy images remains a significant challenge, particularly when using label-free imaging brightfield imaging techniques [72, 157]. In biological applications, "label-free imaging" techniques refer to methods that analyze biological samples without the use of fluorescent dyes or other external labels, enabling the study of cells and biomolecules in their natural state. In contrast, "labels" in machine learning denote annotated data points used to train supervised models, providing the necessary output categories or values that enable the model to learn and make accurate predictions.

4.1 Current Challenges and Limitations

Detecting single cells in brightfield microscopy images presents several significant challenges. One of the primary difficulties is the low contrast between cells and the background, making it hard to distinguish cell boundaries accurately [44]. Unlike fluorescence microscopy, where specific cellular components can be labeled, brightfield images often show cells as transparent objects with minimal intensity differences [87]. This low contrast is further complicated by illumination variations across the field of view and over time during long-term experiments [44]. Additionally, cells in brightfield images can exhibit heterogeneous intensity levels, even within the same cell population [44, 46].

To combat this issue, fluorescent dyes are often used to lift cells out of the background. But relying on fluorescence labeling techniques poses its own set of challenges. While fluorescence microscopy provides high contrast images that clearly delineate cell boundaries,

the labeling process can be time-consuming, expensive, and potentially cytotoxic [95, 40]. This limits the ability to perform live-cell imaging over extended time periods.

Another challenge is the presence of imaging artifacts, such as uneven illumination, out-of-focus effects, air bubbles, and debris, which can be mistaken for cells or obscure actual cell boundaries [157]. The three-dimensional nature of cells also poses problems when projecting them onto a two-dimensional image plane, leading to overlapping cell boundaries and apparent "lateral spillover" effects [157]. This is particularly problematic in dense cell cultures or tissues where individual cells are in close proximity [40].

Furthermore, the morphological diversity of different cell types adds another layer of complexity. Cells can vary greatly in size, shape, and texture, making it challenging to develop a universal detection algorithm [80]. This diversity necessitates the development of adaptive and robust methods that can handle various cell morphologies without extensive parameter tuning [44, 88]. Lastly, the computational demands of processing large datasets from high-throughput microscopy experiments require algorithms that are not only accurate but also efficient in terms of processing time [44, 46]. Modern microscopy experiments can generate vast amounts of image data, necessitating efficient and scalable computational methods for single-cell detection and analysis [131].

One of the biggest limitations researchers face is the acquisition of large, high-quality and diverse microscopy datasets:

1. Time constraints: Collecting extensive brightfield microscopy data is highly time-consuming. Imaging large numbers of samples across multiple conditions and time points can take days or even weeks of continuous microscope operation [166]. This is especially challenging for live-cell imaging experiments that require frequent image acquisition over extended periods.
2. Cost considerations: High-throughput brightfield microscopy setups are expensive, often requiring specialized equipment like automated stages, environmental control systems, and high-performance cameras [137]. Additionally, there are ongoing costs for sample preparation materials, microscope maintenance, and data storage infrastructure.
3. Ethical considerations: When working with biological samples, especially those derived from human subjects or animal models, researchers must navigate complex ethical requirements. This includes obtaining proper informed consent, ensuring confidentiality of sensitive information, and adhering to regulations on the use and storage of biological materials [11, 73, 61]. For human samples, researchers must

often obtain approval from institutional review boards, which can be a lengthy process[1].

4. Laboratory labor requirements: Preparing samples for large-scale brightfield microscopy experiments is labor-intensive. This includes cell culture maintenance, sample mounting, and quality control checks. Additionally, operating microscopes for extended periods and managing the resulting data requires significant personnel time and expertise [119, 152].
5. Data management challenges: The sheer volume of data generated by large-scale brightfield microscopy experiments poses significant storage and analysis challenges. Researchers must implement robust data management systems to organize, store, and process terabytes of image data [79].
6. Standardization issues: Ensuring consistency across large datasets can be difficult, especially when experiments span multiple days or use different microscope setups. Variations in sample preparation, imaging conditions, microscope performance, and microscope operator can introduce unwanted variability into the data [23, 67, 68].

4.2 Potential Benefits and Impact

Diffusion models offer promising potential for generating synthetic brightfield microscopy images, which could significantly benefit cell detection tasks. These models can produce diverse and realistic synthetic data, potentially alleviating the scarcity of labeled training data that often slows down deep learning approaches in microscopy [35, 28]. By learning to model the underlying distribution of cellular appearances, diffusion models may enhance the ability to detect cells across varying imaging conditions and morphologies [8]. This improved robustness is particularly valuable for brightfield microscopy, where cell contrast and appearance can vary immensely. Furthermore, diffusion models have demonstrated success in generating high-quality, fully-annotated microscopy datasets without the need for manual annotations [28, 35]. This capability could dramatically reduce the time and resources required for data preparation while still enabling the training of accurate cell detection models. Additionally, the synthetic data generated by diffusion models has shown potential for improving the performance of downstream tasks, such as image-based profiling and classification [28]. By leveraging these models to augment existing datasets, researchers may be able to develop more robust and generalizable cell detection

[1] Panel on Research Ethics. TCPS 2: CORE-2022 (Course on Research Ethics) Chapter 12. Retrieved from: http://tcps2core.ca/welcome (Accessed September 09th 2024)

algorithms that perform well across a broader range of brightfield imaging conditions and cell types.

Developing more effective methods for single-cell detection in brightfield images could have far-reaching implications for biological research and applications. Improved brightfield cell detection would enable large-scale, label-free imaging analysis, making high-throughput screening and longitudinal studies of live cells easier, without the need for potentially cytotoxic fluorescent labels [165, 60]. More accurate cell detection could reveal subtle variations in morphology and behavior that are masked in population-level analyses, advancing the understanding of cellular heterogeneity [60]. Additionally, enhanced capabilities for analyzing label-free imaging cellular responses could streamline the drug discovery process by enabling more physiologically relevant assays and accelerating compound screening [171, 77]. These advancements would collectively contribute to a more comprehensive and efficient approach to cellular analysis across various fields of biomedical research.

4.3 Ethical Considerations and Responsible Development

While diffusion models offer promising capabilities for generating synthetic microscopy data, their usage raises several important ethical considerations that must be carefully examined. A primary ethical concern is ensuring proper consent and privacy protections when using real microscopy data to train diffusion models. Even if the synthetic images generated are not exact copies of real data, the model may encode private or sensitive information from the training set [56]. Researchers must obtain appropriate consent and permissions to use real microscopy data for model training. Additionally, synthetic data generation should be done in compliance with relevant data protection regulations like GDPR [32]. Diffusion models can potentially amplify biases present in training data or introduce new biases [109]. For microscopy applications, this could manifest as the model generating images that are not representative of the full diversity of cell types, tissue samples, or experimental conditions. Biased synthetic data could then lead to biased or unfair performance of downstream object detection models [162]. Using synthetic microscopy images in scientific research and model development raises questions of scientific integrity and reproducibility. There is a risk that artifacts or unrealistic features in synthetic images could lead to spurious or non-reproducible results in downstream analysis [110]. The ability to generate highly realistic synthetic microscopy images creates

potential for misuse, such as fabricating scientific results or evading detection of manipulated images [117]. While diffusion models are not uniquely susceptible to misuse, their powerful generative capabilities warrant consideration of safeguards and responsible development practices. When synthetic images are used in research or clinical applications, there may be an ethical obligation to disclose their synthetic nature to human experts reviewing the images or to patients whose diagnoses rely on AI systems trained using synthetic data [31]. Clear policies should be developed regarding disclosure and consent. And lastly, training large diffusion models can have a significant computational cost and associated environmental impact [41]. Researchers should consider the energy usage and carbon footprint of synthetic data generation, and explore more efficient architectures or transfer learning approaches where possible.

5 Methodology: Image Generation

This chapter dives into the process of brightfield microscopy image generation using unconditional diffusion models. It outlines the comprehensive methodology, beginning with dataset acquisition and progressing through model architecture design, training procedures, and evaluation techniques. The chapter explores both objective metrics and subjective measures to assess the quality and realness of the generated images. Ultimately, it leads up to the selection of the final model and optimized generation process that will be used to generate synthetic data for the proceeding cell detection step.

5.1 Dataset Acquisition

High-quality image acquisition is crucial for training effective diffusion models in image generation tasks. The quality of the training data directly impacts the model's ability to learn and generate realistic images. As noted by Ho et al. [69], the forward process of adding noise to images relies on high-fidelity starting points to effectively learn the reverse denoising process. Furthermore, Rombach et al. [147] emphasize that the latent space representation in latent diffusion models benefits from clean, high-resolution input images to capture fine details and textures. The importance of image quality extends beyond just resolution; factors such as lighting, background, and camera settings play a significant role in creating optimal training data. Proper image acquisition techniques can help highlight desired features and minimize unwanted artifacts, leading to more robust and accurate diffusion models. As diffusion models continue to advance in generating high-quality synthetic images, ensuring the quality of the training data becomes increasingly important to push the boundaries of what these models can achieve.

5.1.1 Biological Setup

Chinese Hamster Ovary (CHO) cells, specifically the Chinese Hamster Ovary ()-K1 and CHO DG44 variants, were used as the starting point for the experiments in this thesis. These cells are widely used in biotechnology for producing proteins and other biological products [89], and thus serve as a great research base. The cells were "stably transfected" with a plasmid, which is a small circular piece of DNA [194] that has been permanently integrated into the cells' genetics. This plasmid contained several important elements:

1. A Cytomegalovirus () promoter: This is a genetic element that ensures strong and consistent expression of the Enhanced Green Fluorescent Protein () gene [167].
2. A gene coding for Enhanced Green Fluorescent Protein (eGFP): This protein glows green when exposed to blue or ultraviolet light, making it easy to visualize cells that have successfully incorporated the plasmid [146, 145].
3. Two antibiotic selection markers: These genes allow cells that have taken up the plasmid to survive in the presence of specific antibiotics. One marker works in bacteria (used during the plasmid preparation process) and the other in mammalian cells like CHO [146, 145].

After introducing the plasmid, the cells were grown in a special growth medium containing antibiotics. This step ensures that only cells that have successfully incorporated the plasmid survive, as they can resist the antibiotic due to the selection marker. The cells were maintained in this selective environment until over 95% of the population showed strong green fluorescence, indicating successful and stable integration of the eGFP gene.

To prepare for imaging, the cells were counted to determine their concentration. This step is crucial for calculating how much cell suspension to use when seeding the plates for AI training. Cell suspension refers to a liquid mixture where the cells are evenly dispersed in a solution. This means that cells are floating freely in a liquid rather than being attached to a surface or clumped together. The cells were then placed into 96-well plates (a standard format in biological research) at three different densities: 300, 150, and 1 cell per well. This range of densities allows for the capture of images with varying cell numbers and distributions to create a diverse dataset, allowing the model to see varying degrees of cell density.

Finally, the plates were imaged using a high-throughput microscope called CELLAVISTA 3.1 RS HE by Synentec GmbH [168]. This advanced instrument captured two types of

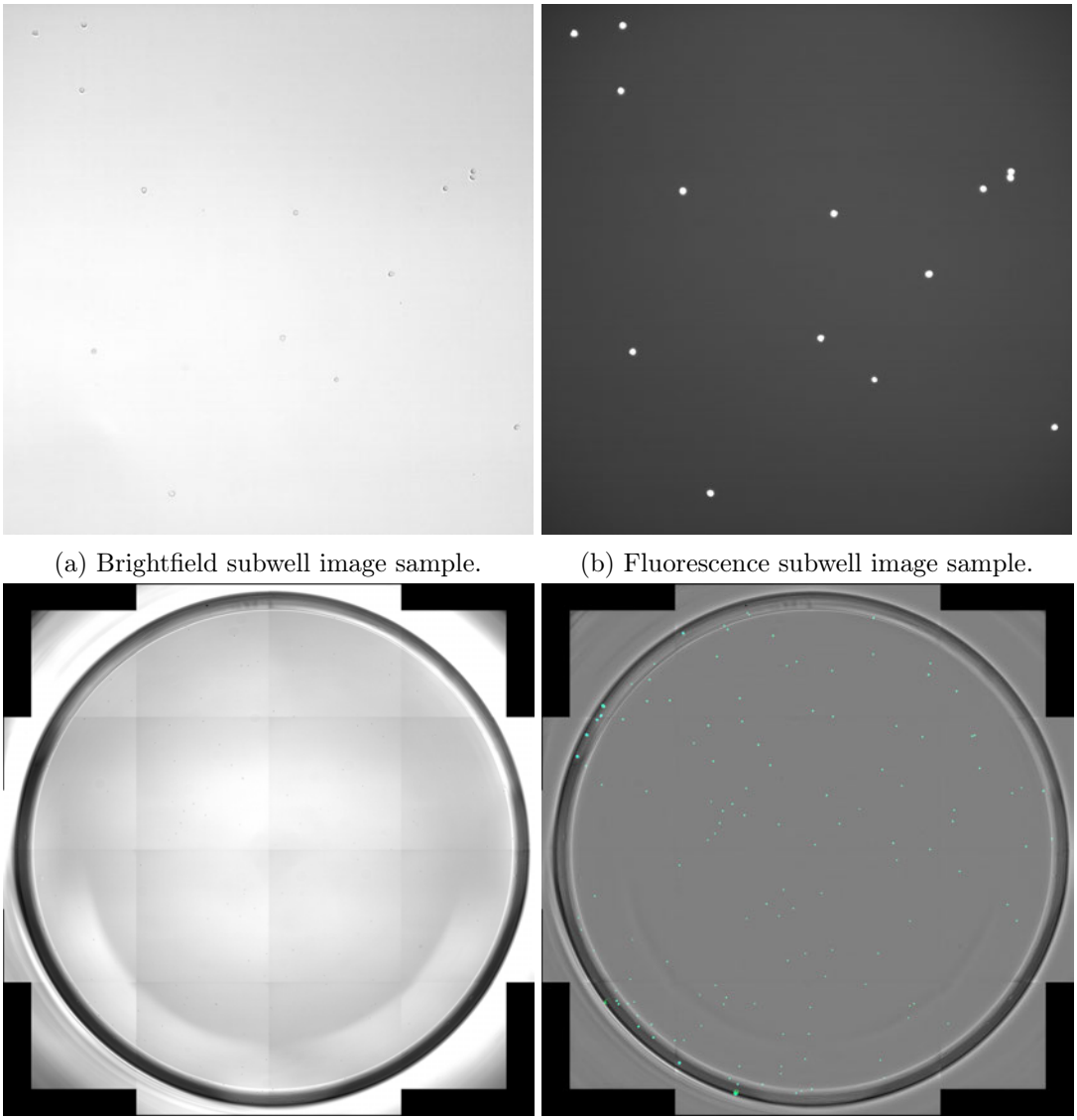
images for each well: *Brightfield* and *Fluorescence* Images. Brightfield images show the overall structure and position of cells, similar to what can be seen through a standard microscope (Fig. 5.1a). Whereas fluorescence images specifically highlight the cells expressing GFP, appearing as bright spots against a dark background (Fig. 5.1b). The images were taken using a 10× magnification objective lens, resulting in 8-bit grayscale images with a resolution of 3056x3056 pixels. The imager, with its 10× lens, captures multiple smaller regions of interest of the well, resulting in 20 total images per well. These images, also called "subwell images", are mechanically stitched together in a circular pattern, resulting in an image of size 8192×8192, due to cropping of overlapping areas (Fig. 5.1c and 5.1d). This approach was chosen to capture as much detail as possible, and to increase the relative area these small cells occupy in high-quality images.

While fluorescence microscopy images may not seem immediately relevant to this thesis focused on brightfield microscopy, they play a crucial role in the data acquisition phase of the object detection model. Although these fluorescence images are not utilized in the image generation process itself, they significantly facilitate the labeling of cells for object detection purposes. The fluorescent markers provide a clear contrast, making it easier to identify and localize individual cells and their boundaries accurately. This enhanced visibility allows for more precise and efficient labeling, and ensures that the increased workload in the cell line development step is worthwhile. A more detailed explanation of this labeling process and its importance will be provided in subchapter 6.1.

5.1.2 Pre-Processing

The Pre-processing stage plays a crucial role in preparing the raw microscopy images for training diffusion models. This phase involves several steps to optimize the dataset, ensuring that the images are suitable for the machine learning task at hand while preserving the essential cellular information.

Following the imaging process described in the previous section, a total of roughly 900 subwell images were acquired, each with dimensions of 3056×3056 pixels. To further increase the relative area occupied by cells in each image and to prevent data loss that might occur from resizing larger images, these subwell images were split into smaller patches of 512×512 pixels. This process resulted in approximately 32,000 patches, significantly increasing the dataset size while maintaining a high image quality. It's worth noting that during the patch creation process, the last column and row of patches



(a) Brightfield subwell image sample.

(b) Fluorescence subwell image sample.

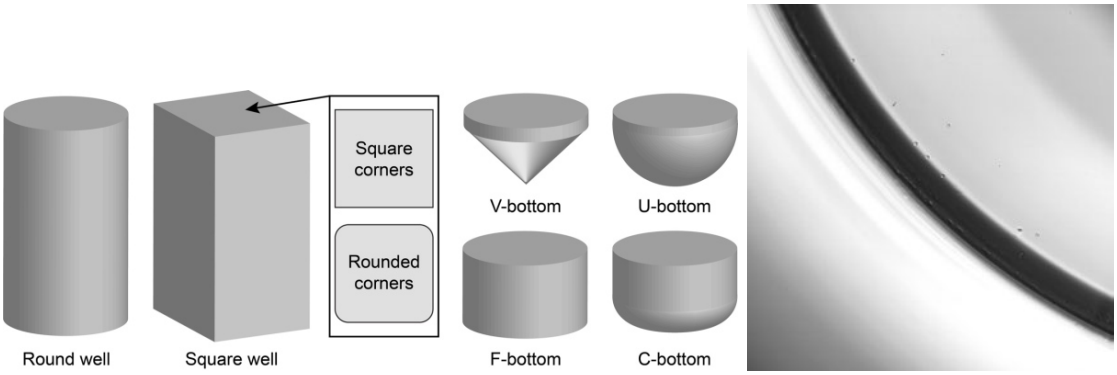
(c) Stitched full well brightfield image, with no additional visualization effects.

(d) Stitched full well brightfield image with the fluorescence channel tinted and overlaid on top. Both have been normalized across all sub well images.

Fig. 5.1: Resulting images from brightfield and fluorescence imaging using the CELLAVISTA 3.1 RS HE. Typically, the fluorescence image is overlaid on top of the brightfield image and tinted to reflect the emission color of the eGFP protein.

required black padding of 16 pixels on the right and bottom edge, to ensure a consistent patch size across the dataset. This minor adjustment ensures uniformity in the input data for the diffusion models.

One of the challenges encountered in the raw images was the presence of optical imaging artifacts caused by the shape of the wells in the microtiter plate. These artifacts manifested as dark arches in many subwell images (Fig. 5.2b), creating a thick border between the inside of the well and the plate itself. They are caused by the refraction and reflection of light on the curved surfaces of the wells, which bend and scatter light in various directions. Additionally, the curved geometry of the wells acts like lenses, focusing and defocusing light, leading to uneven illumination and shadowing effects. This artifact could potentially interfere with the model's ability to detect and analyze cells accurately. The appearance of these well edges varies considerably depending on the exact shape of the well (Fig. 5.2a), impacting the visibility of cells near the edge. To address this issue and maintain the focus on cellular structures, a decision was made to exclude patches containing visible well edges from the dataset. This filtering process reduced the total number of usable patches to approximately 24,000. While this reduction in data quantity might seem substantial, it ensures that the remaining images contain primarily relevant cellular information, which is crucial for training accurate diffusion models.



(a) Schematic of the most common microtiter plate well shapes. (Image source: Auld et al. [14])

(b) Sample subwell image with the visible well edge.

Fig. 5.2: Illustration of various microtiter plate well shapes and a subwell image demonstrating how the well edge can obscure the visibility of cells.

Given the computational constraints and the time limitations of this thesis, a further subset of 10,000 images was randomly selected from the 24,000 filtered patches. This final dataset strikes a balance between having sufficient data for training different diffusion models and managing the computational resources available within the thesis timeframe.

To optimize the training process even further and reduce potential bottlenecks in data transfer, the selected images were converted into a Hugging Face dataset format [96]. This involved transforming the image data into parquet files, a columnar storage format that offers improved efficiency and maintainability during the training phase [180]. This step is particularly important when dealing with large datasets, as it can significantly reduce the time spent on data loading and preprocessing during model training.

By implementing these pre-processing steps, a refined and optimized dataset has been created that is well-suited for training diffusion models.

5.2 Model Architectures

This section details the rationale behind testing different model configurations and presents the specific architectures employed in this study, whereas training details and further hyperparameters are explained in more detail in chapter 5.3.

The exploration of various diffusion model architectures is crucial for understanding the impact of different structural elements on the generation of single-cell microscopy images. Firstly, it allows for a comprehensive understanding of how architectural choices influence the model's ability to capture and reproduce the intricate details of single cells in microscopy images. Secondly, by testing models with and without attention mechanisms, it becomes possible to assess whether the additional computational cost of attention layers actually translates to meaningful improvements in image quality. Lastly, experimenting with larger and smaller U-Net architectures (see 3.1.3) provides insights into the trade-offs between model complexity, computational requirements, and generation quality.

The Hugging Face Diffusers [181] library was utilized for implementing and training four distinct diffusion models: *scc_small_attn*, *scc_small*, *scc_large* and *scc_medium_attn*. This choice was motivated by the library's robust implementation of state-of-the-art diffusion techniques and its flexibility in allowing customization of model architectures and interchangeable noise schedulers to evaluate different diffusion speeds and output quality.

The *scc_small_attn* model serves as the baseline architecture, incorporating a balance of standard convolutional blocks and attention blocks. With 71.4 million trainable parameters out of a total of 95.3 million, this model required a substantial training duration of 7 days, 12 hours, and 40 minutes to reach the maximum training amount of 350 epochs

(see chapter 5.3). The architecture includes attention blocks in both the down-sampling and up-sampling paths, potentially allowing for more complex feature interactions.

In contrast, *scc_small* eliminates all attention blocks, relying solely on standard convolutional operations. This simplification barely reduces the number of trainable parameters to 70.1 million, but significantly shortens the training time to about a third compared to *scc_small_attn*. The comparison between *scc_small_attn* and *scc_small* aims to elucidate the specific contributions of attention mechanisms in the context of single cell image generation.

The *scc_large* model represents a substantial increase in network capacity, featuring additional layers and a higher number of channels in the later stages of the network. With 277 million trainable parameters, this model is about four times larger than *scc_small* and explores the potential benefits of increased model complexity. Although having a bigger architecture with much more trainable parameters, this model had a similar training time as *scc_small*. This confirms that the self-attention blocks require much more computational resources than normal CNN blocks.

Finally, *scc_medium_attn* presents an intermediate approach, incorporating more attention blocks than *scc_small_attn* but maintaining a more moderate increase in overall network size compared to *scc_large*. Due to time constraints and no further visual improvements in performance, the training process of this model was cancelled after 310 epochs. With 114 million trainable parameters, and more self-attention blocks than *scc_small_attn*, training this model for 40 more epochs would have resulted in far more than 8 days of pure training.

The systematic variation in these architectures allows for an analysis of how different structural elements contribute to the model's performance in generating realistic single cell microscopy images. By comparing the outputs and performance metrics of these models, it becomes possible to draw informed conclusions about the optimal architectural choices for this specific application, potentially leading to more accurate and efficient single cell image generation systems.

Tab. 5.1: This table presents a detailed comparison of the diffusion model archictectures.

Model Name	Block Out Channels	Down Block Types[1]	Up Block Types[2]	Total Params	Train-able Params	Epochs	Training Duration
scc_small_-attn	128	DB	UB	95.3M	71.4M	350	7d 12h 40m
	128	DB	UB				
	256	ADB	AUB				
	256	DB	UB				
	512	DB	UB				
scc_small	128	DB	UB	94M	70.1M	350	2d 7h 10m
	128	DB	UB				
	256	DB	UB				
	256	DB	UB				
	512	DB	UB				
scc_large	128	DB	UB	300M	277M	350	2d 11h 30m
	128	DB	UB				
	256	DB	UB				
	256	DB	UB				
	512	DB	UB				
	512	DB	UB				
	1024	DB	UB				
scc_-medium_-attn	128	DB	UB	128M	114M	310	7d 4h 50m
	128	DB	AUB				
	256	ADB	UB				
	256	DB	AUB				
	512	ADB	UB				
	512	DB	UB				

[1] DB = DownBlock, ADB = AttnDownBlock
[2] UB = UpBlock, AUB = AttnUpBlock

5.3 Model Training

The training process for the diffusion models was designed to optimize performance while balancing computational constraints. This section details the training configuration and hyperparameters used across all model architectures.

The training dataset comprised 10,000 images, selected from a larger pool of preprocessed patches as described in section 5.1.2. This dataset size was chosen to provide

a robust representation of the cell imagery while remaining computationally tractable within the time constraints of this thesis. Each model was trained for a maximum of 350 epochs, a duration chosen through testing to allow for convergence while reducing overfitting risks. The learning rate was set at 0.0001, a value commonly used in diffusion model training that provides a balance between convergence speed and stability. The AdamW optimizer was selected for its ability to handle sparse gradients effectively, which is particularly beneficial in the context of diffusion models. The Diffusion Diffusion Inference Model (DDIM) scheduler was employed for its efficiency and improved sampling quality compared to traditional diffusion processes. An Exponential Moving Average (EMA) decay of 0.9999 was applied to the model weights, a technique known to stabilize training and potentially improve generalization. A batch size of 4 was utilized, optimized for the available Graphics Processing Unit (GPU) memory while trying to ensure sufficient stochasticity in gradient updates.

To facilitate model evaluation and enable the selection of optimal checkpoints, the training process incorporated regular checkpointing and validation. Checkpoints were saved every 10 epochs, allowing for analysis of model performance over time. Concurrently, Fréchet Inception Distance (FID) validation was performed every 10 epochs, providing quantitative metrics on the quality and diversity of generated images throughout the training process. For this purpose, a subset of 128 sample images was generated. These images were produced using 25 inference steps and the Diffusion Diffusion Inference Model (DDIM) scheduler, providing a consistent basis for evaluating model progress and performance. Torchmetrics [123] was used to calculate the Fréchet Inception Distance (FID) value.

The training infrastructure leveraged Distributed Data Parallel (DDP) across two NVIDIA A100-SXM4-40GB GPU, significantly accelerating the training process. Mixed precision training with $float16$ was implemented to further enhance computational efficiency without compromising model quality. PyTorch Lightning [45] was used to reduce boilerplate code and facilitate easy scaling to multi-GPU training scenarios. This framework also seamlessly integrated mixed precision training and other optimization techniques.

This comprehensive training configuration was applied consistently across all model architectures described in section 5.2. By maintaining consistency in the training process, meaningful comparisons could be drawn between the performance of different model architectures, isolating the impact of architectural choices on the quality of generated cell images.

5.4 Model Evaluation

The evaluation of the trained diffusion models is a critical step in assessing their performance and suitability for generating high-quality images. In the context of this thesis, where the final model will be used to generate samples for object detection, a comprehensive evaluation approach combining both objective and subjective measures is essential. The use of both objective and subjective metrics is necessary for several reasons. Objective metrics provide quantifiable and reproducible results, allowing for consistent comparisons between different models and across various experiments. They offer a standardized way to assess model performance and track improvements over time. However, these metrics may not always capture the nuanced aspects of image quality that are important for human perception [55, 154]. Subjective metrics, on the other hand, directly incorporate human judgment and can capture subtle qualities that objective measures might miss. They are particularly valuable in assessing the realism and perceptual quality of generated images, which is crucial for applications in microscopy and biological imaging. However, subjective evaluations can be time-consuming, expensive, and potentially biased [130]. By combining both approaches, the strengths of each method can be leveraged to gain a more comprehensive understanding of the model's performance.

5.4.1 Objective Metrics

The primary objective metric used in this evaluation is the FID. FID has become a standard metric in the field of generative models due to its ability to capture both the quality and diversity of generated images [55]. FID measures the distance between the feature distributions of real and generated images. It uses the Inception-v3 network, pretrained on ImageNet, to extract features from both sets of images. The feature vectors are then modeled as multivariate Gaussian distributions, and the Fréchet distance between these distributions is calculated [154]:

$$FID = \|\mu_r - \mu_g\|^2 + Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}) \quad \text{(Eq. 1)}$$

Where μ_r and μ_g are the mean feature vectors for real and generated images, respectively, and Σ_r and Σ_g are their covariance matrices. Tr denotes the trace of a matrix as a scalar value, meaning the sum of the elements on its main diagonal. It is used to measure the difference in the covariances of the real and generated data distributions. FID is preferred over other metrics for several reasons [43, 75]:

1. It correlates well with human perception of image quality.
2. It has become one of the most commonly used metric in the field.
3. It is sensitive to both the quality and diversity of generated images.

However, it's important to note that FID has some limitations. Recent research has shown that it may not always reflect the state-of-the-art perceptual realism of diffusion models as judged by humans [55]. Additionally, its reliance on the Inception-v3 network, which was trained on natural images, may not be ideal for specialized domains like microscopy images [154].

In addition to the Fréchet Inception Distance, several other objective metrics have been proposed for evaluating generative models. However, FID remains the most widely used due to its balance of effectiveness and practicality. Some alternative metrics and brief explanations of why they may be less suitable than FID for this thesis are described below:
Inception Score (IS): This metric also uses the Inception-v3 network to measure both the quality and diversity of generated images. While it was once popular, Inception Score (IS) has several limitations [16, 140]:

- It doesn't compare generated images to real ones, making it less suitable for assessing realism.
- It can be easily fooled by mode collapse, where a model generates a small set of high-quality but limited-diversity images.
- It's less reliable for domains significantly different from ImageNet, such as microscopy images.

Kernel Inception Distance (KID): Similar to FID, Kernel Inception Distance (KID) measures the distance between real and generated image features. However, it's computationally more expensive than FID and it may not provide significant advantages over FID for the specific microscopy image domain [203, 154].
Wasserstein Distance: This metric measures the cost of transforming one distribution into another. It can also be computationally intensive, especially for high-dimensional data like images, while also being less intuitive to interpret [115, 4].
Multi-Scale Structural Similarity Index Measure (MS-SSIM): This metric focuses on structural similarities between images. But it focuses too much on luminance and structural information, potentially overlooking color-related aspects and does not align well with human judgement of image quality [50, 135].

5.4.2 Subjective Measures

To complement the objective evaluation, a subjective assessment was conducted using a survey created with SurveyJS [1] (Fig. A.1 and A.2). The survey (Appendix A) included a set of sample images, consisting of 20 generated images (Tab. A.1) and 10 real images (Tab. A.2) in random order. The images were generated by the final model and hyperparameter set concluded in section 5.5. This survey was distributed to microscopy imaging experts and biologists from Synentec GmbH. The participants were asked to:

1. Decide whether each image was real or generated.
2. State the confidence in their decision.
3. If they decided an image was fake, explain what led to their decision.

This approach allows for a nuanced evaluation of the generated images, capturing aspects that might not be reflected in objective metrics. It provides insights into the model's ability to generate images that are indistinguishable from real microscopy images, which is crucial for its intended application in object detection.

5.5 Final Model Decision

The evaluation of trained diffusion models with different noise schedulers and hyperparameters is a crucial step in optimizing the performance of generative models. This process is particularly important because the chosen combination will significantly impact the quality and characteristics of the generated images, which will subsequently be used for object detection tasks. Testing various hyperparameters and noise schedulers during inference is essential for several reasons. Firstly, different schedulers can affect the trade-off between image quality and generation speed [82]. Secondly, they can influence the plausibility and semantic expression of the generated images [82].

In this study, the top 5 lowest FID checkpoints (see Fig. 5.3) of every model architecture (chapter 5.2) were tested with multiple state-of-the-art noise schedulers and different hyperparameters. The schedulers evaluated included DPM++ 2M, DPM++ 2M Karras, DPM++ 2M SDE, DPM++ 2M SDE Karras, DPM++ SDE, DPM++ SDE Karras, Euler, Euler a, and DDIM. Each scheduler was tested with varying inference steps, ranging from 10 to 45 in increments of 5 for all schedulers except DDIM, which was tested from 30 to 65 in increments of 5. Additional hyperparameters that were explored include:

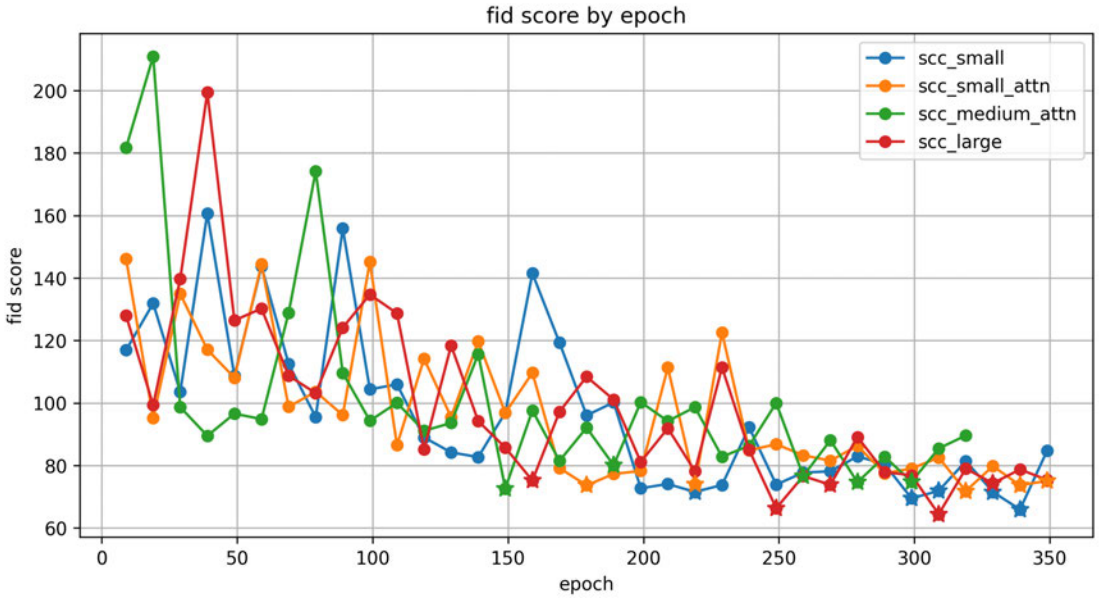


Fig. 5.3: FID metric comparison of the different diffusion architectures throughout the training process. The top 5 lowest FID checkpoints of each model are marked with a star.

- **Solver order**: This parameter, set to either 2 or 3, determines the order of the numerical solver used in the diffusion process. It is recommended to use $solver\_order = 2$ for guided sampling, and $solver\_order = 3$ for unconditional sampling.
- **Prediction type**: Two options were tested: *epsilon* and *v_prediction*. The *epsilon* prediction focuses on predicting the noise added to the image, while *v_prediction* is an alternative formulation that offers more numerical stability and might avoid color-shifting artifacts in higher resolution samples.
- **Timestep spacing**: This parameter, set to either *linspace* or *trailing*, determines how the timesteps are distributed during the diffusion process. *Linspace* spacing distributes timesteps evenly, while *trailing* spacing concentrates more steps towards the end of the process.

Each combination of these hyperparameters was used to generate 8 samples, providing a comprehensive set of images for evaluation. The final model, scheduler, and hyperparameter combination was handpicked based on a careful examination of all generated samples, with the goal of selecting the configuration that produced the best-looking images.

This approach to hyperparameter optimization aligns with recent research in the field. For instance, studies have shown that the choice of noise schedule can significantly impact the plausibility of generated designs, with some schedules improving the rate of plausible designs from 83.4% to 93.5% [82]. By meticulously evaluating these various configurations, it can be ensured that the chosen model and parameters are optimized for generating high-quality, diverse, and semantically coherent images. This optimization is crucial for the subsequent object detection tasks, as the quality and characteristics of the generated images will directly impact the performance and reliability of the detection algorithms.

After thorough evaluation of the various model architectures, noise schedulers, and hyperparameters, a final configuration was selected for the cell detection task. The chosen model is the *scc_small* architecture, which demonstrated a good balance between performance and computational efficiency. This model, despite its relatively smaller size (70.1M trainable parameters), achieved competitive results in terms of image quality and fidelity.

The selected noise scheduler for the final configuration is Euler a, which is known for its ability to produce high-quality samples with a good trade-off between speed and image fidelity. The choice of trailing timestep spacing allows for a more concentrated sampling towards the end of the diffusion process, potentially leading to improved fine details in the generated images. The prediction type was set to epsilon, focusing on predicting the noise added to the image during the diffusion process.

For the generation of the final dataset to be used in the cell detection task, a dynamic approach was implemented. The inference steps were randomly chosen between 35 and 40 for each batch of 16 images. A total of 10,000 images were generated. This variation in inference steps introduces a subtle level of diversity in the generation process, potentially benefiting the robustness of the subsequent detection model.

The final calculated FID score for this configuration was 45.57, which represents a significant improvement over the initial FID scores observed during training (Fig. 5.3). This improvement might be be attributable to the careful selection of the noise scheduler and other hyperparameters, as well as the increased number of inference steps compared to the validation setup. It's worth noting that the final decision was solely influenced by subjective evaluation of the samples. The chosen configuration produced images that were deemed to have the best visual quality and most closely resembled the characteristics of real microscopy images of single cells. This optimized model and generation

pipeline was then used as the foundation for creating a diverse and high-quality dataset of synthetic cell images. These images were crucial in the next phase of the thesis, where they were used to train and evaluate object detection models for single-cell detection in microscopy images.

6 Methodology: Cell Detection

In the field of object detection, datasets typically consist of images paired with corresponding text files containing bounding box information for the objects of interest. Over time, several standardized formats have emerged for storing this label data, each with its own strengths and use cases.

Formerly one of the most widely used formats is the PASCAL VOC format, which has been influential in the development of many object detection models, including SSD, R-CNN, Fast R-CNN, and Faster R-CNN (chapter 3.2.2). The PASCAL VOC format stores annotation data in XML files, providing a structured and human-readable representation of object locations. Each XML file contains detailed information about the image, including filename, size, and object annotations. The bounding box coordinates are represented as $[x\_min, y\_min, x\_max, y\_max]$, defining the top-left and bottom-right corners of the box.

Another popular format is the YOLO format, which is closely associated with the YOLO family of object detection models. YOLO annotations are stored in simple text files, with each line representing a single bounding box and each value in a line being separated by a whitespace. The format is concise and easy to parse, using normalized coordinates: $[class\_id, x\_center, y\_center, width, height]$. This normalization allows for easier scaling and processing of annotations across different image sizes.

The Common Objects in Context (COCO) format, developed by Microsoft, offers a more comprehensive approach to object annotation. Common Objects in Context (COCO) annotations are stored in JSON files, providing a flexible and extensible structure for complex datasets, while also being human-readable. The bounding box information is represented as $[x, y, width, height]$, where (x, y) defines the top-left corner of the box. COCO's format allows for additional features such as segmentation masks and keypoints, making it suitable for a wide range of computer vision tasks beyond object detection.

Each of these formats has its advantages and disadvantages. The PASCAL VOC format's XML structure provides rich metadata and is human-readable, but it can be more verbose and slower to parse for large datasets. The YOLO format's simplicity makes it efficient for processing and storage, but it may lack some of the detailed metadata found in other formats. The COCO format offers the most flexibility and feature-richness, but its complexity can be overwhelming for simpler projects. The choice of annotation format often depends on the specific requirements of the project, the models being used, and the available tools and workflows. For instance, researchers working with R-CNN-based models might prefer the PASCAL VOC format for its compatibility with existing codebases. Those focusing on real-time object detection might opt for the YOLO format due to its efficiency and direct integration with YOLO models. Projects requiring advanced annotations or working with diverse datasets might benefit from the extensibility of the COCO format.

For this research the YOLO format was chosen, as it aligns best with the object detection models that were trained (chapter 6.2).

6.1 Dataset Acquisition

The creation of object detection datasets is significantly more time-consuming than assembling datasets for unconditional diffusion models, primarily due to the need for detailed label data. While the base dataset described in Section 5.1.2 provided the necessary images, it lacked the crucial label data needed for training an object detection model. To address this, a comprehensive labeling process was undertaken to prepare the data for the object detection task.

Tab. 6.1: Baseline and mixed datasets used for object detection model training.

Dataset	Real Images	Generated Images	Val Images	Test Images
scc_real	5000	0	2,527	16,758
scc_10	4500	500	2,527	16,758
scc_30	3500	1500	2,527	16,758
scc_50	2500	2500	2,527	16,758

Four distinct datasets were created (Tab. 6.1), each containing a total of 5,000 labeled images for training. The remaining images were allocated to validation and test splits, maintaining consistency across all four datasets. The composition of these datasets varied

in terms of the ratio between real and generated images. For all datasets, the validation split consisted of 2,527 images, while the test split comprised 16,758 images.

The decision to use a large number of test images in the dataset can be highly beneficial for several reasons:

- Statistical Significance: A large test set provides more statistically significant results. With 16,758 images, the performance metrics derived from the test set are likely to be more robust and reliable, reducing the impact of random variations or outliers.
- Diverse Scenarios: In microscopy, cell appearances can vary greatly due to factors like cell cycle stage, environmental conditions, or imaging parameters. A large test set is more likely to encompass a wide range of scenarios, ensuring that the model's performance is evaluated across diverse conditions.
- Confidence in Generalization: With a substantial test set, there's greater confidence that the model's performance metrics reflect its true generalization capability rather than being overly influenced by the specific composition of a smaller test set.
- Confidence in Small Improvements: When comparing different models or techniques, a large test set allows for the detection of small but statistically significant improvements that might not be apparent with a smaller dataset.

6.1.1 Real Images

As mentioned in Section 5.1.1, the fluorescence images, which were not utilized during the diffusion training process, played a crucial role in the labeling of real images. The availability of these fluorescence images significantly sped up the labeling process for the brightfield microscopy images.

To leverage the fluorescence data, a simple classical image processing algorithm was developed using OpenCV [20]. This algorithm was designed to detect cells in the binarized fluorescence mask (as shown in Fig. 5.1b). Bounding boxes were then extracted from the detected contours. The process can be summarized by the simplified pseudocode of Algorithm 2. This algorithm provides a foundation for automated cell detection, significantly reducing the manual labeling effort. However, to ensure accuracy, the detected bounding boxes were manually checked and, when necessary, slightly adjusted. This approach combines the efficiency of automated detection with the precision of human verification, resulting in a high-quality labeled dataset of real images.

Algorithm 2 Cell Detection and Bounding Box Extraction Pseudocode

```
Input: Fluorescence image F
Output: List of bounding boxes B
B ← ∅
F_bin ← Binarize(F)
contours ← FindContours(F_bin)
for each contour in contours do
    if Area(contour) > min_area then
        box ← BoundingRect(contour)
        B.append(box)
    end if
end for
return B
```

6.1.2 Generated Images

The acquisition of the generated images was discussed in more detail in chapter 5.5, and only missing are the bounding boxes for the cells in these images. To facilitate the labeling process for these images, a YOLOv8m model was quickly trained on the previously labeled real image data. This pre-trained model was then used in conjunction with Roboflow's "Model-Assisted Labeling" [39] feature, where the selected model will run when an image is opened in the annotation tool, streamlining the labeling process.

While this approach significantly accelerated the labeling process compared to complete manual labeling, it was not without its limitations. The model's predictions often required minor adjustments, particularly in cases where it detected dead-looking cells, misclassified generated debris as cells or simply missed them. This process not only aided in efficient labeling but also served as an initial indicator of the quality and realism of the generated images. The need for adjustments and the types of errors encountered provided first valuable insights into the strengths and weaknesses of the diffusion-based image generation process. From the initial set of 10,000 generated images, 2,500 were carefully selected and labeled for inclusion in the mixed datasets. This selection process ensured that only the highest quality generated images were incorporated into the training data, maintaining the integrity of the dataset while exploring the impact of synthetic data on model performance.

6.2 Model Architectures

When selecting an object detection model for single cell detection in brightfield microscopy images, several factors must be taken into account:

1. Accuracy: The model should be capable of detecting cells with high precision and recall, even in challenging scenarios such as crowded fields or low-contrast images.
2. Speed: Real-time or near-real-time performance is often desirable in microscopy applications, especially for live-cell imaging or high-throughput screening.
3. Scalability: The model should be able to handle varying image sizes and cell densities without significant performance degradation.
4. Robustness: Given the potential variability in cell appearance and imaging conditions, the model should be resilient to noise, illumination changes, and other artifacts.
5. Computational requirements: The hardware available for training and inference should be considered, as some models may require substantial computational resources.
6. Ease of use and community support: Models with well-documented implementations and active community support can facilitate easier integration and troubleshooting, especially for researchers with limited experience in machine learning or computer vision.

Based on these considerations and the specific requirements of this thesis, three state-of-the-art models were selected for evaluation: YOLOv8, YOLOv9, and Real-Time Detection Transformer (RT-DETR). Each of these models offers unique advantages and has shown promising results in various object detection benchmarks.

6.2.1 YOLOv8

YOLOv8, developed and published by Ultralytics in january 2023, builds upon the success of YOLOv5 while introducing several key improvements and extensions [84]. The architecture of YOLOv8 is characterized by its efficient backbone, neck, and head components, which work in unison to achieve high accuracy and speed. The backbone of YOLOv8 utilizes a CSPDarknet53 variant, which employs cross-stage partial connections to enhance feature reuse and gradient flow. This design choice contributes to improved accuracy without significantly increasing computational complexity [172]. YOLOv8 uses

CIoU [196] and DFL [98] loss functions as the two main losses. These losses help in detecting objects of varying sizes, especially smaller objects, which is particularly beneficial for cell detection tasks where cell sizes may vary, but are always small compared to the image [172]. The head of YOLOv8 employs an anchor-free detection mechanism (chapter 3.2.1), moving away from the anchor-based approach of earlier YOLO versions.

Additional improvements are available during the training process of YOLOv8. Mosaic Augmentation is an advanced data augmentation technique that combines multiple images into a single training sample, enhancing the model's ability to detect objects in various contexts [18]. YOLOv8's head is decoupled, enabling an independent processing of objectiveness, classification and regression. This allows for more efficient training and inference, as the model can focus on specific tasks without interference from other components [172].

For this thesis, YOLOv8 models of sizes s, m, and x were trained and evaluated. These different sizes offer a range of trade-offs between accuracy and computational efficiency, allowing for a comprehensive assessment of the model's performance in the context of single cell detection.

6.2.2 YOLOv9

The successor of YOLOv8, YOLOv9 was published by Wang et al. [184] in February 2024 and aims to push the boundaries of object detection performance further.

Its architecture is characterized by several innovative features and improvements [184]:

- Programmable Gradient Information (PGI): This novel concept allows the model to generate reliable gradients through auxiliary reversible branches, ensuring that deep features can correctly maintain the key characteristics for the target task.
- Enhanced Backbone: YOLOv9 employs a refined backbone structure that incorporates a more generalized version of the Efficient Layer Aggregation Network (GELAN).
- Information Bottleneck Principle: By leveraging PGI, essential data can be preserved across the network's depth, that otherwise might be lost during passes through successive layers of the network.

- Reversible Functions: To further prevent information degradation –particularly in its deeper layers– YOLOv9 integrates reversible functions into its architecture, ensuring the preservation of vital data for object detection tasks.

For this thesis, YOLOv9 models of sizes c and e were trained and evaluated. These models represent different points on the accuracy-efficiency spectrum, allowing for a comprehensive assessment of YOLOv9's capabilities in single cell detection tasks.

6.2.3 RT-DETR

RT-DETR, or Real-Time Detection Transformer, marks a significant shift from the CNN-based architectures that have dominated the YOLO family of object detection models. Introduced by Zhao et al. [106] in 2023, RT-DETR aims to seamlessly blend the global reasoning capabilities of transformers with the computational efficiency characteristic of CNN-based detectors. This innovative approach represents a promising direction in the field of object detection, particularly for applications such as single cell detection in microscopy images.

The architecture of RT-DETR comprises several key components designed for efficient real-time object detection. At its foundation lies a CNN-based backbone, which extracts multi-scale features from the input image. Specifically, the last three stages of the backbone (S3, S4, S5) are used as input to the subsequent stages [106]. Following the backbone is an efficient hybrid encoder, which transforms the multi-scale features into a sequence of image features. This hybrid encoder employs two main modules: The Attention-based Intra-scale Feature Interaction (AIFI) for processing the S5 features, and the CNN-based Cross-scale Feature-fusion Module (CCFM) for fusing features from S4 and S5 [106]. This feature fusion strategy strikes a balance between computational efficiency and the need to capture multi-scale information. It is particularly relevant in the context of cell detection, where cells may appear at various scales within a single image due to factors such as depth of field or natural size variations. The use of both attention mechanisms and convolutional operations in the encoder allows RT-DETR to capture both local and global contextual information efficiently. Following the encoder, an IoU-aware query selection module chooses a set of initial object queries, which are then refined by a Transformer decoder with auxiliary prediction heads to generate the final bounding boxes and confidence scores.

One of the standout features of RT-DETR is its end-to-end design, eliminating the need for non-maximum suppression (chapter 3.2.2). As mentioned in previous chapters, the removal of NMS in post-processing steps enables the model to generate high-quality predictions directly, while maintaining real-time performance.

For the same reasons mentioned previously, RT-DETR models of sizes x and l were trained and evaluated.

6.3 Model Training

The training process for the object detection models in this study utilized the Ultralytics framework as the primary tool [84]. This choice was made due to its robust implementation of state-of-the-art models and its flexibility in handling various training configurations. All models were fine-tuned on the datasets described in chapter 6.1, rather than being trained from scratch. The base models had been pre-trained on the COCO dataset, a large-scale object detection dataset comprising 80 classes, as discussed in chapter 3.2.3.

6.3.1 Fine-Tuning Approach

The decision to fine-tune pre-trained models instead of training from scratch offers several advantages. Transfer learning, the process of using knowledge gained from solving one problem and applying it to a different but related problem, is at the core of fine-tuning [169]. In the context of object detection, fine-tuning allows the model to leverage general features learned from a large, diverse dataset, like shapes, and adapt them to the specific task at hand. One of the primary benefits of fine-tuning is the significant reduction in training time and computational resources required. Pre-trained models have already learned a rich set of features from a large dataset, which can be particularly beneficial when working with smaller, domain-specific datasets [191]. This is especially relevant in the current study, where the dataset size is relatively small compared to large-scale datasets like COCO. Additionally, fine-tuning often leads to better generalization and performance, particularly when the target dataset is limited in size or diversity. The pre-trained weights serve as a good initialization point, allowing the model to converge faster and potentially to a better optimum than if it were trained from random initialization [91]. Another advantage of fine-tuning is its ability to mitigate overfitting. When

training complex models on smaller datasets, there's a risk of the model memorizing the training data rather than learning generalizable features [66].

6.3.2 Training Configuration

The training process was conducted using a single NVIDIA GeForce RTX 3090 GPU with 24GB of memory. This hardware configuration allowed for efficient training of the various model architectures and sizes. Due to the restricted time frame of this thesis, an extensive hyperparameter search was not conducted. Instead, the training configurations were based on the default settings provided by the Ultralytics framework[1], with a batch size of 32, maximum epochs of 200 and early stopping patience of 35 epochs. This batch size was chosen as a balance between memory constraints and training efficiency. A larger batch size can lead to more stable gradient estimates and potentially faster convergence, but it also requires more memory [163]. The maximum number of epochs was set to 200 to allow sufficient time for model convergence with the limited data available, while the early stopping patience of 35 epochs was implemented to prevent overfitting and unnecessary computation if the model's performance plateaued.

Ultralytics default data augmentation techniques were employed to artificially expand the training dataset and improve the model's ability to generalize. The following augmentation parameters were used:

- HSV-Hue augmentation: 0.015
- HSV-Saturation augmentation: 0.7
- HSV-Value augmentation: 0.4
- Image translation: ±0.1 (fraction)
- Image scale: ±0.5 (gain)
- Left-right flip probability: 0.5
- Mosaic augmentation probability: 1.0

These augmentation techniques were chosen to introduce variability in the training data without distorting the essential characteristics of the cells. The Hue and Saturation augmentations should have no effect on truly grayscale images. The Value augmentation can slightly change the brightness of the images, simulating variations in illumination. Translation and scaling augmentations simulate variations in cell position and size, which are common in microscopy images. The left-right flip augmentation introduces reflection

[1] ultralytics/ultralytics/cfg/default.yaml at main · ultralytics/ultralytics · GitHub

invariance, while the mosaic augmentation combines multiple images, potentially helping the model learn to detect cells in various contexts and densities.

Notably, shear, and perspective transformations were not applied (all set to 0.0) to preserve the natural shape of the cells, which are important features in brightfield microscopy images. Similarly, channel swapping (BGR) was not used, as this transformation might introduce unrealistic variations that do not reflect the expected appearance of cells in the microscopy setup.

6.4 Model Evaluation

Models trained on datasets partially composed of generated images (90%, 70%, and 50% real images) were systematically compared against the baseline (100% real images). This comparison aims to effectively measure any improvements or degradations in model performance resulting from the inclusion of synthetic data.

The primary method for evaluating object detection model performance is through the comparison of standardized metrics (chapter 3.2.3). As already mentioned, while various metrics exist, Mean Average Precision (mAP) stands out as the most commonly and widely accepted metric in the field of object detection [175]. Although mAP@50:95 is often used to determine overall model precision in localization and accuracy, this study places particular emphasis on mAP@50. The rationale behind this focus lies in the specific requirements of single cell detection in brightfield microscopy images.

In the context of this research, the primary objective is to accurately detect the presence and approximate location of individual cells. Unlike tasks that require precise morphological analysis of the detected cells afterwards, the exact tightness of the bounding box is of lesser importance. As long as the cell is correctly identified and roughly localized, the detection can be considered successful for the purpose of this study. The chosen metric (mAP@50) is well-suited for this scenario, as it evaluates the model's performance based on a more lenient Intersection over Union (IoU) threshold of 0.5. In addition, this model is less sensitive to cell size variations compared to higher IoU thresholds. It is important to note that while mAP@50 is the primary focus, the other mAP variants are not disregarded. These metrics provide valuable supplementary information about the models' performance, particularly in cases where more precise localization may be beneficial for downstream analysis or future applications.

While mAP serves as the primary quantitative metric, the evaluation process also considers several qualitative and practical aspects:

1. Inference speed: Assessment of the model's processing time per image, which is crucial for potential real-time applications in microscopy workflows and high-throughput screening.
2. Model size and computational requirements: Evaluation of the trained models' memory footprint and computational demands, considering potential deployment constraints in research or clinical settings.
3. Sample detections: Curated sets of images showcasing successful detections, failure cases, and comparisons between models, providing qualitative insights into model behavior.

The comprehensive evaluation methodology described in this chapter aims to provide a thorough and nuanced understanding of the impact of incorporating generated images into the training data for single cell detection models. By combining quantitative metrics, and qualitative assessments, this approach enables a robust evaluation of the potential benefits and limitations of leveraging unconditional diffusion-based image generation in the context of brightfield microscopy cell detection. The results obtained through this evaluation process will directly address the research question (chapter 1.3), shedding light on whether and to what extent generated images can enhance object detection accuracy or reduce the reliance on large datasets of real images. These findings have significant implications for the fields of biological image analysis and machine learning, potentially offering new strategies for developing robust cell detection models with limited real-world data.

7 Results

This chapter presents the results of the conducted survey (Chapter 5.4.2), visual image differences, and the performance of the trained models on the real and synthetic datasets (Chapter 6.4). The first section delves into the survey results, providing detailed insights into the participants' ability to distinguish between real and synthetic images. The second section offers a visual comparison of the generated and real images, highlighting key differences in color, contrast, brightness, and texture. The third section will present the performance of the trained models (as described in Chapter 6.2) on the various datasets (outlined in Chapter 6.1).

7.1 Survey Results

The survey engaged 11 imaging experts and biologists from Synentec GmbH, leveraging their professional expertise in microscopy and cell biology. Each participant was tasked with classifying 30 images, resulting in a total of 330 individual classifications.

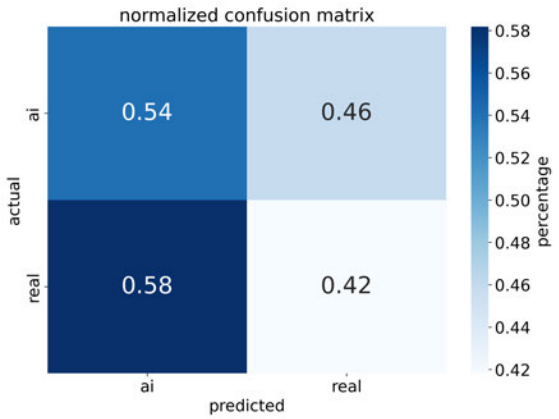


Fig. 7.1: Normalized confusion matrix showing the distribution of correct and incorrect classifications for real and generated microscopy images.

The confusion matrix (Fig. 7.1) offers a concise overview of the participants' ability to distinguish between real and generated microscopy images. Out of 330 total classifications, 165 were correct, yielding an overall accuracy of 50%. This even split between correct and incorrect classifications suggests a significant challenge in reliably differentiating between the two image types. Notably, 54% of generated images were correctly identified as synthetic, while 46% were misclassified as real. Conversely, only 42% of real images were correctly identified, with 58% misclassified as synthetic.

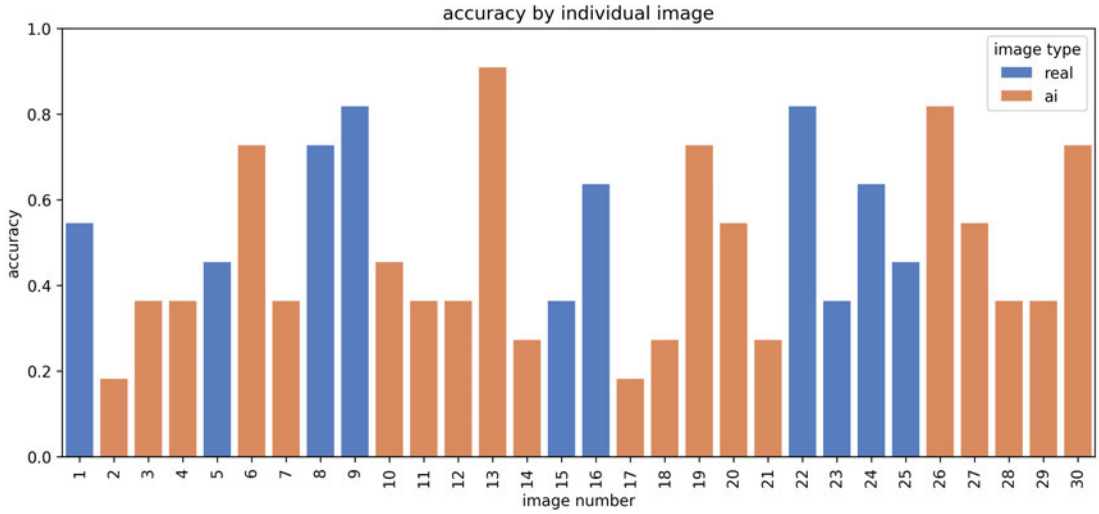


Fig. 7.2: Bar chart displaying accuracy for individual AI-generated and real images.

The bar chart (Fig. 7.2) provides a more nuanced view of classification accuracy for individual images. The data reveals substantial variation across the 30 images, with accuracies ranging from approximately 18% to 90%. This wide range underscores the complexity of the task and the diverse characteristics of the images. Only four images (numbers 9, 13, 22, and 26) achieved an accuracy of 80% or higher, suggesting these images possessed distinctive features that made them more easily classifiable. Conversely, five images (numbers 3, 14, 17, 18, and 21) had accuracies of 30% or lower. The majority of images fell within the 35% to 60% accuracy range, with no discernible pattern emerging, even when differentiating by image type. This distribution further emphasizes the difficulty participants faced in consistently distinguishing between real and generated images.

The word cloud visualization (Fig. 7.3) offers valuable insights into the specific features and characteristics that influenced participants' classifications. The size of each term corresponds to its frequency in the participants' responses, with larger words indicating

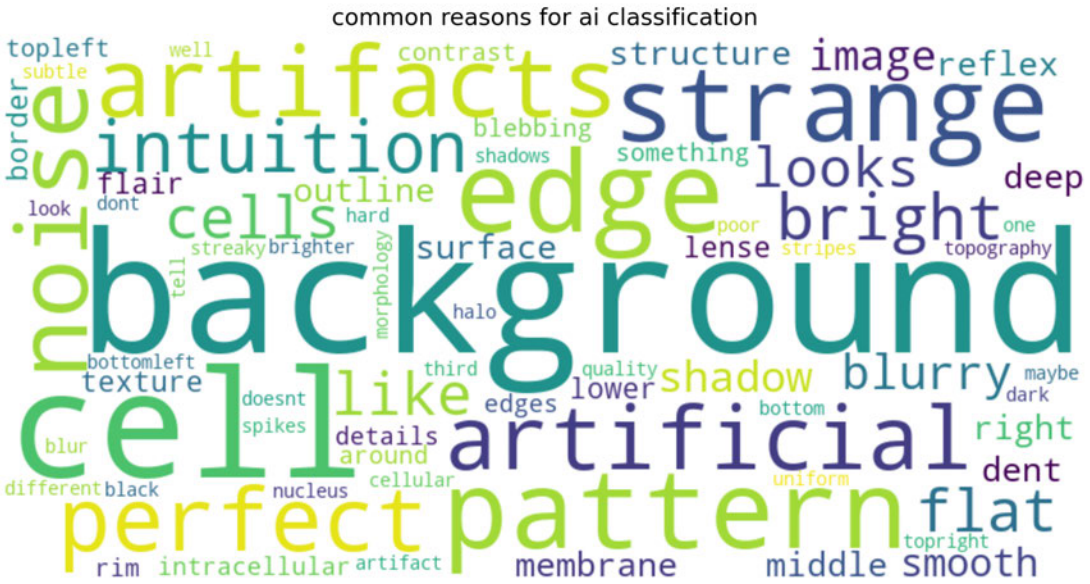


Fig. 7.3: Word cloud depicting common reasons for classifying an image as generated.

more common reasons for classifying an image as generated. The prominence of terms such as “cell”, “background”, and “edge” indicates their frequency in participants’ explanations and highlights the key areas of focus during the decision-making process. Other notable terms like “pattern”, “perfect”, “artifacts” and “noise” suggest common features associated with generated images. These terms provide a window into the participants’ mental models and the visual cues they relied upon to make their judgments. The word cloud also reveals clusters of related terms that offer additional context. Words such as “blurry”, “quality”, “shadows” and “contrast” indicate that image quality and environmental factors played a significant role in participants’ assessments. This suggests that the overall visual fidelity and realistic imperfections were important considerations in distinguishing between real and generated images. Similarly, terms like “pattern”, “structure”, “smooth”, and “surface” imply that participants were attuned to the spatial distribution and arrangement of visual elements, possibly looking for unnatural regularities or repetitions that might indicate a synthetic origin. The prominence of “background” and related terms like “noise” and “artifact” underscores the importance of the image context in the classification process. Participants were not solely focused on the cells themselves but also considered the surrounding areas and overall image characteristics. This comprehensive evaluation of images captures the intricacies involved in the task and highlights the diverse visual elements that influenced the decision-making.

In conclusion, the survey results reveal the intricate challenge of distinguishing between real and generated microscopy images of single cells. The near-even split in overall accuracy, combined with the wide range of individual image accuracies and the diverse terminology used by participants, demonstrates the sophistication of the generated images and the difficulty of the classification task. These findings highlight the potential of unconditional diffusion-based image generation techniques to produce highly realistic microscopy images that can challenge even expert observers. The results also underscore the importance of considering multiple visual aspects when evaluating image authenticity and the potential need for more advanced tools or training to reliably differentiate between real and synthetic microscopy images in scientific contexts.

7.2 Visual Comparison of Generated and Real Images

The comparison between generated images (Appendix A.1) and real brightfield microscopy images (Appendix A.2) reveals subtle differences. This analysis employs both visual inspection and quantitative metrics to provide a comprehensive evaluation of the unconditional diffusion model's performance in replicating brightfield microscopy images. The metric-based approach includes average brightness, average contrast, color channel ratios, and a color bias metric, which complement and substantiate the visual observations.

The color bias metric is calculated as the difference between the highest and lowest color channel ratios. The metrics for the generated images were calculated using the 2500 images included in the *scc_50* dataset, whereas the metrics for the real images were calculated using the entire *scc_real* dataset. The images were converted from a float range of 0–1 to an integer range of 0–255 for the metric calculations.

One of the most striking differences is the overall color palette. The generated images exhibit a faint of greenish tint that seems to be overlaid on top of the actual images. This observation is supported by the metric-based analysis, which reveals a very small color bias of 0.0008 in the generated images compared to 0.0000 in the real images. Although this difference is small, it aligns with the perceived color tint. The color channel ratios (R: 0.3330, G: 0.3338, B: 0.3332) for generated images deviate slightly from the ideal grayscale distribution of 0.3333 for each channel, further confirming the subtle color variations introduced by the diffusion model.

In contrast, the real microscopy images maintain a consistent grayscale palette, typical of standard brightfield microscopy, as evidenced by their perfect 0.3333 ratio across all color channels. This difference in color range suggests that the diffusion model may have introduced some color variations not present in the original training data.

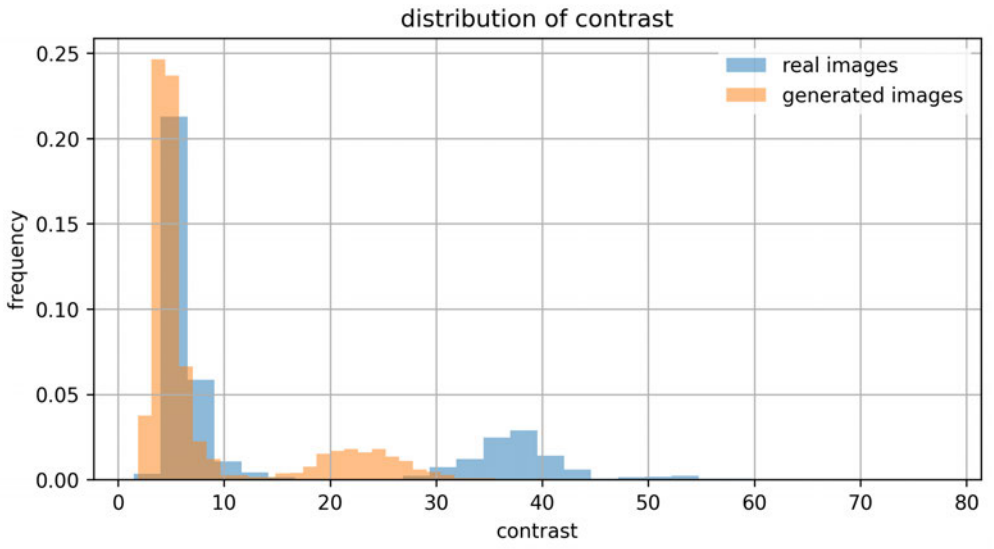


Fig. 7.4: Density histogram depicting the contrast levels of generated and real images.

The contrast levels also differ between the two sets, both visually and metrically (Fig. 7.4). The generated images have a lower average contrast value of 8.21, whereas real images have an average contrast of 14.51, so nearly twice as high. This discrepancy is also slightly visible in the generated images, which appear a little bit washed out or lacking in detail compared to the real images. The real images, on the other hand, exhibit a higher contrast range, with more pronounced differences between light and dark areas, hinting at clearer and more defined cell boundaries.

Brightness levels follow a similar pattern to contrast (Fig. 7.5). The metric-based analysis reveals an average brightness of 206.1 for real images compared to 132.21 for generated images, indicating a noticeable overexposure in the real images. Ideally, the brightness levels should be scattered around the middle of the available range (128) to avoid overexposure or underexposure for grayscale images with 8-bit depth (pixel values of 0–255). The overexposure in the real images can lead to loss of detail and information in the brighter areas, which is less likely in the generated images due to their better balanced brightness levels.

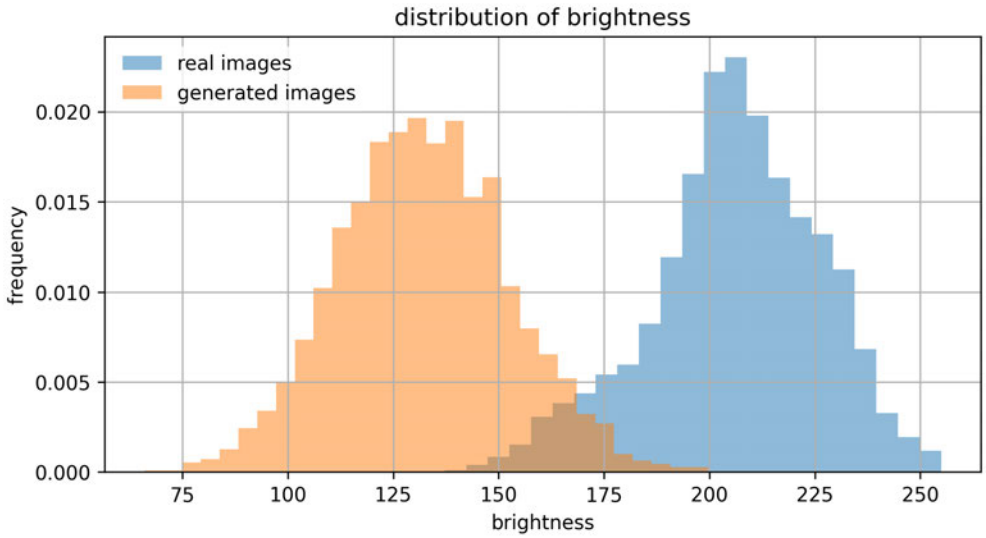


Fig. 7.5: Density histogram illustrating the brightness levels of generated and real images.

The texture is another area where differences are apparent. The generated images sometimes display a subtle, almost canvas-like structure or noise across the background, which is rarely that noticeable in the real microscopy images. This added texture in the generated set might be an artifact of the diffusion process or an attempt by the model to replicate microscopic details it interpreted from the training data.

Lastly, as mentioned in chapter 5.1.2, the model was able to reliably generate images with the same 16 pixel black padding. This padding can be generated on every side and serves as a clear indicator of the model's ability to replicate this specific feature, ensuring consistency across the synthetic and real dataset.

These differences, both visual and metric-based, highlight both the capabilities and current limitations of the unconditional diffusion model in replicating brightfield microscopy images. While the model has captured many key features, there are still slight differences that distinguish the generated images from their real counterparts.

The metric-based analysis provides quantitative support for many of the visual observations, particularly regarding brightness, contrast, and color variations. However, it also reveals that some perceived differences, such as the color tint, are subtler than they might appear visually, giving an initial explanation for the difficulty in distinguishing between real and generated images in the survey. Generally, this underscores the value of combining both approaches for a more comprehensive evaluation.

7.3 Single Cell Detection Performance

A detailed comparison of the performance metrics for all evaluated models across the different training datasets can be found in Tab. 7.1. Several key observations can be made from the performance data:

1. Inference Speed: The YOLOv8 models demonstrate the fastest inference times, ranging from approximately 17ms to 21ms. YOLOv9 models show slightly slower inference, with the compact version at 23ms and the efficient version at 34ms. The RT-DETR models have the slowest inference times, with the large version at 34ms and the extra-large at 40ms.
2. Model Size and Parameters: As expected, there is a trade-off between model size and performance. The YOLOv8s model has the smallest footprint at 42.7MB, while the RT-DETR-x model is the largest at 257.9MB. This variation in model size corresponds to the number of parameters, ranging from 11.2M for YOLOv8s to 68.2M for YOLOv8 extra-large.
3. mAP Performance: The mAP metrics reveal interesting patterns across the different IoU thresholds and training datasets. Generally, the models trained on the real dataset slightly outperform those trained on synthetic datasets, particularly at higher IoU thresholds. However, the performance gap narrows at lower IoU thresholds, with some synthetic-trained models achieving comparable or even slightly better results in certain cases.

Analyzing the requirements of a specific use case is crucial when selecting the most suitable model for single cell detection. The trade-offs between inference speed, model size, and detection accuracy must be carefully considered to align with the project's constraints and objectives. Ultimately, the choice of model should be tailored to the specific needs of the application, balancing performance metrics with practical considerations such as computational resources and real-time processing requirements.

Tab. 7.1: Detailed comparison of the performance metrics for all evaluated models across the different training datasets.

Model	Dataset	Inference Speed	Total Params	Model Size	mAP @50	mAP @75	mAP @50:95
YOLOv8s	*scc_real*	~ 17ms	11.2M	42.7MB	0.8947	0.8095	0.6557
	scc_10				0.8941	0.8053	0.6470
	scc_30				0.9043	0.8079	0.6448
	scc_50				0.8932	0.7891	0.6325
YOLOv8m	*scc_real*	~ 19ms	25.9M	98.9MB	0.9038	0.8203	0.6637
	scc_10				0.9035	0.8093	0.6558
	scc_30				0.8945	0.8076	0.6462
	scc_50				0.8930	0.8040	0.6456
YOLOv8x	*scc_real*	~ 21ms	68.2M	260.5MB	0.9038	0.8120	0.6639
	scc_10				0.9032	0.8055	0.6531
	scc_30				0.9031	0.8085	0.6549
	scc_50				0.8935	0.7961	0.6425
YOLOv9c	*scc_real*	~ 23ms	25.5M	97.8MB	0.9047	0.8215	0.6645
	scc_10				0.9042	0.8070	0.6531
	scc_30				0.9042	0.8106	0.6568
	scc_50				0.8935	0.7951	0.6422
YOLOv9e	*scc_real*	~ 34ms	58.1M	222.4MB	0.9037	0.8201	0.6629
	scc_10				0.9041	0.8141	0.6650
	scc_30				0.8946	0.8069	0.6485
	scc_50				0.8938	0.8037	0.6453
RT-DETR-l	*scc_real*	~ 34ms	33.0M	126.0MB	0.9146	0.8169	0.6614
	scc_10				0.9147	0.8071	0.6574
	scc_30				0.9017	0.7780	0.6240
	scc_50				0.9036	0.7843	0.6298
RT-DETR-x	*scc_real*	~ 40ms	67.5M	257.9MB	0.9164	0.8257	0.6748
	scc_10				0.9144	0.8083	0.6565
	scc_30				0.9032	0.7830	0.6328
	scc_50				0.9045	0.8032	0.6437

7.3.1 Quantitative Metrics

mAP@50

Fig. 7.6 illustrates the mAP@50 values for all models across the four datasets. At this lower IoU threshold, the performance differences between models trained on real and synthetic datasets are relatively small. In fact, some models trained on synthetic data achieve marginally higher mAP@50 scores than their counterparts trained on real data. For instance, the RT-DETR-l model trained on the 10% synthetic dataset achieves the highest mAP@50 of 0.9147, slightly outperforming the same model trained on real data (0.9146). Similarly, the YOLOv9e model trained on the 10% synthetic dataset achieves a

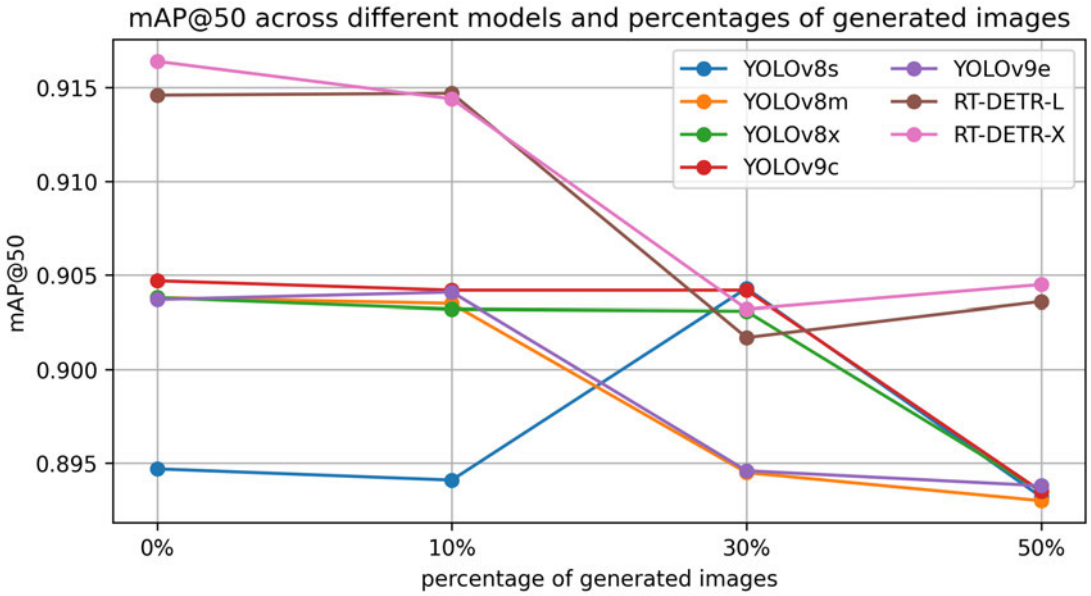


Fig. 7.6: mAP@50 values for all models trained on the four datasets.

mAP@50 of 0.9041, marginally higher than the 0.9037 achieved by the real-data trained version. The YOLOv8s model trained on the 30% synthetic dataset also outperforms its real-data counterpart by roughly 1%. The largest difference between a model trained on the real data and a model trained on synthetic data is 1.19% (RT-DETR-x). Generally, the mAP@50 values for all models are relatively high, indicating strong performance in detecting cells at this IoU threshold, with all models achieving scores above 89%. Also interesting to observe is the visible grouping of the models, especially on the real and 10% synthetic datasets. As expected by models sizes and architectures, YOLOv8s performs

the worst, whereas the other two YOLOv8 models and both YOLOv9 models perform similarly. The RT-DETR models, on the other hand, start as the best performing models, but show the greatest performance drop, particularly on the 30% synthetic dataset.

mAP@75

The mAP@75 metric, shown in Fig. 7.7, reveals a more pronounced difference between models trained on real and synthetic datasets. At this higher IoU threshold, models trained on real data consistently outperform their synthetic-trained counterparts across all architectures. The RT-DETR-x model trained on real data achieves the highest

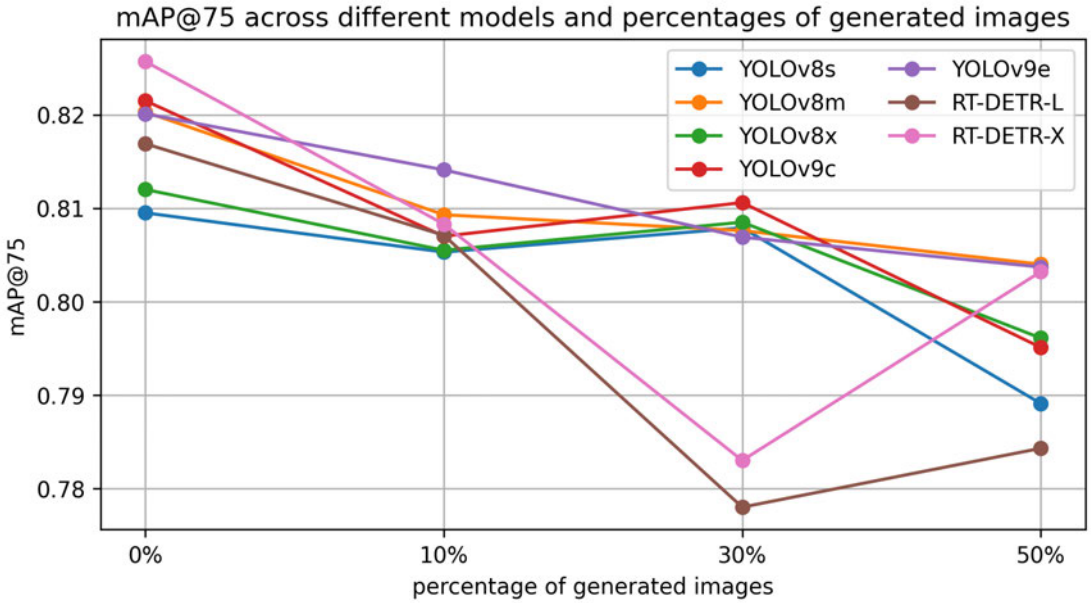


Fig. 7.7: mAP@75 values for all models trained on the four datasets.

mAP@75 of 0.8257, followed closely by the YOLOv9 compact model at 0.8215. The performance gap between real and synthetic-trained models becomes more evident, with the synthetic models generally showing a decrease in mAP@75 as the proportion of generated images in the training set increases. The largest gap between a real and synthetic-trained model grew to 4.27% (RT-DETR-x). Additionally, both transformer models (RT-DETR) perform better on 50% synthetic data compared to 30% synthetic data. As for the model grouping visible in the mAP@50 results, the same pattern can not be observed here. All YOLO models perform similarly on all datasets, whereas

the RT-DETR models again show a significant performance drop on the 30% synthetic dataset.

mAP@50:95

Lastly, the mAP@50:95 metric, illustrated in Fig. 7.8, provides the most comprehensive view of model performance across a range of IoU thresholds. This metric further reinforces the observations made from the mAP@75 analysis. The RT-DETR-x model trained on real data achieves the highest mAP@50:95 of 0.6748, demonstrating its superior performance across various IoU thresholds. Interestingly, the YOLOv9e model trained on the

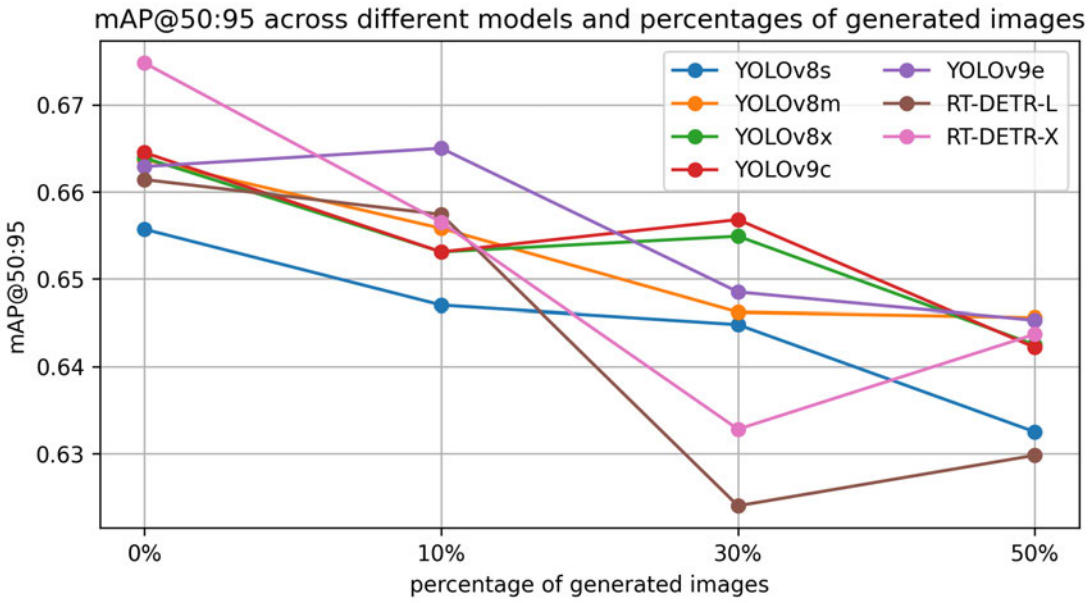


Fig. 7.8: mAP@50:95 values for all models trained on the four datasets.

10% synthetic dataset achieves the second-highest mAP@50:95 of 0.6650, slightly outperforming its real-data trained counterpart (0.6629). As expected by now, the RT-DETR-x model again shows the largest performance gap between real and synthetic-trained models, with a difference of 4.2%. And again, the RT-DETR models show a significant performance drop on the 30% synthetic dataset, which slightly increases again on the 50% synthetic dataset (observable for all three metrics).

7.3.2 Qualitative Assessment

This chapter presents a detailed analysis of the performance of three object detection models —YOLOv8s, YOLOv9e, and RT-DETR-l —on sample batches from the single-cell detection datasets (Appendix B). The analysis focuses on the models' ability to accurately detect and localize cells in brightfield microscopy images, highlighting both their strengths and limitations. The limitations of this thesis do not allow for a more detailed analysis of sample detections of every trained model.

YOLOv8s Sample Detections

The YOLOv8s model, despite being the smallest and least complex of the three architectures examined, demonstrated impressive performance on the *scc_30* dataset. Fig. B.1 illustrates the ground truth labels and model predictions for a representative sample batch. The ground truth labels, as depicted in Fig. B.1a, showcase the complexity of the cell detection task. The labeled images contain a variety of cell presentations, including:

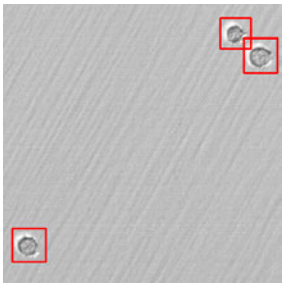


Fig. 7.9: Overlapping boxes.

1. Fully visible cells within the image boundaries
2. Partially visible cells at the image edges
3. Overlapping cells with intersecting bounding boxes

This diversity in cell presentation poses significant challenges for object detection models, requiring them to accurately identify and localize cells under various conditions. Despite its relative simplicity, the YOLOv8s model exhibited remarkable performance on the sample batch. As shown in Fig. B.1b, the model successfully detected all cells present in the images with 100% confidence. This high level of confidence across all detections suggests that the model has learned robust features for cell identification, even in challenging scenarios such as partial visibility and overlapping cells.

YOLOv9e Sample Detections

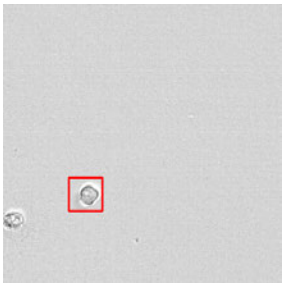


Fig. 7.10: Missing box.

The YOLOv9e model, trained on the *scc_10* dataset, also showcased strong performance in cell detection tasks. Fig. B.2 presents the ground truth labels and model predictions for a sample batch from this datasets' test split. An important observation from the ground truth labels (Fig. B.2a) is the presence of a labeling inconsistency. Specifically, in the image located in the first row and third column, one cell lacks a bounding box label. This missing label highlights a critical aspect of real-world machine learning applications: datasets are often imperfect, and models must learn to generalize effectively despite such inconsistencies. The sample batch for the YOLOv9e model predominantly features fully visible cells, which may present a somewhat easier detection scenario compared to the YOLOv8s sample batch. This difference in sample characteristics underscores the importance of diverse and representative datasets in training robust object detection models. The YOLOv9e model demonstrated excellent performance on the sample batch, as evidenced in Fig. B.2b. The model also successfully detected all cells with high confidence, including the cell with the missing label. This suggests that the YOLOv9e architecture has developed a strong internal representation of cell features, allowing it to generalize beyond the provided labels. Such robustness is crucial for real-world applications where perfect labeling cannot be guaranteed.

RT-DETR-l Sample Detections

The RT-DETR-l model, also trained on the *scc_10* dataset, exhibited strong performance in cell detection tasks. Fig. B.3 illustrates the ground truth labels and model predictions for a sample batch. The sample batch for this model represents a combination of the previous two batches, by including partially visible cells at image boundaries, multiple cells per image and a missing label for one of the cells (Fig. B.3a). Key observations of the models' performance on this sample batch are as follows (Fig. B.3b):

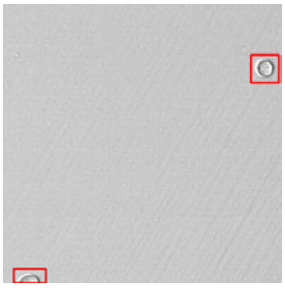


Fig. 7.11: Split box.

- Successful detection of all visible cells, including those with missing labels

- High confidence scores for most detections
- Detection of the cell with the missing label, albeit with a lower confidence score of 0.3
- A false positive detection at the left border of one image, also with a confidence score of 0.3

The model's ability to detect the cell with the missing label, similar to the YOLOv9e model, indicates robust feature learning and generalization capabilities. However, the lower confidence score for this detection suggests some uncertainty, which is reasonable given the lack of a corresponding ground truth label during training. The potential false positive detection at the image border with a low confidence score of 0.3 highlights an important consideration in object detection tasks: the trade-off between sensitivity and specificity. While high sensitivity is crucial for detecting all cells, it may occasionally lead to false positives, especially in challenging areas such as image boundaries where partial cells or artifacts may be present.

Comparative Analysis

Comparing the performance of the three models —YOLOv8s, YOLOv9e, and RT-DETR-l —on their respective sample batches reveals several important insights:

- All three models demonstrated strong cell detection capabilities, successfully identifying cells under various conditions, including partial visibility and overlapping instances.
- The similar performance of the three models is consistent with their comparable mAP scores, confirming that evaluating models based on quantitative metrics and qualitative assessments can provide a comprehensive view of their capabilities.
- The models showed robustness to labeling inconsistencies, detecting cells even when ground truth labels were missing. This suggests effective feature learning and generalization beyond the provided annotations.
- The YOLOv8s model, despite being the smallest and least complex, achieved 100% confidence in its detections on the sample batch.
- This high confidence, however, should be interpreted cautiously and verified across a larger, more diverse sample batch to ensure it's not a result of a particularly easy sample batch.

- The YOLOv9e and RT-DETR-l models, trained on the *scc_10* dataset, showed similar performance in terms of detection accuracy. However, the RT-DETR-l model exhibited more nuanced confidence scoring, particularly for challenging cases like the cell with the missing label and the potential false positive at the image border.
- The potential false positive detection by the RT-DETR-l model highlights the ongoing challenge in object detection of balancing sensitivity and specificity. This trade-off is particularly crucial in medical and biological applications where both false negatives and false positives can have significant consequences.

These observations underscore the importance of comprehensive evaluation beyond simple accuracy metrics. Factors such as confidence calibration, robustness to labeling errors, and performance on edge cases (e.g. partially visible cells) play crucial roles in determining a model's suitability for real-world applications in cell detection and analysis.

8 Discussion

The results presented in the previous chapter offer valuable insights into the research questions (Chapter 1.3) and have significant implications for the fields of machine learning, computer vision, and computational biology.

8.1 Interpretation of Results

The survey results involving imaging experts and biologists revealed a remarkable level of difficulty in distinguishing between real and generated microscopy images. With an overall accuracy of 50%, the survey demonstrates the high quality and realism achieved by the unconditional diffusion model in generating synthetic brightfield microscopy images. This finding aligns with recent advancements in generative AI, particularly in the domain of image synthesis, where state-of-the-art models have shown remarkable capabilities in replicating complex visual patterns, textures, and sceneries. The results of the survey further underscore the challenge faced by human experts in this task, with only a handful of images being consistently classified correctly. A 50% accuracy rate resembles random guessing, giving a clear answer to the first sub research question (SRQ1): the synthetic images are of sufficient quality to be indistinguishable from real images to human experts.

Additionally, the word cloud analysis of participants' reasoning provides valuable insights into the visual cues and features that experts rely on when evaluating microscopy images. The prominence of terms such as "cell", "background", and "edge" indicates that both the primary subjects (cells) and the overall image context play crucial roles in the assessment process. The frequent mention of image quality-related terms (e.g., "quality", "blurry", "contrast") suggests that the diffusion model has successfully replicated not only the content but also the typical imperfections and variations found in real microscopy images. This level of detail contributes significantly to the realism of the generated images and

highlights the potential of diffusion models in creating diverse and representative synthetic datasets for machine learning applications in microscopy. The visual comparison between generated and real images revealed subtle but noteworthy differences, particularly in color tones, brightness and contrast levels. The more equally balanced brightness values in the generated images, might actually improve the performance of object detection models. Whereas lower average contrast values could potentially decrease the performance, as the models might struggle to detect cells with less pronounced edges. Additionally, color tints pose a significant problem, since brightfield microscopy is by definition grayscale. Using synthetic images with color tints in combination with real images introduces a bias to the model, that is not present in the real data. There is a need to further investigate the actual impact of these color tints on the performance of object detection models. These issues also underscores the importance of continued refinement in generative models to more accurately capture the specific visual properties of brightfield microscopy. Generally, the differences are hard to spot for the human eye, but might be of technical relevance and importance for the performance of object detection models.

The performance analysis of various object detection models trained on real and synthetic datasets yielded several significant findings. Generally, models trained on synthetic datasets achieved comparable performance to those trained on real data, particularly at lower IoU thresholds. This suggests that synthetic images can serve as a viable alternative or supplement to real data for training object detection models in microscopy applications. The observation that some models trained on synthetic data marginally outperformed their counterparts trained on real data at mAP@50 is particularly encouraging, indicating the potential of synthetic data to enhance model robustness and generalization. However, the performance gap between real and synthetic-trained models became more pronounced at higher IoU thresholds, as evidenced by the mAP@75 and mAP@50:95 metrics. This trend suggests that while synthetic data can effectively capture general cell features and locations, it may fall short in replicating the precise boundaries and fine details of cells. This limitation could be attributed to subtle differences in cell morphology or edge definition between real and generated images, highlighting an area for future improvement in the image generation process. Generally, a slight performance decrease was observable when increasing the percentage of synthetic images in the training set. This decrease was more pronounced in the case of the RT-DETR models, indicating a greater impact of synthetic data. These results suggest that sub research question two (SRQ2) can be answered by stating that the proportion of synthetic

data in the training set slightly influences the magnitude and direction of its impact on model performance. Especially, the transformer-based models seem more sensitive to the proportion of synthetic data in the training set. But it is important to note that the difference in performance is limited to a few percentage points and might not be that noticeable in real-world applications.

The qualitative assessment of sample detections from YOLOv8s, YOLOv9e, and RT-DETR-l models provided valuable insights into their practical performance. All three models demonstrated strong cell detection capabilities across various challenging scenarios, including partially visible and overlapping cells. Their ability to detect cells even in the presence of labeling inconsistencies is particularly noteworthy, suggesting robust feature learning and generalization beyond the provided annotations. This resilience to imperfect ground truth data is crucial for real-world applications where labeling errors are common. But these sample detections and their results have to be analyzed with a grain of salt, as the sample size is small and might not represent the actual distribution of the data. They should be analyzed and evaluated in combination with the quantitative results.

The observed differences in confidence scoring between models, particularly the more nuanced approach of the RT-DETR-l model, highlight the importance of considering not just detection accuracy but also the reliability of confidence estimates. The potential false positive detection by RT-DETR-l at an image border underscores the ongoing challenge of balancing sensitivity and specificity in object detection tasks, especially in medical and biological applications where both false negatives and false positives can have significant consequences.

To sum everything up and answer the central research question (RQ): The introduction of synthetic brightfield microscopy images to a training dataset has an impact on the performance, although the magnitude of this impact is very limited and not always positive. Depending on the object detection model, the proportion of synthetic data in the training set, and the evaluation metric, the performance can slightly increase or decrease. The results suggest that synthetic data can be a valuable tool for training object detection models in microscopy applications, providing a cost-effective and scalable alternative to real-world datasets. However, further refinement of the image generation process and model training strategies is necessary to fully leverage the potential of synthetic data and really improve the accuracy and robustness of cell detection models.

8.2 Impact on Biological Research and Applications

The findings of this study have significant implications for biological research and applications, particularly in the realm of brightfield microscopy and cell analysis. By demonstrating the viability of using synthetic brightfield microscopy images for training object detection models, this research opens up new possibilities for advancing scientific understanding and improving experimental methodologies.

One of the most immediate impacts is the potential for research groups and laboratories to generate their own data, given sufficient computational resources and expertise. This capability could lead to a substantial time and cost savings, as well as increased flexibility in experimental design and data acquisition. The ability to generate synthetic data is particularly valuable for research groups with limited access to high-quality microscopy images, democratizing advanced cell detection techniques and potentially accelerating scientific progress in resource-constrained environments.

The use of generated images also has important implications for cell viability and experimental design. By reducing the need for fluorescence dyes, which can be harmful to cells, researchers can minimize potential confounding factors in their experiments. This approach aligns with the principles of ethical research and may lead to more accurate results in studies of cell behavior and function.

Another significant advantage of synthetic image generation is the ability to create scenarios and cell configurations that are challenging to achieve in laboratory settings or occur rarely in nature. This capability addresses a common limitation in machine learning applications, where rare events are often underrepresented in training data. By specifically generating these uncommon scenarios, researchers can develop more robust and generalizable cell detection models, potentially improving performance in real-world applications where such rare events may be critical.

The technique also offers the possibility of generating many images of rare or expensive cell lines in silico. This could dramatically reduce costs associated with certain types of research and enables studies that might otherwise be too expensive or logistically challenging to conduct. Furthermore, the ability to generate diverse datasets could accelerate the process of assay development, drug discovery, and analysis of cell behavior, potentially leading to faster breakthroughs in biomedical research and pharmaceutical development.

However, it is crucial to note that transparency is paramount when using generated images in research. To main scientific integrity and avoid ethical issues, researchers should clearly disclose the use of synthetic data in their studies. This transparency will ensure that the scientific community can properly evaluate the validity and generalizability of findings based on synthetic data.

In conclusion, the impact of this research on biological studies and applications is far-reaching. By providing a method to generate high-quality synthetic microscopy images, this work paves the way for more efficient, cost-effective, and ethically sound research practices. As the technique is further refined and adopted, it has the potential to accelerate scientific discoveries and innovation in cell biology and related fields.

8.3 Limitations and Future Directions

While this study has yielded promising results, several limitations must be acknowledged, and some paths for future research remain to be explored.

The most significant limitations and opportunities for improvement lie in the image generation process. The current architecture of the diffusion model, which produces multi-channel images (RGB), introduces undesirable color tints that are inconsistent with true brightfield microscopy. Future work should focus on adjusting the model architecture to accept and generate single-channel grayscale images only. This modification would likely improve the fidelity of the generated images and potentially enhance the performance of object detection models trained on synthetic data.

The generalizability of the trained diffusion model is another area of concern. Currently, the model is tailored to a specific biological application and dataset, which limits its broader applicability. Future research should explore the development of more versatile models, possibly through the use of conditional diffusion techniques, that can generate images for a wider range of biological applications and cell types. This would involve gathering and incorporating real microscopy images from diverse sources, including different microscopes and cell lines, to create a more comprehensive training dataset.

The computational requirements for training the diffusion model posed a significant constraint in this study. Future work should investigate the relationship between the amount of training data and the quality of generated images, to determine the optimal balance between computational resources and image fidelity. Additionally, exploring alternative

architectures for the diffusion model could potentially yield improvements in both efficiency and output quality.

Regarding the evaluation of generated images, it is important to note that the experts and biologists involved in the survey were aware that some images were generated. This knowledge may have influenced their judgements, potentially biasing the results. Future studies that also employ surveys should consider implementing double-blind evaluation protocols and more participants to eliminate this potential source of bias.

In terms of object detection, there is moderate room for improvement in both data quality and model performance. Enhancing the quality of label data should be a priority, as this could significantly impact the accuracy of trained models. Furthermore, a more extensive exploration of the hyperparameter space for the various object detection models could yield performance improvements. Future research should also consider evaluating additional state-of-the-art object detection models to identify the most effective approaches for cell detection in brightfield microscopy images.

The performance discrepancies observed between models trained on real and synthetic data, particularly at higher IoU thresholds, suggest that further refinement of the image generation or object detection process is necessary. Future work should consider improving the replication of fine cellular details and precise boundaries in generated images. This could involve developing more sophisticated loss functions or incorporating additional biological constraints into the generation process.

Additionally, due to time constrains, as mentioned in Chapter 5.1.2, image patches containing the well edge were removed from the initial dataset. Hence, neither the diffusion model was able to generate them, nor the object detection models were trained on them. Since these artifacts are common in brightfield microscopy, cells are often located, and are especially hard to detect near the well edge, future research should explore methods to incorporate these challenging scenarios into the training process.

Lastly, while this study focused on brightfield microscopy, future research could further explore the application of similar techniques to other imaging modalities in biology and medicine. This could include fluorescence microscopy, electron microscopy, or even medical imaging techniques such as MRI or CT scans.

By pursuing the suggested research directions, future studies can build upon this work to further advance the field of computational biology and enhance the capabilities of AI-assisted microscopy analysis.

9 Conclusion

This thesis set out to investigate the impact of leveraging unconditional diffusion-based image generation on single cell detection accuracy in brightfield microscopy images. The primary objective was to evaluate whether the introduction of synthetic brightfield microscopy images to a training dataset could positively impact the performance of object detection models in single cell detection tasks.

9.1 Summary of Findings

The research yielded several significant findings. First, the survey results involving imaging experts and biologists revealed a remarkable level of difficulty in distinguishing between real and generated microscopy images, with an overall accuracy of 50%. This demonstrates the high quality and realism achieved by the unconditional diffusion model in generating synthetic brightfield microscopy images.

Second, the performance analysis of various object detection models trained on real and synthetic datasets showed that models trained on synthetic datasets achieved comparable performance to those trained on real data, particularly at lower IoU thresholds. Some models trained on synthetic data even marginally outperformed their counterparts trained on real data at mAP@50. However, the performance gap between real and synthetic-trained models became more pronounced at higher IoU thresholds and increased percentage of synthetic data.

Third, the visual comparison between generated and real images revealed subtle but noteworthy differences, particularly in color tones, contrast levels, and background textures. While these differences highlight some limitations of the current image generation process, the increased variability in the synthetic images may actually prove beneficial for developing more robust cell detection models.

9.2 Outlook and Recommendations

The findings of this research have significant implications for the fields of machine learning, computer vision, and computational biology. By demonstrating the viability of using synthetic brightfield microscopy images for training object detection models, this study opens up new possibilities for advancing scientific understanding and improving experimental methodologies in cell biology and related fields.

The ability to generate high-quality synthetic microscopy images could lead to substantial time and cost savings in research, as well as increased flexibility in experimental design and data acquisition. This capability is particularly valuable for research groups with limited access to high-quality microscopy images, potentially democratizing advanced cell detection techniques and accelerating scientific progress in resource-constrained environments.

However, several limitations and areas for future research remain. The most significant opportunities for improvement lie in the image generation process. Future work should focus on adjusting the model architecture to accept and generate single-channel grayscale images, which could negate the most pronounced visual differences between real and synthetic images: the color tints. Additionally, exploring the development of more versatile models through conditional diffusion techniques could expand the applicability of this approach to a wider range of biological applications and cell types.

Further research should also investigate methods to improve the replication of fine cellular details and precise boundaries in generated images, as well as incorporate challenging scenarios such as cells located near well edges into the training process. These enhancements could help close the performance gap observed between real and synthetic-trained models at higher IoU thresholds.

In conclusion, this thesis has demonstrated the potential of using synthetic brightfield microscopy images for cell detection, while also highlighting the challenges and opportunities for improvement in this emerging field. As generative AI techniques continue to advance, their integration with biological research holds promise for accelerating scientific discoveries and innovation in cell biology and related disciplines. The findings of this study provide a foundation for future research in this exciting intersection of computer science and biology, paving the way for more efficient, cost-effective, and ethically sound research practices in microscopy and cell analysis.

Bibliography

[1] surveyjs/survey-library, July 2024. URL https://github.com/surveyjs/survey-library. last accessed: 2024-09-10.

[2] Ahmed Husham Al-Badri, Nor Azman Ismail, Khamael Al-Dulaimi, Ghalib Ahmed Salman, and Md. Sah Salam. Adaptive Non-Maximum Suppression for improving performance of Rumex detection. *Expert systems with applications*, 219:119634–119634, June 2023. doi: 10.1016/j.eswa.2023.119634. MAG ID: 4318757264 S2ID: 7b08132dab6bb1287ce80eae1f595a176ff15cd5.

[3] Alex Krizhevsky, Alex Krizhevsky, Ilya Sutskever, Ilya Sutskever, Geoffrey E. Hinton, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. 25:1097–1105, December 2012. MAG ID: 2163605009.

[4] Alex Lipp and Pieter Vermeesch. Short communication: The Wasserstein distance as a dissimilarity metric for comparing detrital age spectra and other geological distributions. *Geochronology*, 5(1):263–270, May 2023. doi: 10.5194/gchron-5-263-2023. MAG ID: 4376866702 S2ID: 96dc66b36b31af94248413f43cafe10a69c58421.

[5] Alexander Neubeck, Alexander Neubeck, Luc Van Gool, and L. Van Gool. Efficient Non-Maximum Suppression. *International Conference on Pattern Recognition*, 3:850–855, August 2006. doi: 10.1109/icpr.2006.479. MAG ID: 2144506857 S2ID: 52ca4ed04d1d9dba3e6ae30717898276735e0b79.

[6] Ali Ghaznavi, Ali Ghaznavi, Renata Rychtáriková, Renata Rychtáriková, Mohammadmehdi Saberioon, Mohammadmehdi Saberioon, Dalibor Štys, and Dalibor Štys. Cell segmentation from telecentric bright-field transmitted light microscopy images using a Residual Attention U-Net: A case study on HeLa line. *Computers in Biology and Medicine*, pages 105805–105805, June 2022. doi: 10.1016/j.compbiomed.2022.105805. ARXIV_ID: 2203.12290 MAG ID: 4283688005 S2ID: 932db20b32bdb59faf4a643b9489caa8c6de6420.

[7] Alina Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, Ivan Krasin, J. Pont-Tuset, Shahab Kamali, S. Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and V. Ferrari. the open images dataset v4. *International Journal of Computer Vision*, 2020. doi: 10.1007/s11263-020-01316-z. ARXIV_ID: 1811.00982 MAG ID: 3015310352 S2ID: 5ac18d505ed6d10e8692cbb7d33f6852e6782692.

[8] Alon Saguy, Tav Nahimov, M. Lehrman, Onit Alalouf, and Yoav Shechtman. This microtubule does not exist: Super-resolution microscopy image generation by a diffusion model. *bioRxiv*, July 2023. doi: 10.1101/2023.07.06.548004. MAG ID: 4383532693 S2ID: 92d6b164855ce882e74ee932499c7c92b1447ae0.

[9] Andreas Lugmayr, Martin Danelljan, Andrés Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using Denoising Diffusion Probabilistic Models. *Computer Vision and Pattern Recognition*, June 2022. doi: 10.1109/cvpr52688.2022.01117. ARXIV_ID: 2201.09865 MAG ID: 4312497550 S2ID: 1e91fa21b890a8f5d615578f4ddf46c3cb394691.

[10] Andrew Shepley, Andrew Shepley, Gregory Falzon, Greg Falzon, Gregory Falzon, Paul Kwan, and Paul Kwan. Confluence: A Robust Non-IoU Alternative to Non-Maxima Suppression in Object Detection. *arXiv: Computer Vision and Pattern Recognition*, 2020. doi: 10.1109/tpami.2023.3273210. ARXIV_ID: 2012.00257 MAG ID: 3110079064 S2ID: 0d5d38b6dad5361ca373fbd681902e55f68bbae7.

[11] Anne Cambon-Thomsen, A. Cambon-Thomsen, Emmanuelle Rial-Sebbag, and E. Rial-Sebbag. SERIES "GENETICS OF ASTHMA AND COPD IN THE POSTGENOME ERA" Edited by E. von Mutius, M. Kabesch and F. Kauffmann Number 6 in this Series Trends in ethical and legal frameworks for the use of human biobanks. *European Respiratory Journal*, January 2007. doi: 10.1183/09031936.00165006. MAG ID: 2186909180 S2ID: e744f40a504c1a2211549676a42a81c33a3faf65.

[12] Ashish Vaswani, Ashish Vaswani, Noam Shazeer, Noam Shazeer, Niki Parmar, Niki Parmar, Jakob Uszkoreit, Jakob Uszkoreit, Llion Jones, Llion Jones, Aidan N. Gomez, Aidan N. Gomez, Łukasz Kaiser, Lukasz Kaiser, Illia Polosukhin, and Illia Polosukhin. Attention is All you Need. *Neural Information Processing Systems*, 30:5998–6008, June 2017. ARXIV_ID: 1706.03762 MAG ID: 2963403868 S2ID: 204e3073870fae3d05bcbc2f6a8e263d9b72e776.

[13] Assaf Arbelle, Assaf Arbelle, Tammy Riklin Raviv, and Tammy Riklin Raviv. Microscopy cell segmentation via adversarial neural networks. *IEEE In-*

ternational Symposium on Biomedical Imaging, pages 645–648, April 2018. doi: 10.1109/isbi.2018.8363657. ARXIV_ID: 1709.05860 MAG ID: 2963194837 S2ID: c5bacfa2bbd469f45dc4334729d56894a3257d2d.

[14] Douglas S Auld, Peter A Coassin, Nathan P Coussens, Paul Hensley, Carleen Klumpp-Thomas, Sam Michael, G Sitta Sittampalam, O Joseph Trask, Bridget K Wagner, Jeffrey R Weidner, et al. Microplate selection and recommended practices in high-throughput screening and quantitative biology. *Assay Guidance Manual [Internet]*, 2020.

[15] Ayoub Benali Amjoud and Mustapha Amrouch. Object Detection Using Deep Learning, CNNs and Vision Transformers: A Review. *IEEE Access*, 11:35479–35516, January 2023. doi: 10.1109/access.2023.3266093. MAG ID: 4364323147 S2ID: 7bc29d5b422109a1b85b05a141b9207d9dd937a8.

[16] Barratt, Shane, Shane Barratt, Sharma, Rishi, and Rishi Sharma. A Note on the Inception Score. January 2018. doi: 10.48550/arxiv.1801.01973. MAG ID: 4293775665.

[17] Bo Li, Z. Song, Liujia Shi, Yutao Li, Duli Yu, and Xinglin Guo. Label-free viability detection of T-cells based on 2D bright-field microscopic images and deep learning. *Biophysical Society of Guang Dong Province Academic Forum - Precise Photons and Life Health*, April 2023. doi: 10.1117/12.2673564. MAG ID: 4362550041 S2ID: 01cf87efa4b52b866a5a0d644ea85d99caedd311.

[18] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. YOLOv4: Optimal speed and accuracy of object detection. *ArXiv*, abs/2004.10934, 2020. URL https://api.semanticscholar.org/CorpusID:216080778.

[19] M Boshart, F Weber, G Jahn, K Dorschhler, B Fleckenstein, and W Schaffner. A very strong enhancer is located upstream of an immediate early gene of human cytomegalovirus. *Cell*, 41(2):521–530, June 1985. ISSN 00928674. doi: 10.1016/S0092-8674(85)80025-8. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867485800258.

[20] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[21] Brian B. Moser, Arundhati S. Shanbhag, Federico Raue, Stanislav Frolov, Sebastián M. Palacio, and Andreas Dengel. Diffusion Models, Image Super-Resolution And

Everything: A Survey. *arXiv.org*, 2024. doi: 10.48550/arxiv.2401.00736. ARXIV_-ID: 2401.00736 S2ID: 30d31dd034fd65c0bf4df6f24c6225eab39317f1.

[22] C. Corneanu, Raghudeep Gadde, and Aleix M. Martínez. LatentPaint: Image Inpainting in Latent Space with Diffusion Models. *IEEE Workshop/Winter Conference on Applications of Computer Vision*, 2024. doi: 10.1109/wacv57701.2024.00428. S2ID: f7cc49a184218a83d710aeb907761d99f626d9dc.

[23] Juan C Caicedo, Sam Cooper, Florian Heigwer, Scott Warchal, Peng Qiu, Csaba Molnar, Aliaksei S Vasilevich, Joseph D Barry, Harmanjit Singh Bansal, Oren Kraus, Mathias Wawer, Lassi Paavolainen, Markus D Herrmann, Mohammad Rohban, Jane Hung, Holger Hennig, John Concannon, Ian Smith, Paul A Clemons, Shantanu Singh, Paul Rees, Peter Horvath, Roger G Linington, and Anne E Carpenter. Data-analysis strategies for image-based cell profiling. *Nature Methods*, 14 (9):849–863, September 2017. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.4397. URL http://www.nature.com/articles/nmeth.4397.

[24] John Canny. A Computational Approach toEdgeDetection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, January 1986. doi: 10.1109/tpami.1986.4767851. MAG ID: 2182096904 S2ID: fcf9fc4e23b45345c2404ce7d6cb0fc9dea2c9ec.

[25] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *ArXiv*, abs/2005.12872, 2020. URL https://api.semanticscholar.org/CorpusID:218889832.

[26] Chawin Ounkomol, Chawin Ounkomol, Sharmishtaa Seshamani, Sharmishtaa Seshamani, Sharmishtaa Seshamani, Mary M. Maleckar, Mary M. Maleckar, Forrest Collman, Forrest Collman, Johnson Gr, and Gregory R. Johnson. Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. *Nature Methods*, 15(11):917–920, September 2018. doi: 10.1038/s41592-018-0111-2. MAG ID: 2949493305.

[27] Eric M. Christiansen, Samuel J. Yang, D. Michael Ando, Ashkan Javaherian, Gaia Skibinski, Scott Lipnick, Elliot Mount, Alison O'Neil, Kevan Shah, Alicia K. Lee, Piyush Goyal, William Fedus, Ryan Poplin, Andre Esteva, Marc Berndl, Lee L. Rubin, Philip Nelson, and Steven Finkbeiner. In Silico Labeling: Predicting Fluorescent Labels in Unlabeled Images. *Cell*, 173(3):792–803.e19,

April 2018. ISSN 00928674. doi: 10.1016/j.cell.2018.03.040. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867418303647.

[28] Jan Oscar Cross-Zamirski, Praveen Anand, Guy Williams, Elizabeth Mouchet, Yinhai Wang, and Carola-Bibiane Schönlieb. Class-Guided Image-to-Image Diffusion: Cell Painting from Brightfield Images with Class Labels, March 2023. URL http://arxiv.org/abs/2303.08863. arXiv:2303.08863 [cs, q-bio].

[29] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 1:886–893 vol. 1, 2005. URL https://api.semanticscholar.org/CorpusID:206590483.

[30] Danny Salem, Danny Salem, Yifeng Li, Yifeng Li, Pengcheng Xi, Pengcheng Xi, Hilary Phenix, Hilary Phenix, Miroslava Čuperlović-Culf, Miroslava Cuperlovic-Culf, Mads Kærn, and Mads Kærn. YeastNet: Deep-Learning-Enabled Accurate Segmentation of Budding Yeast Cells in Bright-Field Microscopy. *Applied Sciences*, 11(6):2692, January 2021. doi: 10.3390/app11062692. MAG ID: 3208772529.

[31] Danton Char, Danton S. Char, Nigam H. Shah, Nigam H. Shah, David Magnus, and David Magnus. Implementing Machine Learning in Health Care - Addressing Ethical Challenges. *The New England Journal of Medicine*, 378(11):981–983, March 2018. doi: 10.1056/nejmp1714229. MAG ID: 2789970635 S2ID: adfc508b9b3d4fc3903aa383a290dc68fb8bbe5a.

[32] David Restrepo Amariles, A. Troussel, and Rajaa El Hamdani. Compliance Generation for Privacy Documents under GDPR: A Roadmap for Implementing Automation and Machine Learning. *arXiv.org*, 2020. ARXIV_ID: 2012.12718 S2ID: 0d1e9281676c32bbdf2acbe43ae1d2144913f105.

[33] David Svoboda, David Svoboda, Vladimír Ulman, and Vladimír Ulman. Generation of synthetic image datasets for time-lapse fluorescence microscopy. pages 473–482, June 2012. doi: 10.1007/978-3-642-31298-4_56. MAG ID: 180508781.

[34] David Svoboda, David Svoboda, Vladimír Ulman, and Vladimír Ulman. Towards a Realistic Distribution of Cells in Synthetically Generated 3D Cell Populations. pages 429–438, September 2013. doi: 10.1007/978-3-642-41184-7_44. MAG ID: 261791069.

[35] Dennis Eschweiler and Johannes Stegmaier. Denoising Diffusion Probabilistic Models for Generation of Realistic Fully-Annotated Microscopy Image Data Sets. *PLoS Comput. Biol.*, January 2023. doi: 10.48550/arxiv.2301.10227. ARXIV_ID: 2301.10227 MAG ID: 4318149353 S2ID: 98ae411778f53138070441ec8f98377bca1e8670.

[36] Prafulla Dhariwal and Alex Nichol. Diffusion Models Beat GANs on Image Synthesis, June 2021. URL http://arxiv.org/abs/2105.05233. arXiv:2105.05233 [cs, stat].

[37] Diederik P. Kingma, Diederik P. Kingma, Max Welling, and Max Welling. Auto-Encoding Variational Bayes. *International Conference on Learning Representations*, January 2014. ARXIV_ID: 1312.6114 MAG ID: 1959608418 S2ID: 5f5dc5b9a2ba710937e2c413b37b053cd673df02.

[38] Dominic Waithe, Dominic Waithe, John Brown, Jill M. Brown, Jill M. Brown, Katharina Reglinski, Katharina Reglinski, Katharina Reglinski, Isabel Diez-Sevilla, Isabel Diez-Sevilla, David Roberts, David J. Roberts, Christian Eggeling, Christian Eggeling, Christian Eggeling, and Christian Eggeling. Object detection networks and augmented reality for cellular detection in fluorescence microscopy. *Journal of Cell Biology*, 219(10), 2020. doi: 10.1083/jcb.201903166. MAG ID: 3082961189 S2ID: 6465c0e203ef0cd4ad23cfd65b2b5cf5cc389362.

[39] B. Dwyer, J. Nelson, and T. Hansen. Roboflow (Version 1.0), 2024. URL https://roboflow.com. last accessed: 2024-09-10.

[40] E. D. Ferreira and G. F. Silveira. Classification and counting of cells in brightfield microscopy images: an application of convolutional neural networks. *Scientific Reports*, 2024. doi: 10.1038/s41598-024-59625-z. S2ID: 27078ac4ab60c12c1c98e9aab000c6b3673eff22.

[41] Emma Strubell, Emma Strubell, Ananya Ganesh, Ananya Ganesh, Andrew McCallum, and Andrew McCallum. Energy and Policy Considerations for Deep Learning in NLP. *Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, June 2019. doi: 10.18653/v1/p19-1355. ARXIV_ID: 1906.02243 MAG ID: 2963809228 S2ID: d6a083dad7114f3a39adc65c09bfbb6cf3fee9ea.

[42] Erick Moen, Erick K. Moen, Erick Moen, Dylan Bannon, Dylan Bannon, Tetsuichi Kudo, Takamasa Kudo, William D. Graf, William Graf, Markus W. Covert, Markus W. Covert, David Van Valen, and David Van Valen. Deep learning

for cellular image analysis. *Nature Methods*, 16(12):1233–1246, May 2019. doi: 10.1038/s41592-019-0403-1. MAG ID: 2946901414.

[43] Eyal Betzalel, Coby Penso, and Ethan Fetaya. Evaluation Metrics for Generative Models: An Empirical Study. *Machine Learning and Knowledge Extraction*, 2024. doi: 10.3390/make6030073. S2ID: 161184980c12086b3267f9f00d0e4adb59e1ccf0.

[44] F. Mualla. Automatic Unstained Cell Detection in Bright Field Microscopy (Automatische Detektion ungefärbter Zellen in der Hellfeld-Mikroskopie). 2016. S2ID: 63b5dcdf97b4b69d5112d0e060c79cb6a3ad7064.

[45] William Falcon, Jirka Borovec, Adrian Wälchli, Nic Eggert, Justus Schock, Jeremy Jordan, Nicki Skafte, Ir1dXD, Vadim Bereznyuk, Ethan Harris, Tullie Murrell, Peter Yu, Sebastian Præsius, Travis Addair, Jacob Zhong, Dmitry Lipin, So Uchida, Shreyas Bapat, Hendrik Schröter, Boris Dayma, Alexey Karnachev, Akshay Kulkarni, Shunta Komatsu, Martin.B, Jean-Baptiste SCHIRATTI, Hadrien Mary, Donal Byrne, Cristobal Eyzaguirre, Cinjon, and Anton Bakhtin. PyTorchLightning/pytorch-lightning: 0.7.6 release, May 2020. URL https://zenodo.org/record/3828935.

[46] Felix Buggenthin, Felix Buggenthin, Carsten Marr, Carsten Marr, Michael Schwarzfischer, Michael Schwarzfischer, Philipp S. Hoppe, Philipp S. Hoppe, Oliver Hilsenbeck, Oliver Hilsenbeck, Timm Schroeder, Timm Schroeder, Fabian J. Theis, and Fabian J. Theis. An automatic method for robust and fast cell detection in bright field images from high-throughput microscopy. *BMC Bioinformatics*, 14(1):297–297, October 2013. doi: 10.1186/1471-2105-14-297. MAG ID: 2072561351 S2ID: 465d34e7d10b37e9a91cd4b0cc6969042e9c83af.

[47] Pedro F. Felzenszwalb, David A. McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. URL https://api.semanticscholar.org/CorpusID:14327585.

[48] Pedro F. Felzenszwalb, Ross B. Girshick, and David A. McAllester. Cascade object detection with deformable part models. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2241–2248, 2010. URL https://api.semanticscholar.org/CorpusID:6735187.

[49] Pedro F. Felzenszwalb, Ross B. Girshick, David A. McAllester, and Deva Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:1627–1645, 2010. URL https://api.semanticscholar.org/CorpusID:3198903.

[50] Fitri N. Rahayu, Fitri N. Rahayu, Ulrich Reiter, Ulrich Reiter, Touradj Ebrahimi, Touradj Ebrahimi, Andrew Perkis, Andrew Perkis, U. Peter Svensson, and U. Peter Svensson. SS-SSIM and MS-SSIM for digital cinema applications. *electronic imaging*, 7240(1):72400, February 2009. doi: 10.1117/12.805805. MAG ID: 2043031986 S2ID: fcf8ae342bbfb9fab1721992ba33769fa4507939.

[51] Flavio Schneider. ArchiSound: Audio Generation with Diffusion. *arXiv.org*, January 2023. doi: 10.48550/arxiv.2301.13267. ARXIV_ID: 2301.13267 MAG ID: 4318902838 S2ID: 627f56001dd37d7b805bfeb87ffafc6985aeb940.

[52] Gabriel della Maggiora, L. A. Croquevielle, Nikita Desphande, Harry Horsley, Thomas Heinis, and Artur Yakimovich. Conditional Variational Diffusion Models. *arXiv.org*, 2023. doi: 10.48550/arxiv.2312.02246. ARXIV_ID: 2312.02246 S2ID: 0aa948324bca48784d2a78a799c92b101687c89e.

[53] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YOLOX: Exceeding YOLO Series in 2021, 2021. URL https://arxiv.org/abs/2107.08430. Version Number: 2.

[54] Geert Litjens, Geert Litjens, Thijs Kooi, Thijs Kooi, Babak Ehteshami Bejnordi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Francesco Ciompi, Mohsen Ghafoorian, Mohsen Ghafoorian, Jeroen van der Laak, Jeroen van der Laak, Bram van Ginneken, Bram van Ginneken, Clara I. Sánchez, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, December 2017. doi: 10.1016/j.media.2017.07.005. MAG ID: 2592929672.

[55] George Juraj Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, JohnMark Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Neural Information Processing Systems*, June 2023. doi: 10.48550/arxiv.2306.04675. ARXIV_ID: 2306.04675 MAG ID: 4380135944 S2ID: caf6ca4ccfabf903003cdf927fb7e883342fdcad.

[56] Georgios Kaissis, Georgios Kaissis, Marcus R. Makowski, Marcus R. Makowski, Marcus R. Makowski, Daniel Rueckert, Daniel Rueckert, Daniel Rückert, Rickmer Braren, and Rickmer Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6):305–311, June 2020. doi: 10.1038/s42256-020-0186-1. MAG ID: 3033511014 S2ID: 5123717445799d8137327f4041e8d5a5a2c91379.

[57] R J Geraghty, A Capes-Davis, J M Davis, J Downward, R I Freshney, I Knezevic, R Lovell-Badge, J R W Masters, J Meredith, G N Stacey, P Thraves, and M Vias. Guidelines for the use of cell lines in biomedical research. *British Journal of Cancer*, 111(6):1021–1046, September 2014. ISSN 0007-0920, 1532-1827. doi: 10.1038/bjc.2014.166. URL https://www.nature.com/articles/bjc2014166.

[58] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:142–158, 2016. URL https://api.semanticscholar.org/CorpusID:13980455.

[59] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. 2014. doi: 10.48550/ARXIV.1406.2661. URL https://arxiv.org/abs/1406.2661. Publisher: arXiv Version Number: 1.

[60] Gregory P. Way, Heba Sailem, Steven Shave, Richard Kasprowciz, and Neil O. Carragher. Evolution and Impact of High Content Imaging. September 2023. doi: 10.1016/j.slasd.2023.08.009. MAG ID: 4386401812.

[61] Gregory Pappas, Gregory Pappas, Adnan A. Hyder, and Adnan A. Hyder. Exploring ethical considerations for the use of biological and physiological markers in population-based surveys in less developed countries. *Globalization and Health*, 1 (1):16–16, November 2005. doi: 10.1186/1744-8603-1-16. MAG ID: 2137103412.

[62] Chenxu Guo, Francis K. Fordjour, Shang Jui Tsai, James C. Morrell, and Stephen J. Gould. Choice of selectable marker affects recombinant protein expression in cells and exosomes. *Journal of Biological Chemistry*, 297(1):100838, July 2021. ISSN 00219258. doi: 10.1016/j.jbc.2021.100838. URL https://linkinghub.elsevier.com/retrieve/pii/S0021925821006360.

[63] Gyuhyun Lee, Gyuhyun Lee, Jeong-Woo Oh, Jeong-Woo Oh, Mi-Sun Kang, Mi-Sun Kang, Nam-Gu Her, Nam-Gu Her, Myoung-Hee Kim, Myoung-Hee Kim, Won-Ki

Jeong, and Won-Ki Jeong. DeepHCS: Bright-Field to Fluorescence Microscopy Image Conversion Using Deep Learning for Label-Free High-Content Screening. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 335–343, September 2018. doi: 10.1007/978-3-030-00934-2_38. MAG ID: 2892296896 S2ID: 6221dc0f12e1513610a3a4de8b1fc4adff628825.

[64] Gyuhyun Lee, Gyuhyun Lee, Jeong-Woo Oh, Jeong-Woo Oh, Jeong-Woo Oh, Nam-Gu Her, Nam-Gu Her, Won-Ki Jeong, and Won-Ki Jeong. DeepHCS++: Bright-field to fluorescence microscopy image conversion using multi-task learning with adversarial losses for label-free high-content screening. *Medical Image Analysis*, 70: 101995, 2021. doi: 10.1016/j.media.2021.101995. MAG ID: 3133306532 S2ID: 6ba26fb5c73049bf60d365a19be4cc209c4f917e.

[65] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37:1904–1916, 2014. URL https://api.semanticscholar.org/CorpusID:436933.

[66] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking ImageNet Pre-training, 2018. URL https://arxiv.org/abs/1811.08883. Version Number: 1.

[67] Helen Pearson and Helen Pearson. Image manipulation: CSI: cell biology. *Nature*, 434(7036):952–953, April 2005. doi: 10.1038/434952a. MAG ID: 2087961582 S2ID: a449b680632b31ed046387164822ab80b689ac0e.

[68] Helen Pearson and Helen Pearson. The good, the bad and the ugly. *Nature*, 447 (7141):138–140, May 2007. doi: 10.1038/447138a. MAG ID: 2050115286 S2ID: c4de4c0f026a4bc79f6bb323a4db7b8bac943cb0.

[69] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models, December 2020. URL http://arxiv.org/abs/2006.11239. arXiv:2006.11239 [cs, stat].

[70] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen Video: High Definition Video Generation with Diffusion Models, October 2022. URL http://arxiv.org/abs/2210.02303. arXiv:2210.02303 [cs].

[71] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video Diffusion Models. 2022. doi: 10.48550/ARXIV.2204.03458. URL https://arxiv.org/abs/2204.03458. Publisher: arXiv Version Number: 2.

[72] Huixia Ren, Huixia Ren, Huixia Ren, Huixia Ren, Mengdi Zhao, Mengdi Zhao, Bo Liu, Bo Liu, Bo Liu, Bo Liu, Ruixiao Yao, Ruixiao Yao, Ruixiao Yao, Ruixiao Yao, Lei Qi, Qi Liu, Qi Liu, Zhipeng Ren, Zhipeng Ren, Zirui Wu, Zirui Wu, Zongmao Gao, Zongmao Gao, Xiu Yang, Xiaojing Yang, Chao Tang, and Chao Tang. Cellbow: a robust customizable cell segmentation program. 8(3):245–255, 2020. doi: 10.1007/s40484-020-0213-6. MAG ID: 3083174864.

[73] I Galende and Octavio Miguel Rivero-Lezcano. Ethical considerations about the collection of biological samples for genetic analysis in clinical trials. *Research Ethics*, 19(2):220–226, January 2023. doi: 10.1177/17470161231152077. MAG ID: 4319788163 S2ID: afcea223cd5d53938877055d1befd4126f0c19b6.

[74] Isabella Castiglioni, Isabella Castiglioni, Leonardo Rundo, Leonardo Rundo, Marina Codari, Marina Codari, Giovanni Di Leo, Giovanni Di Leo, Christian Salvatore, Christian Salvatore, Matteo Interlenghi, Matteo Interlenghi, Francesca Gallivanone, Francesca Gallivanone, Andrea Cozzi, Andrea Cozzi, Natascha Claudia D'Amico, Natascha Claudia D'Amico, Francesco Sardanelli, and Francesco Sardanelli. AI applications to medical images: From machine learning to deep learning. *Physica Medica*, 83:9–24, 2021. doi: 10.1016/j.ejmp.2021.02.006. MAG ID: 3135096391.

[75] Ivana Marin, Ivana Marin, S. Gotovac, Sven Gotovac, Mladen Russo, and Mladen Russo. Evaluation of Generative Adversarial Network for Human Face Image Synthesis. *International Conference on Software, Telecommunications and Computer Networks*, pages 1–6, 2020. doi: 10.23919/softcom50211.2020.9238203. MAG ID: 3095975757 S2ID: 697c3cdf8575f9a69cc1869050fa04926bf65842.

[76] Jan Hosang, Jan Hosang, Rodrigo Benenson, Rodrigo Benenson, Bernt Schiele, and Bernt Schiele. Learning non-maximum suppression. *arXiv: Computer Vision and Pattern Recognition*, May 2017. doi: 10.1109/cvpr.2017.685. ARXIV_ID: 1705.02950 MAG ID: 2950806363 S2ID: f94feceb5b725c6b303b758a0e5e90215b0174d3.

[77] Jan Oscar Cross-Zamirski, Jan Oscar Cross-Zamirski, Elizabeth Mouchet, Elizabeth Mouchet, Guy Williams, Guy Williams, Carola-Bibiane Schönlieb, Carola-Bibiane Schönlieb, Riku Turkki, Riku Turkki, Yinhai Wang, and Yinhai Wang. Label-free prediction of cell painting from brightfield images. *Scientific Reports*. doi: 10.1038/s41598-022-12914-x. MAG ID: 4282967624 S2ID: 590fc45e167068cac9a8a98eee95725531a5a352.

[78] Jascha Sohl-Dickstein, Jascha Sohl-Dickstein, Eric A. Weiss, Eric L. Weiss, Niru Maheswaranathan, Niru Maheswaranathan, Surya Ganguli, and Surya Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. *arXiv: Learning*, March 2015. ARXIV_ID: 1503.03585 MAG ID: 2129069237 S2ID: 2dcef55a07f8607a819c21fe84131ea269cc2e3c.

[79] Jason R. Swedlow, Jason R. Swedlow, Ilya G. Goldberg, Ilya G. Goldberg, Erik Brauner, Erik Brauner, Peter K. Sorger, and Peter K. Sorger. Informatics and Quantitative Analysis in Biological Imaging. *Science*, 300(5616):100–102, April 2003. doi: 10.1126/science.1082602. MAG ID: 2147385427 S2ID: ae4c01a3463feb9d2af915e20f510d2749956f88.

[80] Jean-Baptiste Lugagne, Jean-Baptiste Lugagne, Srajan Jain, Srajan Jain, Pierre Ivanovitch, Pierre Ivanovitch, Zacchari Ben Meriem, Zacchary Ben Meriem, Clément Vulin, Clément Vulin, Chiara Fracassi, Chiara Fracassi, Grégory Batt, Gregory Batt, Pascal Hersen, and Pascal Hersen. Identification of individual cells from z-stacks of bright-field microscopy images. *Scientific Reports*, 8(1):11455–11455, July 2018. doi: 10.1038/s41598-018-29647-5. MAG ID: 2883693400 S2ID: ab274d2a2e7ea58fbbde709fd375878bb017e7a1.

[81] Jean-Baptiste Lugagne, Jean-Baptiste Lugagne, Haonan Lin, Haonan Lin, Mary J. Dunlop, and Mary J. Dunlop. DeLTA: Automated cell segmentation, tracking, and lineage reconstruction using deep learning. *PLOS Computational Biology*, 16(4): 1007673, April 2020. doi: 10.1371/journal.pcbi.1007673. MAG ID: 3015600294.

[82] Jiajie Fan, L. Vuaille, Thomas Bäck, and Hao Wang. On the Noise Scheduling for Generating Plausible Designs with Diffusion Models. *arXiv.org*, 2023. doi: 10.48550/arxiv.2311.11207. ARXIV_ID: 2311.11207 S2ID: 7ff5027b5cad8d4128259a1eb348d60aabece852.

[83] Jianming Zhang, Jianming Zhang, Benben Huang, Benben Huang, Zi Ye, Zi Ye, Li-Dan Kuang, Li-Dan Kuang, Xin Ning, and Xin Ning. Siamese anchor-free ob-

ject tracking with multiscale spatial attentions. *Scientific Reports*, 11(1):22908, November 2021. doi: 10.1038/s41598-021-02095-4. MAG ID: 3215098936 S2ID: 5bc4139440539691099855 4d83ec22cf3ebf0fab.

[84] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLOv8, 2023. URL https://github.com/ultralytics/ultralytics. orcid: 0000-0001-5950-6979, 0000-0002-7603-6750, 0000-0003-3783-7069 license: AGPL-3.0.

[85] Johannes Gilg, Torben Teepe, Fabian Herzog, Philipp Wolters, and G. Rigoll. Do We Still Need Non-Maximum Suppression? Accurate Confidence Estimates and Implicit Duplication Modeling with IoU-Aware Calibration. *IEEE Workshop/Winter Conference on Applications of Computer Vision*, 2023. doi: 10.1109/wacv57701.2024.00478. ARXIV_ID: 2309.03110 S2ID: 8913e4970e5d4152e7a7acf76e6220a8118b0071.

[86] Juan C Caicedo, Juan C. Caicedo, Jonathan Roth, Jonathan R. Roth, Allen Goodman, Allen Goodman, Tim Becker, Tim Becker, Tim Becker, Kyle W. Karhohs, Kyle W. Karhohs, Matthieu Broisin, Matthieu Broisin, Csaba Molnár, Csaba Molnar, Claire McQuin, Claire McQuin, Shantanu Singh, Shantanu Singh, Fabian J. Theis, Fabian J. Theis, Anne E. Carpenter, and Anne E. Carpenter. Evaluation of Deep Learning Strategies for Nucleus Segmentation in Fluorescence Images. *Cytometry Part A*, 95(9):952–965, July 2019. doi: 10.1002/cyto.a.23863. MAG ID: 2961912654.

[87] Jyrki Selinummi, Jyrki Selinummi, Pekka Ruusuvuori, Pekka Ruusuvuori, Irina Podolsky, Irina Podolsky, Irina Podolsky, Adrian Ozinsky, Adrian Ozinsky, Elizabeth S. Gold, Elizabeth S. Gold, Olli Yli-Harja, Olli Yli-Harja, Alan Aderem, Alan Aderem, Ilya Shmulevich, and Ilya Shmulevich. Bright Field Microscopy as an Alternative to Whole Cell Fluorescence in Automated Analysis of Macrophage Images. *PLOS ONE*, 4(10), October 2009. doi: 10.1371/journal.pone.0007497. MAG ID: 1980189947 S2ID: 26af3e90694e8ed14e25a11f0492af6b2e39ca87.

[88] Kaisa Liimatainen, Kaisa Liimatainen, Lauri Kananen, Lauri Kananen, Leena Latonen, Leena Latonen, Leena Latonen, Pekka Ruusuvuori, and Pekka Ruusuvuori. Iterative unsupervised domain adaptation for generalized cell detection from brightfield z-stacks. *BMC Bioinformatics*, 20(1):1–10, February 2019. doi: 10.1186/s12859-019-2605-z. MAG ID: 2920164639 S2ID: 0a8529b7ec564970e9fbd1da639c30571b30eb1a.

[89] Jee Yon Kim, Yeon-Gu Kim, and Gyun Min Lee. CHO cells in biotechnology for production of recombinant proteins: current state and further potential. *Applied Microbiology and Biotechnology*, 93(3):917–930, February 2012. ISSN 0175-7598, 1432-0614. doi: 10.1007/s00253-011-3758-5. URL http://link.springer.com/10.1007/s00253-011-3758-5.

[90] Tae Kyung Kim and James H. Eberwine. Mammalian cell transfection: the present and the future. *Analytical and Bioanalytical Chemistry*, 397(8):3173–3178, August 2010. ISSN 1618-2642, 1618-2650. doi: 10.1007/s00216-010-3821-6. URL http://link.springer.com/10.1007/s00216-010-3821-6.

[91] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do Better ImageNet Models Transfer Better?, 2018. URL https://arxiv.org/abs/1805.08974. Version Number: 3.

[92] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[93] Hei Law and Jia Deng. CornerNet: Detecting objects as paired keypoints. *International Journal of Computer Vision*, 128:642 – 656, 2018. URL https://api.semanticscholar.org/CorpusID:51923817.

[94] A. Lehmussola, P. Ruusuvuori, J. Selinummi, T. Rajala, and O. Yli-Harja. Synthetic Images of High-Throughput Microscopy for Validation of Image Analysis Methods. *Proceedings of the IEEE*, 96(8):1348–1360, August 2008. ISSN 0018-9219, 1558-2256. doi: 10.1109/JPROC.2008.925490. URL http://ieeexplore.ieee.org/document/4567415/.

[95] Leonor Morgado, Estibaliz G'omez-de-Mariscal, H. Heil, and Ricardo Henriques. The rise of data-driven microscopy powered by machine learning. *Journal of Microscopy*, 2024. doi: 10.1111/jmi.13282. ARXIV_ID: 2401.05282 S2ID: 6d2db990a34439c4ca806563bedca4f25b4bb2f2.

[96] Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault

Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.emnlp-demo.21.

[97] Rui Li, Gabriel della Maggiora, Vardan Andriasyan, Anthony Petkidis, Artsemi Yushkevich, Mikhail Kudryashev, and Artur Yakimovich. Microscopy image reconstruction with physics-informed denoising diffusion probabilistic model, 2023. URL https://arxiv.org/abs/2306.02929. Version Number: 2.

[98] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection, 2020. URL https://arxiv.org/abs/2006.04388. Version Number: 1.

[99] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2016. URL https://api.semanticscholar.org/CorpusID:10716717.

[100] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. 2017. doi: 10.48550/ARXIV.1708.02002. URL https://arxiv.org/abs/1708.02002. Publisher: arXiv Version Number: 2.

[101] Jihong Liu, Fei Gao, Lvheng Zhang, and Haixu Yang. A Saturation Artifacts Inpainting Method Based on Two-Stage GAN for Fluorescence Microscope Images, June 2024. URL https://www.preprints.org/manuscript/202406.0267/v1.

[102] W. Liu, Dragomir Anguelov, D. Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single shot MultiBox detector. In *European conference on computer vision*, 2015. URL https://api.semanticscholar.org/CorpusID:2141740.

[103] Ye Liu, Sophia J. Wagner, and Tingying Peng. Multi-Modality Microscopy Image Style Transfer for Nuclei Segmentation, 2021. URL https://arxiv.org/abs/2111.12138. Version Number: 1.

[104] Feixiao Long, Feixiao Long, and Feixiao Long. Microscopy cell nuclei segmentation with enhanced U-Net. *BMC Bioinformatics*, 21(1):8, January 2020. doi: 10.1186/s12859-019-3332-1. MAG ID: 2999770235 S2ID: 75bdaacf355d0a6d52d9a4a8790e1fdc936eb867.

[105] Chixiang Lu, Kai Chen, Heng Qiu, Xiaojun Chen, Gu Chen, Xiaojuan Qi, and Haibo Jiang. Diffusion-based deep learning method for augmenting ultrastructural imaging and volume electron microscopy. *Nature Communications*, 15(1):4677, June 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-49125-z. URL https://www.nature.com/articles/s41467-024-49125-z.

[106] Wenyu Lv, Shangliang Xu, Yian Zhao, Guanzhong Wang, Jinman Wei, Cheng Cui, Yuning Du, Qingqing Dang, and Yi Liu. DETRs beat yolos on real-time object detection. *ArXiv*, abs/2304.08069, 2023. URL https://api.semanticscholar.org/CorpusID:258179840.

[107] Ricardo Macarron, Martyn N. Banks, Dejan Bojanic, David J. Burns, Dragan A. Cirovic, Tina Garyantes, Darren V. S. Green, Robert P. Hertzberg, William P. Janzen, Jeff W. Paslay, Ulrich Schopfer, and G. Sitta Sittampalam. Impact of high-throughput screening in biomedical research. *Nature Reviews Drug Discovery*, 10(3):188–195, March 2011. ISSN 1474-1776, 1474-1784. doi: 10.1038/nrd3368. URL https://www.nature.com/articles/nrd3368.

[108] Madeleine S. Durkee, Madeleine S. Durkee, Rebecca Abraham, Rebecca Abraham, Marcus R. Clark, Marcus R. Clark, Maryellen L. Giger, and Maryellen L. Giger. Artificial Intelligence and Cellular Segmentation in Tissue Microscopy Images. *American Journal of Pathology*, 191(10):1693–1701, June 2021. doi: 10.1016/j.ajpath.2021.05.022. MAG ID: 3170493236 S2ID: cc45b38fb5ced246f872c8ae5ec6ed11ef209ba9.

[109] Malsha V. Perera and Vishal M. Patel. Analyzing Bias in Diffusion-based Face Generation Models. *2023 IEEE International Joint Conference on Biometrics (IJCB)*, 2023. doi: 10.1109/ijcb57857.2023.10449200. ARXIV_ID: 2305.06402 S2ID: 94831cbd104369092b08f3711e6ac95c5f5f2c7b.

[110] Maria Littmann, Maria Littmann, Katharina Selig, Katharina Selig, Liel Cohen-Lavi, Liel Cohen-Lavi, Yotam Frank, Yotam Frank, Peter Hönigschmid, Peter Hönigschmid, Evans Kataka, Evans Kataka, Anja Mösch, Anja Mösch, Kun Qian, Kun Qian, Kun Qian, Avihai Ron, Avihai Ron, Sebastian Schmid, Sebastian

Schmid, Adam Sorbie, Adam Sorbie, Liran Szlak, Liran Szlak, Ayana Dagan-Wiener, Ayana Dagan-Wiener, Nir Ben-Tal, Nir Ben-Tal, Peter Gmeiner, Masha Y. Niv, Masha Y. Niv, Daniel Razansky, Daniel Razansky, Björn W. Schuller, Björn Schuller, Donna P. Ankerst, Donna P. Ankerst, Tomer Hertz, Tomer Hertz, Burkhard Rost, and Burkhard Rost. Validity of machine learning in biology and medicine increased through collaborations across fields of expertise. *Nature Machine Intelligence*, 2(1):18–24, January 2020. doi: 10.1038/s42256-019-0139-8. MAG ID: 2999864839 S2ID: 653a1af33e753b22730b74f09e7a4e52b3044a65.

[111] Marin Scalbert, Marin Scalbert, Florent Couzinié-Devy, Florent Couzinie-Devy, Riadh Fezzani, and Riadh Fezzani. Generic Isolated Cell Image Generator. *Cytometry Part A*, 95(11), November 2019. doi: 10.1002/cyto.a.23899. MAG ID: 2979953123.

[112] Mark Everingham, Mark Everingham, Luc Van Gool, Luc Van Gool, Christopher Williams, Christopher K. I. Williams, Christopher K. I. Williams, Christopher Williams, John Winn, John Winn, Andrew Zisserman, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. doi: 10.1007/s11263-009-0275-4. MAG ID: 2031489346.

[113] Mark Everingham, Mark Everingham, S. M. Ali Eslami, S. M. Eslami, Luc Van Gool, Luc Van Gool, Christopher K. I. Williams, Christopher K. I. Williams, Christopher K. I. Williams, Christopher Williams, John Winn, John Winn, Andrew Zisserman, and Andrew Zisserman. The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision*, 111(1):98–136, January 2015. doi: 10.1007/s11263-014-0733-5. MAG ID: 2037227137 S2ID: 616b246e332573af1f4859aa91440280774c183a.

[114] Erik Meijering. Cell Segmentation: 50 Years Down the Road [Life Sciences]. *IEEE Signal Processing Magazine*, 29(5):140–145, September 2012. ISSN 1053-5888. doi: 10.1109/MSP.2012.2204190. URL http://ieeexplore.ieee.org/document/6279591/.

[115] Meysam Cheramin, Meysam Cheramin, Meysam Cheramin, Jianqiang Cheng, Jianqiang Cheng, Ruiwei Jiang, Ruiwei Jiang, Kai Pan, and Kai Pan. Computationally Efficient Approximations for Distributionally Robust Optimization Under Moment and Wasserstein Ambiguity. *INFORMS journal on computing*,

January 2022. doi: 10.1287/ijoc.2021.1123. MAG ID: 4210665097 S2ID: 159ee687cf79ffa2ec6a6ca3875c49bff8c85b7e.

[116] Midjourney. Midjourney, 2022. URL https://www.midjourney.com/home/. last accessed: 2024-09-10.

[117] Yisroel Mirsky, Wenke Lee, and Wenke Lee. The Creation and Detection of Deepfakes: A Survey. *arXiv: Computer Vision and Pattern Recognition*, April 2020. doi: 10.1145/3425780. ARXIV_ID: 2004.11138 MAG ID: 3019200173 S2ID: 92b4c8deecee703569b9e909dfb88aa70e691219.

[118] Mohammed Ali, Mohammed Ali, O. N. Misko, Oleg Misko, Sten-Oliver Salumaa, Sten-Oliver Salumaa, Mikhail Papkov, Mikhail Papkov, Kaupo Palo, Kaupo Palo, Dmytro Fishman, Dmytro Fishman, Leopold Parts, and Leopold Parts. Evaluating Very Deep Convolutional Neural Networks for Nucleus Segmentation from Brightfield Cell Microscopy Images. *SLAS discovery : advancing life sciences R & D*, 26 (9):1125–1137, June 2021. doi: 10.1177/24725552211023214.

[119] Mojca Mattiazzi Ušaj, Mojca Mattiazzi Usaj, Erin B. Styles, Erin B. Styles, Adrian J. Verster, Adrian J. Verster, Helena Friesen, Helena Friesen, Charles Boone, Charles Boone, Brenda Andrews, and Brenda J. Andrews. High-Content Screening for Quantitative Cell Biology. *Trends in Cell Biology*, 26(8):598–611, August 2016. doi: 10.1016/j.tcb.2016.03.008. MAG ID: 2340102874 S2ID: 6bfc1c0c7d3fc3cd9ebb9517cc0de73e841b7f74.

[120] Navaneeth Bodla, Navaneeth Bodla, Bharat Singh, Bharat Singh, Rama Chellappa, Rama Chellappa, Larry S. Davis, Larry S. Davis, and Larry S. Davis. Soft-NMS — Improving Object Detection with One Line of Code. *IEEE International Conference on Computer Vision*, pages 5562–5570, October 2017. doi: 10.1109/iccv.2017.593. ARXIV_ID: 1704.04503 MAG ID: 2964121718 S2ID: 53c0aa8d33d240197caff824a6225fb223c1181c.

[121] Alex Nichol and Prafulla Dhariwal. Improved Denoising Diffusion Probabilistic Models. 2021. doi: 10.48550/ARXIV.2102.09672. URL https://arxiv.org/abs/2102.09672. Publisher: arXiv Version Number: 1.

[122] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. 2021. doi:

10.48550/ARXIV.2112.10741. URL https://arxiv.org/abs/2112.10741. Publisher: arXiv Version Number: 3.

[123] Nicki Skafte Detlefsen, Jiri Borovec, Justus Schock, Ananya Harsh, Teddy Koker, Luca Di Liello, Daniel Stancl, Changsheng Quan, Maxim Grechkin, and William Falcon. TorchMetrics - Measuring Reproducibility in PyTorch, February 2022. URL https://github.com/Lightning-AI/torchmetrics.

[124] Olga Russakovsky, Olga Russakovsky, Jia Deng, Jia Deng, Hao Su, Hao Su, Jonathan Krause, Jonathan Krause, Sanjeev Satheesh, Sanjeev Satheesh, Siying Ma, Sean Ma, Zhiheng Huang, Zhiheng Huang, Andrej Karpathy, Andrej Karpathy, Aditya Khosla, Aditya Khosla, Michael S. Bernstein, Michael S. Bernstein, Alexander C. Berg, Alexander C. Berg, Feifei Li, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, December 2015. doi: 10.1007/s11263-015-0816-y. MAG ID: 2117539524.

[125] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, T. Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A Space-Time Diffusion Model for Video Generation. *arXiv.org*, 2024. doi: 10.48550/arxiv.2401.12945. ARXIV_ID: 2401.12945 S2ID: 94f7d8bce3bb848d127c8f113afc5bb0243579df.

[126] OpenAI. Dall-e 2, 2022. URL https://openai.com/index/dall-e-2. last accessed: 2024-09-10.

[127] OpenAI. Dall-e 3, 2023. URL https://openai.com/index/dall-e-3. last accessed: 2024-09-10.

[128] Patrick Esser, Sumith Kulal, A. Blattmann, Rahim Entezari, Jonas Muller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. *arXiv.org*, 2024. doi: 10.48550/arxiv.2403.03206. ARXIV_ID: 2403.03206 S2ID: 41a66997ce0a366bba3becf7c3f37c9aebb13fbd.

[129] Pavel Kopel, Tomas Vicar, Jan Balvan, Jan Balvan, Josef Jaroš, Josef Jaroš, Florian Jug, Florian Jug, Radim Kolář, Radim Kolar, Michal Masařík, Michal Masařík, Michal Masarik, Jaromír Gumulec, and Jaromír Gumulec. Cell segmentation methods for label-free contrast microscopy: review and comprehensive comparison. *BMC*

Bioinformatics, 20(1):360–360, June 2019. doi: 10.1186/s12859-019-2880-8. MAG ID: 2953421783 S2ID: 01dd8c73df2b37b2e02a9e25612ed86a06f25cc5.

[130] Pedram Mohammadi, A. Ebrahimi-Moghadam, and S. Shirani. Subjective and Objective Quality Assessment of Image: A Survey. *arXiv.org*, 2014. ARXIV_ID: 1406.7799 MAG ID: 2952473915 S2ID: 9edc6b3745a45d1bb84b39d74d88c82e0e1c24b0.

[131] Philipp Angerer, Philipp Angerer, Lukas M. Simon, Lukas M. Simon, Lukas M. Simon, Sophie Tritschler, Sophie Tritschler, Florian Wolf, F. Alexander Wolf, David Fischer, David S. Fischer, Fabian J. Theis, and Fabian J. Theis. Single cells make big data: New challenges and opportunities in transcriptomics. *Current Opinion in Systems Biology*, 4:85–91, August 2017. doi: 10.1016/j.coisb.2017.07.004. MAG ID: 2735067671 S2ID: 5c8a622955c7da0dc68919d9b331ffb8f65d5500.

[132] Henry Pinkard, Zachary Phillips, Arman Babakhani, Daniel A. Fletcher, and Laura Waller. Deep learning for single-shot autofocus microscopy. *Optica*, 6(6):794, June 2019. ISSN 2334-2536. doi: 10.1364/OPTICA.6.000794. URL https://opg.optica.org/abstract.cfm?URI=optica-6-6-794.

[133] Piotr Dollár, Piotr Dollar, Christian Wojek, Christian Wojek, Bernt Schiele, Bernt Schiele, Pietro Perona, and Pietro Perona. Pedestrian detection: A benchmark. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 304–311, June 2009. doi: 10.1109/cvpr.2009.5206631. MAG ID: 2107775979 S2ID: 3e083dc8aeb7983a5cdff146985363d38caf0886.

[134] Piotr Dollár, Piotr Dollar, Christian Wojek, Christian Wojek, Bernt Schiele, Bernt Schiele, Pietro Perona, and Pietro Perona. Pedestrian Detection: An Evaluation of the State of the Art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):743–761, April 2012. doi: 10.1109/tpami.2011.155. MAG ID: 2031454541 S2ID: 34e0ba2daabfa4d3d22913ade8265aff50b5f917.

[135] Rafał Mantiuk, Rafat K. Mantiuk, Rafal Mantiuk, Kil Joong Kim, Kil Joong Kim, Allan G. Rempel, Allan G. Rempel, Wolfgang Heidrich, and Wolfgang Heidrich. HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on Graphics*, 30(4): 40, July 2011. doi: 10.1145/2010324.1964935. MAG ID: 2156566307 S2ID: b69d13c6a1b848c7f4816a73fbdb1d396a03928c.

[136] Arun Rai. Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science*, 48(1):137–141, January 2020. ISSN 0092-0703, 1552-7824. doi: 10.1007/s11747-019-00710-5. URL http://link.springer.com/10.1007/s11747-019-00710-5.

[137] Rainer Pepperkok, Rainer Pepperkok, Jan Ellenberg, and Jan Ellenberg. High-throughput fluorescence microscopy for systems biology. *Nature Reviews Molecular Cell Biology*, 7(9):690–696, July 2006. doi: 10.1038/nrm1979. MAG ID: 2158706178 S2ID: 1f57e8d422456da7f47cc12047af1eb82099a7ae.

[138] Satwik Rajaram, Benjamin Pavie, Nicholas E F Hac, Steven J Altschuler, and Lani F Wu. SimuCell: a flexible framework for creating synthetic microscopy images. *Nature Methods*, 9(7):634–635, July 2012. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.2096. URL https://www.nature.com/articles/nmeth.2096.

[139] Rasmus Rothe, Rasmus Rothe, Matthieu Guillaumin, Matthieu Guillaumin, Luc J. Van Gool, and Luc Van Gool. Non-maximum Suppression for Object Detection by Passing Messages Between Windows. *Asian Conference on Computer Vision*, 9003:290–306, November 2014. doi: 10.1007/978-3-319-16865-4_19. MAG ID: 2126096326 S2ID: d5851ecc786b643ed203fce98276bf67270e3add.

[140] Suman V. Ravuri and Oriol Vinyals. Seeing is not necessarily believing: Limitations of biggans for data augmentation. 2019. URL https://api.semanticscholar.org/CorpusID:195527523.

[141] Joseph Redmon and Ali Farhadi. YOLO9000: Better, faster, stronger. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6525, 2016. URL https://api.semanticscholar.org/CorpusID:786357.

[142] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. *ArXiv*, abs/1804.02767, 2018. URL https://api.semanticscholar.org/CorpusID:4714433.

[143] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2015. URL https://api.semanticscholar.org/CorpusID:206594738.

[144] Ricard Durall, Ricard Durall, Avraam Chatzimichailidis, Avraam Chatzimichailidis, Peter Labus, Peter Labus, Peter Labus, Peter Labus, Janis Keuper, and Janis Keuper. Combating Mode Collapse in GAN training: An Empirical Analysis using Hessian Eigenvalues. *arXiv: Learning*, 2020. doi: 10.5220/0010167902110218. ARXIV_ID: 2012.09673 MAG ID: 3112865286 S2ID: 3ad13bd6713d74eec7faa964050f7d61089440a0.

[145] Riccardo A. G. Cinelli, Riccardo A. G. Cinelli, Aldo Ferrari, Aldo Ferrari, Vittorio Pellegrini, Vittorio Pellegrini, Mudit Tyagi, Mudit Tyagi, Mauro Giacca, Mauro Giacca, Fabio Beltram, and Fabio Beltram. The Enhanced Green Fluorescent Protein as a Tool for the Analysis of Protein Dynamics and Localization: Local Fluorescence Study at the Single-molecule Level. *Photochemistry and Photobiology*. doi: 10.1562/0031-8655(2000)071<0771:tegfpa>2.0.co;2. MAG ID: 4253752195.

[146] Roger Y. Tsien and Roger Y. Tsien. THE GREEN FLUORESCENT PROTEIN. *Annual Review of Biochemistry*, 67(1):509–544, January 1998. doi: 10.1146/annurev.biochem.67.1.509. MAG ID: 2127068865 S2ID: 4757163862c2a582af6a38a9f81a30732d4a5c54.

[147] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models, April 2022. URL http://arxiv.org/abs/2112.10752. arXiv:2112.10752 [cs].

[148] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yang Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xia Yin, and Zhao Zhang. Make-An-Audio: Text-To-Audio Generation with Prompt-Enhanced Diffusion Models. *International Conference on Machine Learning*, January 2023. doi: 10.48550/arxiv.2301.12661. ARXIV_ID: 2301.12661 MAG ID: 4318718996 S2ID: 6d1433f3342fbee85ad1e2809e62734aec5c3853.

[149] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv:1505.04597 [cs]*, May 2015. URL http://arxiv.org/abs/1505.04597. arXiv: 1505.04597.

[150] Ross Girshick and Ross Girshick. Fast R-CNN. pages 1440–1448, December 2015. doi: 10.1109/iccv.2015.169. MAG ID: 1536680647.

[151] Ross Girshick, Ross Girshick, Jeff Donahue, Jeff Donahue, Trevor Darrell, Trevor Darrell, Jitendra Malik, and Jitendra Malik. Rich Feature Hierarchies for Accurate

Object Detection and Semantic Segmentation. pages 580–587, June 2014. doi: 10.1109/cvpr.2014.81. MAG ID: 2102605133.

[152] S. Asha, G. Gopakumar, and Sai Subrahmanyam Gorthi. Saliency and ballness driven deep learning framework for cell segmentation in bright field microscopic images. *Engineering applications of artificial intelligence*, 118:105704–105704, February 2023. doi: 10.1016/j.engappai.2022.105704. MAG ID: 4311141309 S2ID: a1e954fee1777273e7f6765fc7c49c594a5cb15a.

[153] Saad Ullah Akram, Saad Ullah Akram, Juho Kannala, Juho Kannala, Lauri Eklund, Lauri Eklund, Janne Heikkilä, and Janne Heikkilä. Cell segmentation proposal network for microscopy image analysis. *LABELS/DLMIA@MICCAI*, pages 21–29, October 2016. doi: 10.1007/978-3-319-46976-8_3. MAG ID: 2522761331 S2ID: a2e06c347c192c94bcae153c36199d1272f7408f.

[154] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking FID: Towards a Better Evaluation Metric for Image Generation. *arXiv.org*, 2023. doi: 10.48550/arxiv.2401.09603. ARXIV_ID: 2401.09603 S2ID: ee9017b716c49006f423595624593b87df0a491b.

[155] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. 2022. doi: 10.48550/ARXIV.2205.11487. URL https://arxiv.org/abs/2205.11487. Publisher: arXiv Version Number: 1.

[156] Samuel A. Barnett and Samuel A. Barnett. Convergence Problems with Generative Adversarial Networks (GANs). *arXiv: Learning*, June 2018. ARXIV_ID: 1806.11382 MAG ID: 2810842121 S2ID: c30773c84a4be787e310e8c24900b6d42b4a826d.

[157] Sandhya Prabhakaran, Clarence Yapp, Gregory J. Baker, Johanna Beyer, Young Hwan Chang, Allison Creason, Robert F. Krueger, Jeremy Muhlich, Nathan Heath Patterson, Kevin Sidak, Damir Sudar, Andy Taylor, Luke Ternes, Jakob Troidl, Yi Xie, Artem Sokolov, and Darren R. Tyson. Addressing persistent challenges in digital image analysis of cancerous tissues. *bioRxiv*, July

2023. doi: 10.1101/2023.07.21.548450. MAG ID: 4385189573 S2ID: fb78a54f8d16ec82ae73de71b0b750d19a22a9a1.

[158] Johannes Schindelin, Ignacio Arganda-Carreras, Erwin Frise, Verena Kaynig, Mark Longair, Tobias Pietzsch, Stephan Preibisch, Curtis Rueden, Stephan Saalfeld, Benjamin Schmid, Jean-Yves Tinevez, Daniel James White, Volker Hartenstein, Kevin Eliceiri, Pavel Tomancak, and Albert Cardona. Fiji: an open-source platform for biological-image analysis. *Nature Methods*, 9(7):676–682, July 2012. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.2019. URL http://www.nature.com/articles/nmeth.2019.

[159] Shan E. Ahmed Raza, Shan E Ahmed Raza, Linda Cheung, Linda Cheung, Linda Cheung, Muhammad Shaban, Muhammad Shaban, Simon Graham, Simon Graham, D. B. A. Epstein, David B. A. Epstein, David Epstein, Stella Pelengaris, Stella Pelengaris, Michael Khan, Michael Khan, Nasir M. Rajpoot, and Nasir M. Rajpoot. Micro-Net: A unified model for segmentation of various objects in microscopy images. *Medical Image Analysis*, 52:160–173, February 2019. doi: 10.1016/j.media.2018.12.003. MAG ID: 2798643036.

[160] Shaofeng Zhang, Jinfa Huang, Qiang Zhou, Zhibin Wang, Fan Wang, Jiebo Luo, and Junchi Yan. Continuous-Multiple Image Outpainting in One-Step via Positional Query and A Diffusion-based Approach. *arXiv.org*, 2024. doi: 10.48550/arxiv.2401.15652. ARXIV_ID: 2401.15652 S2ID: d47277e0ae524d456b17c1d8246d558e3bf2a06d.

[161] Shouxin Ren, Shaoqing Ren, Kai He, Kaiming He, Ross Girshick, Ross Girshick, Jian Sun, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, June 2017. doi: 10.1109/tpami.2016.2577031. MAG ID: 639708223.

[162] Shuang Hao, Wenfeng Han, Tao Jiang, Yiping Li, Haonan Wu, Chunlin Zhong, Zhangjun Zhou, and He Tang. Synthetic Data in AI: Challenges, Applications, and Ethical Implications. *arXiv.org*, 2024. doi: 10.48550/arxiv.2401.01629. ARXIV_ID: 2401.01629 S2ID: 7a9933f52e4717d18c19bd360ed977dea6c74402.

[163] Samuel L. Smith, Pieter-Jan Kindermans, Chris Ying, and Quoc V. Le. Don't Decay the Learning Rate, Increase the Batch Size, 2017. URL https://arxiv.org/abs/1711.00489. Version Number: 2.

[164] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. 2020. doi: 10.48550/ARXIV.2010.02502. URL https://arxiv.org/abs/2010.02502. Publisher: arXiv Version Number: 4.

[165] Srijit Seal, Maria-Anna Trapotsi, O. Spjuth, Shantanu Singh, Jordi Carreras-Puigvert, Nigel Greene, Andreas Bender, and A. Carpenter. A Decade in a Systematic Review: The Evolution and Impact of Cell Painting. *bioRxiv*, 2024. doi: 10.1101/2024.05.04.592531. S2ID: 2c861d37ae678b8e37cc3a4064d50a9373f991ab.

[166] Stavroula Skylaki, Stavroula Skylaki, Oliver Hilsenbeck, Oliver Hilsenbeck, Timm Schroeder, and Timm Schroeder. Challenges in long-term imaging and quantification of single-cell dynamics. *Nature Biotechnology*, 34(11):1137–1144, November 2016. doi: 10.1038/nbt.3713. MAG ID: 2554348690 S2ID: 8e7490d155a1684db2c1fb1a9128b5a51f1c59bf.

[167] Susanne Bäck, Susanne Bäck, Amanda M. Dossat, Amanda M. Dossat, Amanda Dossat, Ilmari Parkkinen, Ilmari Parkkinen, Pyry Koivula, Pyry Koivula, Mikko Airavaara, Mikko Airavaara, Christopher T. Richie, Christopher T. Richie, Yun-Hsiang Chen, Yun-Hsiang Chen, Yun-Hsiang Chen, Yun-Hsiang Chen, Yun Wang, Yun Wang, Brandon K. Harvey, and Brandon K. Harvey. Neuronal Activation Stimulates Cytomegalovirus Promoter-Driven Transgene Expression. *Molecular therapy. Methods & clinical development*, 14:180–188, September 2019. doi: 10.1016/j.omtm.2019.06.006. MAG ID: 2953668380 S2ID: 875f002ea906f9eddd774cdf4229cdabc6a6c979.

[168] SYNENTEC GmbH. URL https://synentec.com/cellavista-4/. last accessed: 2024-09-10.

[169] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A Survey on Deep Transfer Learning. In Věra Kůrková, Yannis Manolopoulos, Barbara Hammer, Lazaros Iliadis, and Ilias Maglogiannis, editors, *Artificial Neural Networks and Machine Learning – ICANN 2018*, volume 11141, pages 270–279. Springer International Publishing, Cham, 2018. ISBN 978-3-030-01423-0 978-3-030-01424-7. doi: 10.1007/978-3-030-01424-7_27. URL http://link.springer.com/10.1007/978-3-030-01424-7_27. Series Title: Lecture Notes in Computer Science.

[170] Tasnima Sadekova, Tasnima Sadekova, Vladimir Gogoryan, Vladimir Gogoryan, Ivan Vovk, Ivan Vovk, Vadim Popov, B. B. Popov, Mikhail Kudinov, Mikhail

Kudinov, Jiansheng Wei, and Wei Jiang. A Unified System for Voice Cloning and Voice Conversion through Diffusion Probabilistic Modeling. *Interspeech*, September 2022. doi: 10.21437/interspeech.2022-10879. MAG ID: 4297841465 S2ID: 1dd0c640b537003ef1cbc15c01c80ab6d6b159d3.

[171] Terra M. Kuhn, M. Paulsen, and Sara Cuylen-Haering. Accessible high-speed image-activated cell sorting. *Trends in Cell Biology*, 2024. doi: 10.1016/j.tcb.2024.04.007. S2ID: 16c0c408aa42ca25b2012bedbbae143cb7211e25.

[172] Juan Terven and Diana Cordova-Esparza. A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS. *Machine Learning and Knowledge Extraction*, 5(4):1680–1716, November 2023. ISSN 2504-4990. doi: 10.3390/make5040083. URL http://arxiv.org/abs/2304.00501. arXiv:2304.00501 [cs].

[173] Thorsten Falk, Thorsten Falk, Dominic Mai, Dominic Mai, Robert Bensch, Robert Bensch, Özgün Çiçek, Özgün Çiçek, Ahmed Abdulkadir, Ahmed Abdulkadir, Yassine Marrakchi, Yassine Marrakchi, Anton Böhm, Anton Böhm, Jan Deubner, Jan Deubner, Zoë Jäckel, Zoe Jäckel, Katharina Seiwald, Katharina Seiwald, Alexander Dovzhenko, Alexander Dovzhenko, Olaf Tietz, Olaf Tietz, Cristina Dal Bosco, Cristina Dal Bosco, Seán Walsh, Sean Walsh, Deniz Saltukoglu, Deniz Saltukoglu, Tuan Leng Tay, Tuan Leng Tay, Marco Prinz, Marco Prinz, Klaus Palme, Klaus Palme, Klaus Palme, Matias Simons, Matias Simons, Ilka Diester, Ilka Diester, Thomas Brox, Thomas Brox, Olaf Ronneberger, and Olaf Ronneberger. U-Net: deep learning for cell counting, detection, and morphometry. *Nature Methods*, 16 (1):67–70, January 2019. doi: 10.1038/s41592-018-0261-2. MAG ID: 2900936384.

[174] Patrick Trampert, Dmitri Rubinstein, Faysal Boughorbel, Christian Schlinkmann, Maria Luschkova, Philipp Slusallek, Tim Dahmen, and Stefan Sandfeld. Deep Neural Networks for Analysis of Microscopy Images—Synthetic Data Generation and Adaptive Sampling. *Crystals*, 11(3):258, March 2021. ISSN 2073-4352. doi: 10.3390/cryst11030258. URL https://www.mdpi.com/2073-4352/11/3/258.

[175] Tsung-Yi Lin, Tsung-Yi Lin, Michael Maire, Michael Maire, Serge Belongie, Serge Belongie, James Hays, James Hays, Pietro Perona, Pietro Perona, Deva Ramanan, Deva Ramanan, Piotr Dollár, Piotr Dollar, C. Lawrence Zitnick, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. pages 740–755, September 2014. doi: 10.1007/978-3-319-10602-1_48. MAG ID: 1861492603.

[176] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, M. Kudinov, and Jiansheng Wei. Diffusion-Based Voice Conversion with Fast Maximum Likelihood Sampling Scheme. *International Conference on Learning Representations*, 2021. ARXIV_ID: 2109.13821 S2ID: d49a230b7718bd82fd7816d9d78e3ebd49118d2a.

[177] Abdul Vahab, Maruti S Naik, Prasanna G Raikar, and SR Prasad. Applications of object detection system. *International Research Journal of Engineering and Technology (IRJET)*, 6(4):4186–4192, 2019.

[178] Paul A. Viola and Michael J. Jones. Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, 1:I–I, 2001. URL https://api.semanticscholar.org/CorpusID:2715202.

[179] Paul A. Viola and Michael J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57:137–154, 2001. URL https://api.semanticscholar.org/CorpusID:2796017.

[180] Deepak Vohra. Apache Parquet. In *Practical Hadoop Ecosystem*, pages 325–335. Apress, Berkeley, CA, 2016. ISBN 978-1-4842-2198-3 978-1-4842-2199-0. doi: 10.1007/978-1-4842-2199-0_8. URL http://link.springer.com/10.1007/978-1-4842-2199-0_8.

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

[182] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. YOLOv10: Real-time end-to-end object detection. *ArXiv*, abs/2405.14458, 2024. URL https://api.semanticscholar.org/CorpusID:269983404.

[183] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7464–7475, 2022. URL https://api.semanticscholar.org/CorpusID:250311206.

[184] Chien-Yao Wang, I-Hau Yeh, and Hongpeng Liao. YOLOv9: Learning what you want to learn using programmable gradient information. *ArXiv*, abs/2402.13616, 2024. URL https://api.semanticscholar.org/CorpusID:267770251.

[185] Weidi Xie, Weidi Xie, J. Alison Noble, J. Alison Noble, Andrew Zisserman, and Andrew Zisserman. Microscopy cell counting and detection with fully convolutional regression networks. *Computer methods in biomechanics and biomedical engineering. Imaging & visualization*, 6(3):283–292, May 2018. doi: 10.1080/21681163.2016.1149104. MAG ID: 2347064614 S2ID: f83c5daa07093e4c27af17f8f39a66e6196dcb74.

[186] William S. Peebles and Saining Xie. Scalable Diffusion Models with Transformers. *IEEE International Conference on Computer Vision*, 2022. doi: 10.1109/iccv51070.2023.00387. ARXIV_ID: 2212.09748 S2ID: 736973165f98105fec3729b7db414ae4d80fcbeb.

[187] Florian M Wurm. Production of recombinant protein therapeutics in cultivated mammalian cells. *Nature Biotechnology*, 22(11):1393–1398, November 2004. ISSN 1087-0156, 1546-1696. doi: 10.1038/nbt1026. URL https://www.nature.com/articles/nbt1026.

[188] Fuyong Xing and Lin Yang. Robust Nucleus/Cell Detection and Segmentation in Digital Pathology and Microscopy Images: A Comprehensive Review. *IEEE Reviews in Biomedical Engineering*, 9:234–263, 2016. ISSN 1937-3333, 1941-1189. doi: 10.1109/RBME.2016.2515127. URL http://ieeexplore.ieee.org/document/7373572/.

[189] Xipeng Pan, Xipeng Pan, Xipeng Pan, Dengxian Yang, Dengxian Yang, Lingqiao Li, Lingqiao Li, Zhenbing Liu, Zhenbing Liu, Zhenbing Liu, Huihua Yang, Huihua Yang, Huihua Yang, Zhiwei Cao, Zhiwei Cao, Yubei He, Yubei He, Zhen Ma, Zhen Ma, Yiyi Chen, and Yiyi Chen. Cell detection in pathology and microscopy images with multi-scale fully convolutional neural networks. *World Wide Web*, 21(6):1721–1743, January 2018. doi: 10.1007/s11280-017-0520-7. MAG ID: 2784592326 S2ID: 3b533356f5e26d6bcba2a0e9526f7f6681e8740d.

[190] Yanan Song, Yanan Song, Yanan Song, Quan-Ke Pan, Quan-Ke Pan, Liang Gao, Liang Gao, Biao Zhang, and Biao Zhang. Improved non-maximum suppression for object detection using harmony search algorithm. *Applied Soft Computing*, 81:105478, August 2019. doi: 10.1016/j.asoc.2019.05.005. MAG ID: 2944770233 S2ID: 63aa796e91e78885c180b9cb3cd1e3185272f7cb.

[191] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? 2014. doi: 10.48550/ARXIV.1411.1792. URL https://arxiv.org/abs/1411.1792. Publisher: arXiv Version Number: 1.

[192] Youssef Kossale, Youssef Kossale, Mohammed Airaj, Mohammed Airaj, Aziz Darouichi, and Aziz Darouichi. Mode Collapse in Generative Adversarial Networks: An Overview. *OPTIMA*, October 2022. doi: 10.1109/icoa55659.2022.9934291. MAG ID: 4308733509 S2ID: 0befed4e1ddd6edafc011a4d7ce34b4b6a9a8927.

[193] Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, A. Kot, and Bihan Wen. SinSR: Diffusion-Based Image Super-Resolution in a Single Step. *arXiv.org*, 2023. doi: 10.48550/arxiv.2311.14760. ARXIV_ID: 2311.14760 S2ID: 4435f5cf602625862d81eb45454cd5f91d43afa0.

[194] Yusuf B. Johari, Yusuf B. Johari, Joseph M Scarrott, Joseph M Scarrott, Thilo H Pohle, Thilo H Pohle, Ping Liu, Ping Liu, Ayda Mayer, Ayda Mayer, Adam Brown, Adam J Brown, David C. James, and David C James. Engineering of the CMV promoter for controlled expression of recombinant genes in HEK293 cells. *Biotechnology Journal*, pages e2200062–e2200062, April 2022. doi: 10.1002/biot.202200062. MAG ID: 4225109427 S2ID: 73a6d3de391692f9a09bb495f50dc12734a0870f.

[195] Hao Zhang, Chunyu Fang, Xinlin Xie, Yicong Yang, Wei Mei, Di Jin, and Peng Fei. High-throughput, high-resolution deep learning microscopy based on registration-free generative adversarial network. *Biomedical Optics Express*, 10(3):1044, March 2019. ISSN 2156-7085, 2156-7085. doi: 10.1364/BOE.10.001044. URL https://opg.optica.org/abstract.cfm?URI=boe-10-3-1044.

[196] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression, 2019. URL https://arxiv.org/abs/1911.08287. Version Number: 1.

[197] Zhengxia Zou, Zhengxia Zou, Zhenwei Shi, Zhenwei Shi, Yuhong Guo, Yuhong Guo, Yuhong Guo, Jieping Ye, Jieping Ye, Jieping Ye, and Jieping Ye. Object Detection in 20 Years: A Survey. *arXiv: Computer Vision and Pattern Recognition*, May 2019. doi: 10.1109/jproc.2023.3238524. ARXIV_ID: 1905.05055 MAG ID: 2944165510 S2ID: bd040c9f76d3b0b77e2065089b8d344c9b5d83d6.

[198] Zhichao Liu, Zhichao Liu, Luhong Jin, Luhong Jin, Jincheng Chen, Jincheng Chen, Qiuyu Fang, Qiuyu Fang, Sergey Ablameyko, Sergey Ablameyko, Zhaozheng Yin, Zhaozheng Yin, Yingke Xu, and Yingke Xu. A survey on applications of deep

learning in microscopy image analysis. *Computers in Biology and Medicine*, 134: 104523–104523, May 2021. doi: 10.1016/j.compbiomed.2021.104523. MAG ID: 3173032895.

[199] Zhifang Guo, Jianguo Mao, Rui Tao, Long Yan, Kazushige Ouchi, Hong Liu, and Xiangdong Wang. Audio Generation with Multiple Conditional Diffusion Model. *AAAI Conference on Artificial Intelligence*, August 2023. doi: 10.48550/arxiv.2308.11940. ARXIV_ID: 2308.11940 MAG ID: 4386148299 S2ID: 3e27dc211b1e1845d415b1fb9116130fd057f539.

[200] Zhifeng Kong, Zhifeng Kong, Zhifeng Kong, Wei Ping, Wei Ping, Wei Ping, Jiaji Huang, Jiaji Huang, Jiaji Huang, Kaixuan Zhao, Kexin Zhao, Kexin Zhao, Bryan Catanzaro, and Bryan Catanzaro. DiffWave: A Versatile Diffusion Model for Audio Synthesis. *arXiv: Audio and Speech Processing*, 2020. ARXIV_ID: 2009.09761 MAG ID: 3087665158 S2ID: 34bf13e58c7226d615afead0c0f679432502940e.

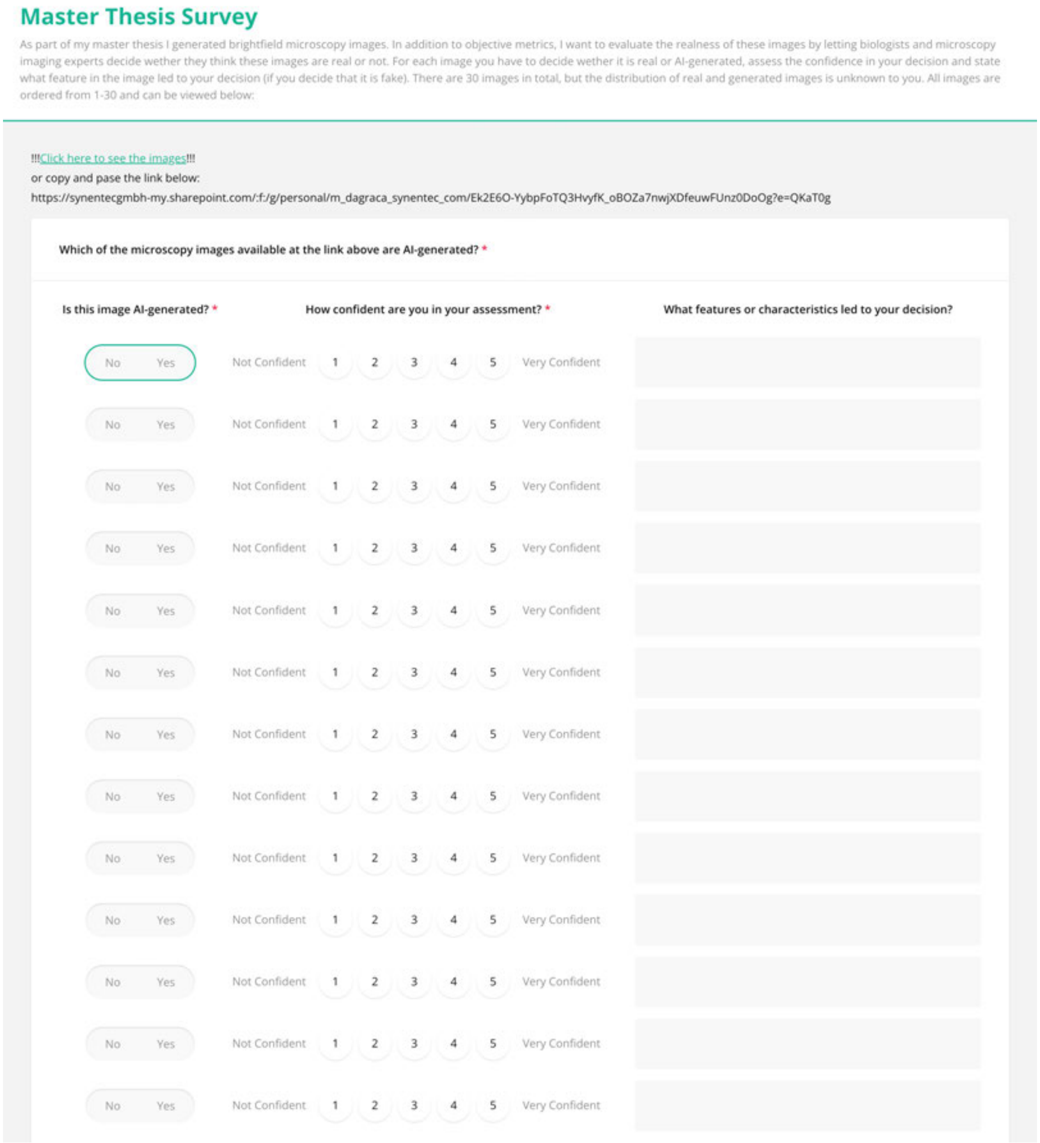
[201] Zhong-Qiu Zhao, Zhong-Qiu Zhao, Peng Zheng, Peng Zheng, Shou-Tao Xu, Shou-Tao Xu, Xindong Wu, and Xindong Wu. Object Detection With Deep Learning: A Review. *IEEE Transactions on Neural Networks*, 30(11):3212–3232, January 2019. doi: 10.1109/tnnls.2018.2876865. ARXIV_ID: 1807.05511 MAG ID: 2884367402 S2ID: 7998468d99ab07bb982294d1c9b53a3bf3934fa6.

[202] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. *ArXiv*, abs/2010.04159, 2020. URL https://api.semanticscholar.org/CorpusID:222208633.

[203] Zixiao Wang, Farzan Farnia, Zhenghao Lin, Yunheng Shen, and Bei Yu. On the Distributed Evaluation of Generative Models. 2023. ARXIV_ID: 2310.11714 S2ID: 9a7f7dc8a2eb9a153ed07325b47bc77b4c2128ce.

[204] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. ResShift: Efficient Diffusion Model for Image Super-resolution by Residual Shifting. *Neural Information Processing Systems*, July 2023. doi: 10.48550/arxiv.2307.12348. ARXIV_ID: 2307.12348 MAG ID: 4385291271 S2ID: d4a60ef37125fcde198781c2eb578a9c9dc78c1c.

A Survey



Master Thesis Survey

As part of my master thesis I generated brightfield microscopy images. In addition to objective metrics, I want to evaluate the realness of these images by letting biologists and microscopy imaging experts decide wether they think these images are real or not. For each image you have to decide wether it is real or AI-generated, assess the confidence in your decision and state what feature in the image led to your decision (if you decide that it is fake). There are 30 images in total, but the distribution of real and generated images is unknown to you. All images are ordered from 1-30 and can be viewed below:

!!!Click here to see the images!!!
or copy and pase the link below:
https://synentecgmbh-my.sharepoint.com/:f:/g/personal/m_dagraca_synentec_com/Ek2E6O-YybpFoTQ3HvyfK_oBOZa7nwjXDfeuwFUnz0DoOg?e=QKaT0g

Which of the microscopy images available at the link above are AI-generated? *

Is this image AI-generated? *	How confident are you in your assessment? *	What features or characteristics led to your decision?
No Yes	Not Confident 1 2 3 4 5 Very Confident	
No Yes	Not Confident 1 2 3 4 5 Very Confident	
No Yes	Not Confident 1 2 3 4 5 Very Confident	
No Yes	Not Confident 1 2 3 4 5 Very Confident	
No Yes	Not Confident 1 2 3 4 5 Very Confident	
No Yes	Not Confident 1 2 3 4 5 Very Confident	
No Yes	Not Confident 1 2 3 4 5 Very Confident	
No Yes	Not Confident 1 2 3 4 5 Very Confident	
No Yes	Not Confident 1 2 3 4 5 Very Confident	
No Yes	Not Confident 1 2 3 4 5 Very Confident	
No Yes	Not Confident 1 2 3 4 5 Very Confident	
No Yes	Not Confident 1 2 3 4 5 Very Confident	
No Yes	Not Confident 1 2 3 4 5 Very Confident	

Fig. A.1: Survey form – Part 1.

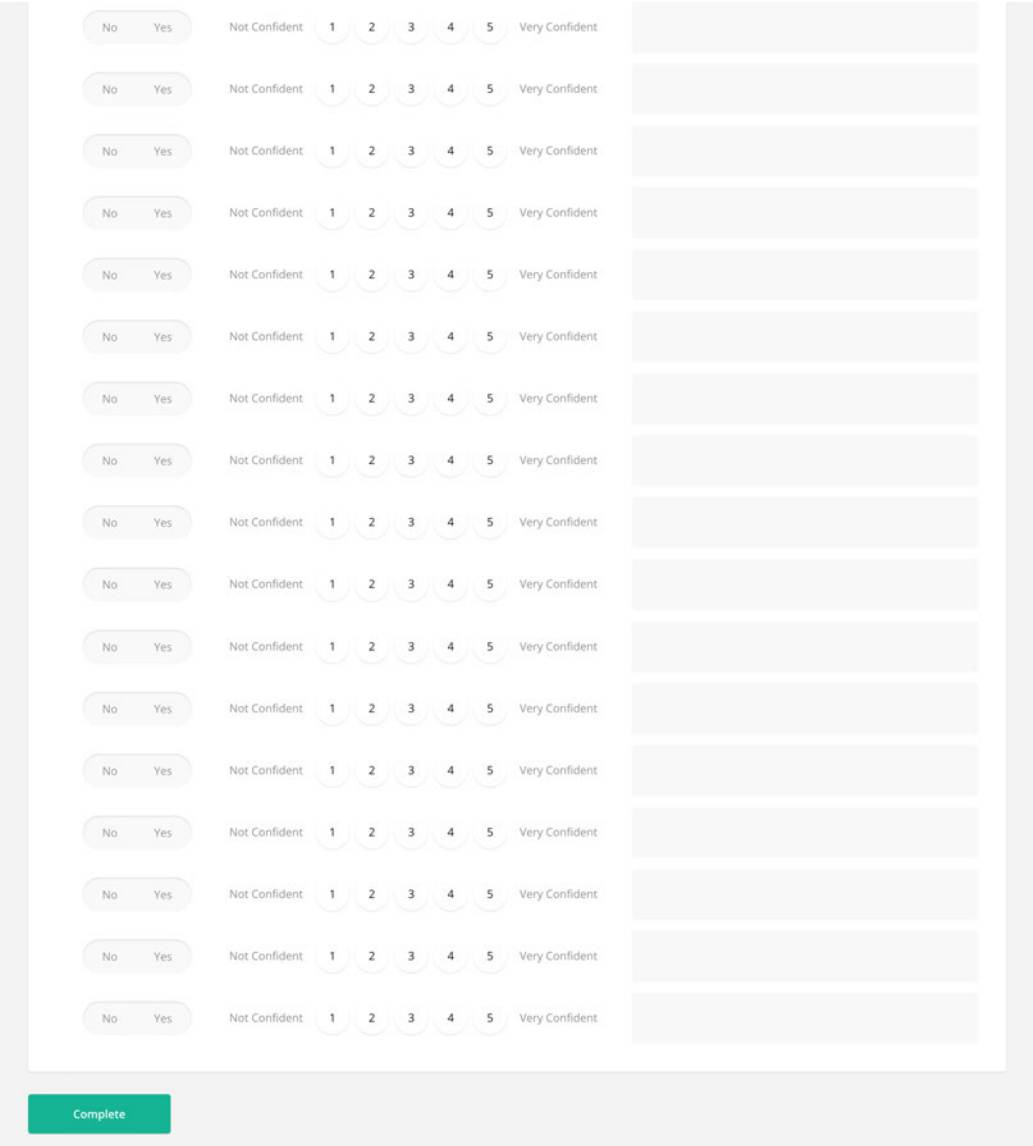
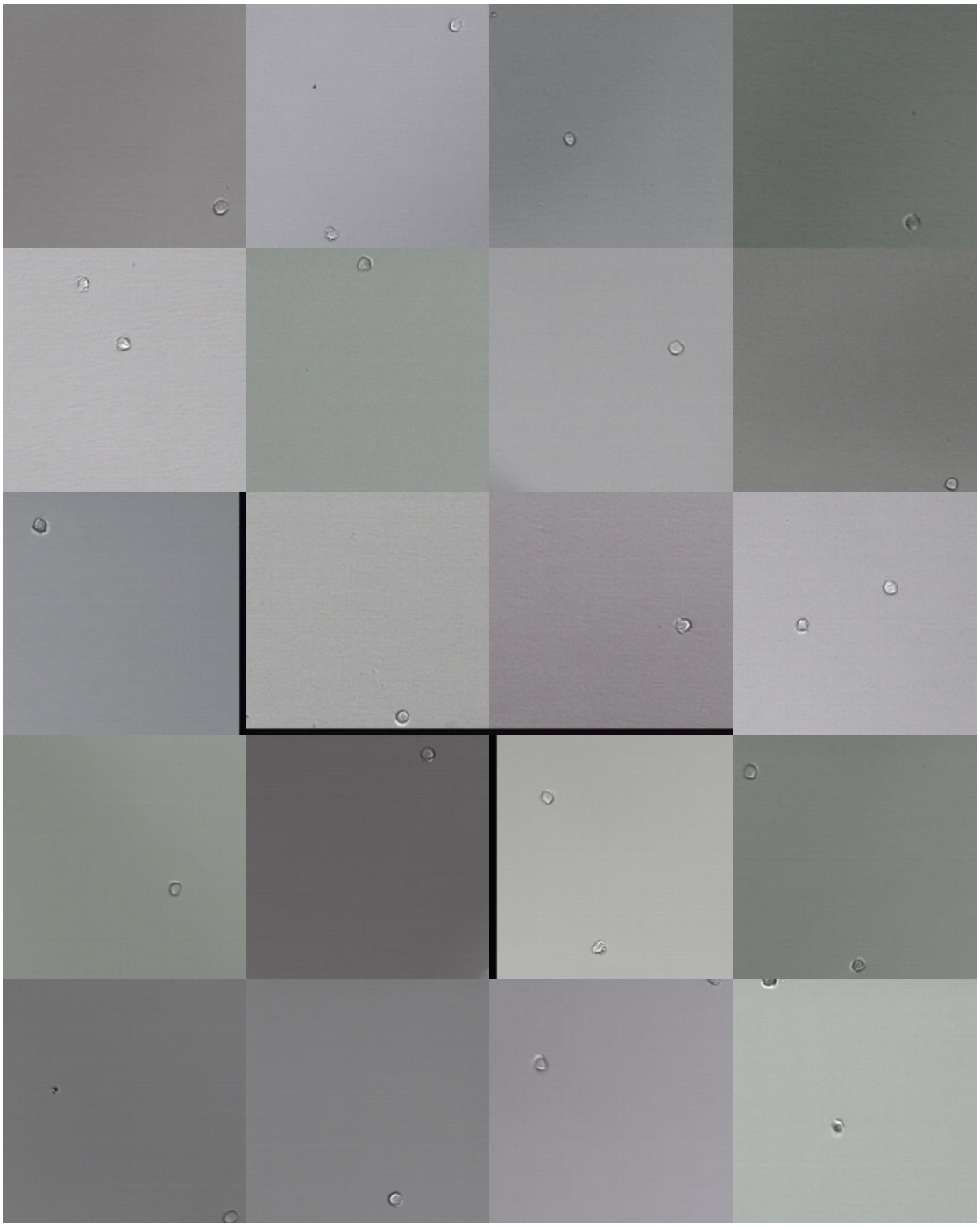
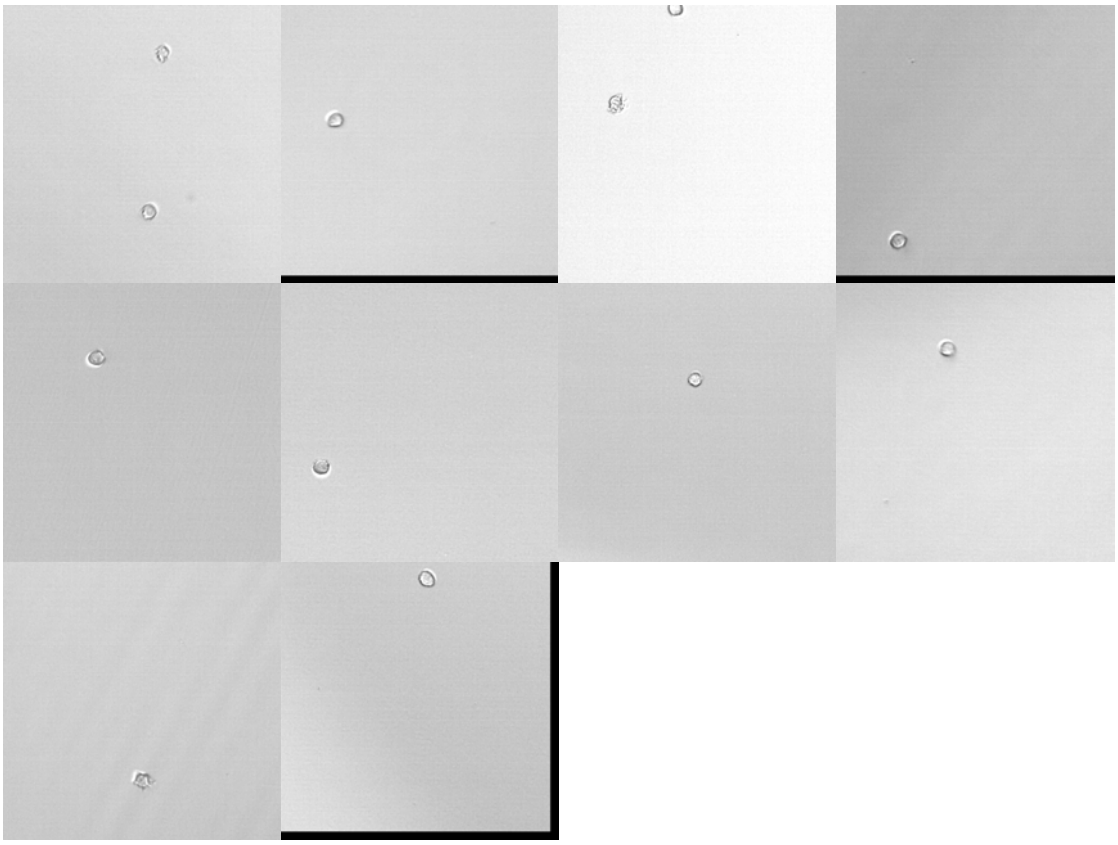


Fig. A.2: Survey form – Part 2.

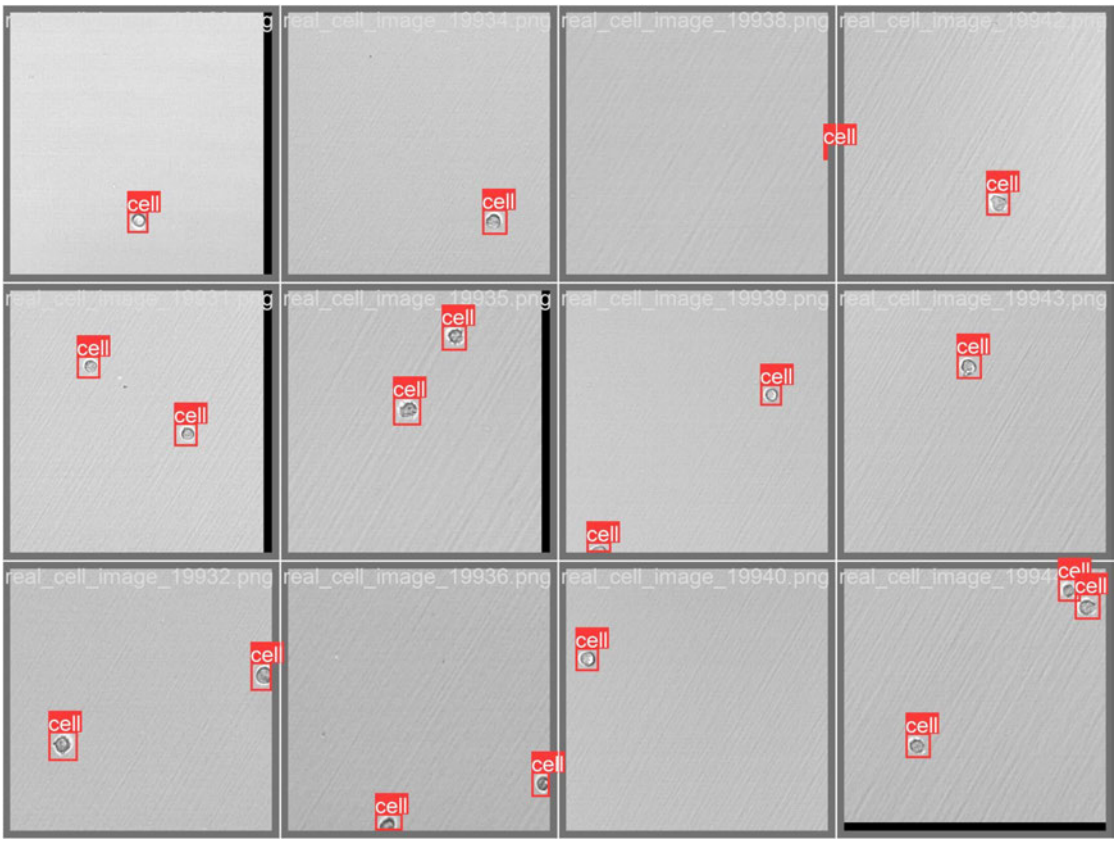
Tab. A.1: Generated images employed in the survey.



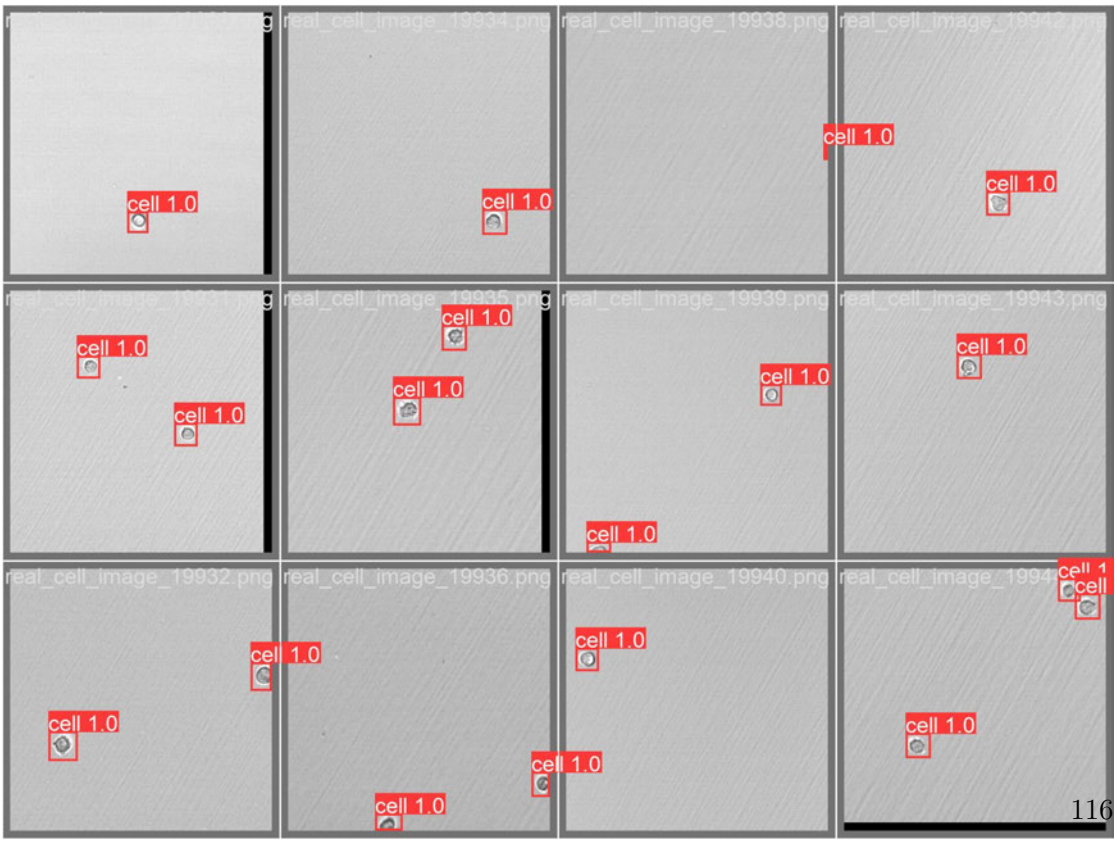
Tab. A.2: Real images employed in the survey.



B Sample Detections



(a) Labels of a sample batch.



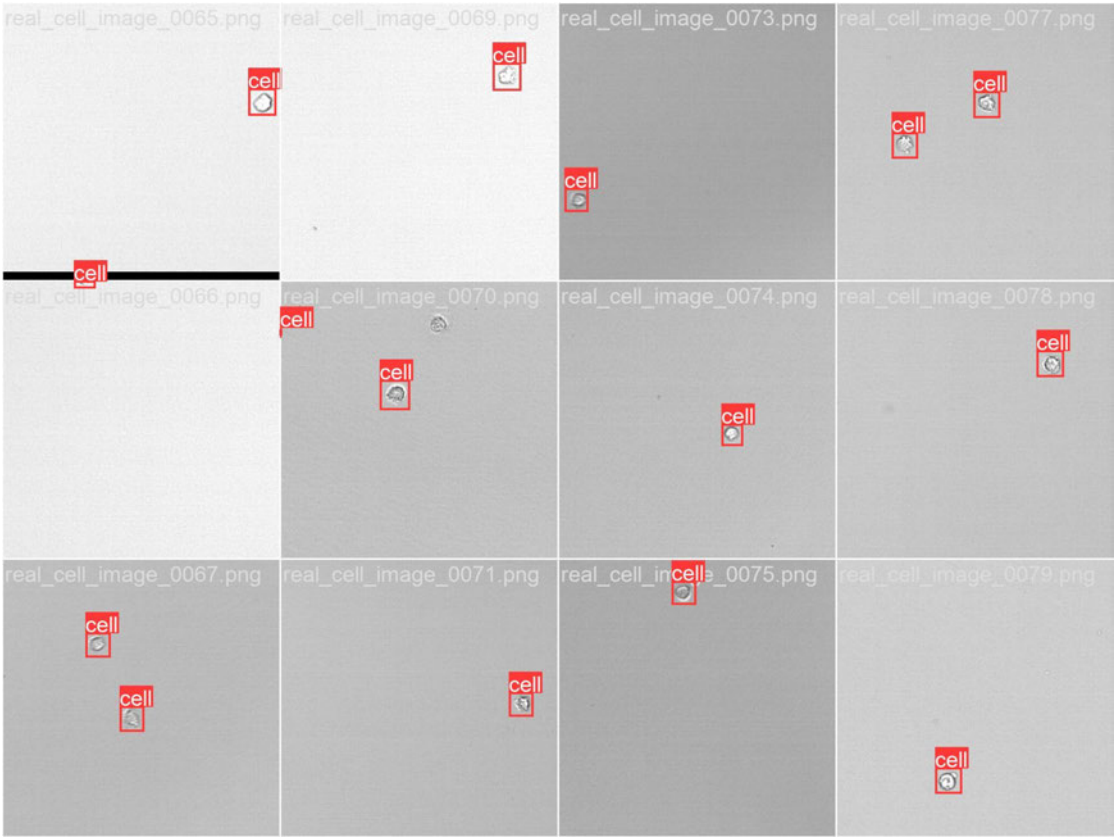
(b) YOLOv8s predictions of a sample batch.

Fig. B.1: Comparison of labels and predictions from the YOLOv8s model that was trained on the *scc_30* dataset.

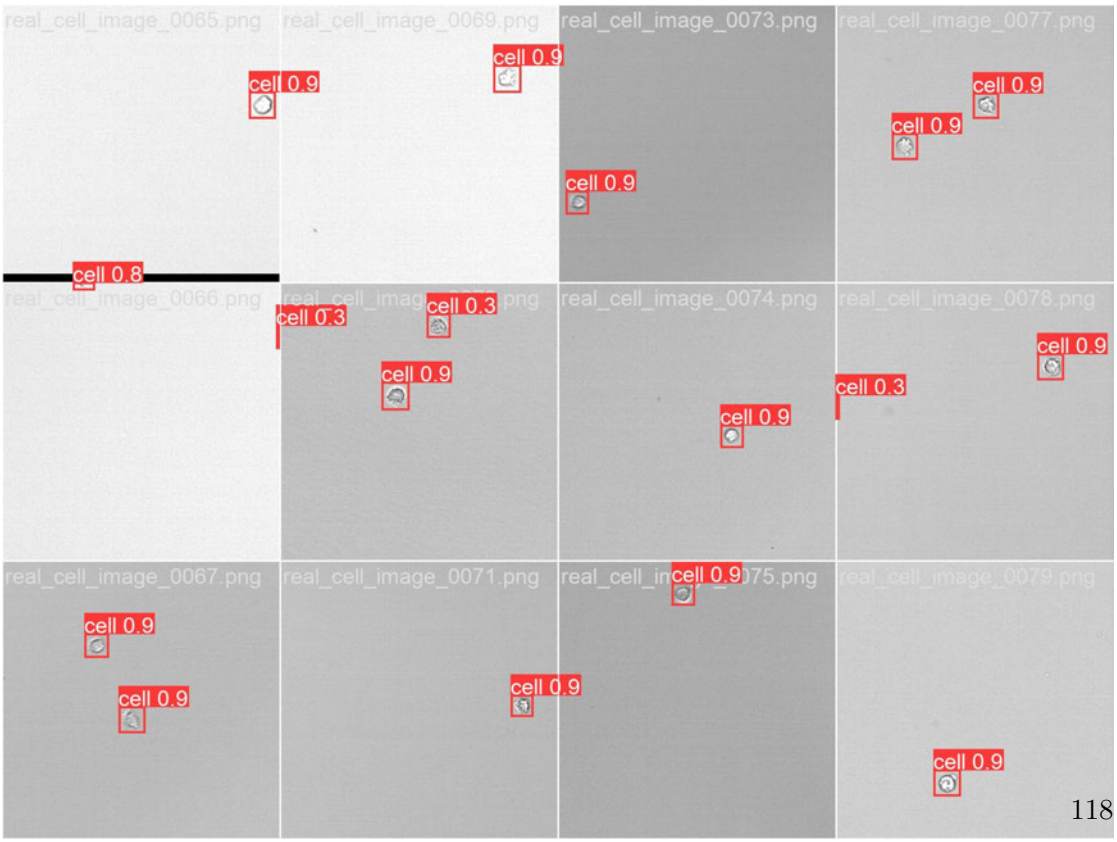
(a) Labels of a sample batch.

(b) YOLOv9e predictions of a sample batch.

Fig. B.2: Comparison of labels and predictions from the YOLOv9e model that was trained on the *scc_10* dataset.



(a) Labels of a sample batch.



(b) RT-DETR-l predictions of a sample batch.

Fig. B.3: Comparison of labels and predictions from the RT-DETR-l model that was trained on the *scc_10* dataset.

Glossary

antibiotic selection marker A gene that confers resistance to a specific antibiotic, used to select for cells that have successfully incorporated a plasmid [62]

brightfield microscopy A basic light microscopy technique where light is transmitted through a specimen, creating contrast as the light is absorbed or scattered by dense areas in the sample

cell line A population of cells derived from a single cell and therefore consisting of cells with a uniform genetic makeup [57]

cell viability The proportion of live cells within a population

chinese hamster ovary cells A cell line derived from the ovary of a Chinese hamster, commonly used in biological and medical research [187]

cytomegalovirus A genetic element used to drive strong and consistent expression of genes in mammalian cells [19]

enhanced green fluorescent protein A genetically engineered version of the green fluorescent protein that produces a brighter fluorescence and higher expression in mammalian cells [145]

fluorescence microscopy A type of light microscopy that uses fluorescence to generate an image. Fluorescent molecules are excited by light of a specific wavelength and emit light of a longer wavelength

high-throughput screening A method for scientific experimentation that uses automation to rapidly conduct a large number of tests [107]

label-free imaging Microscopy techniques that allow for the visualization of samples without the use of labels or stains [17]

microtiter plate A flat plate with multiple "wells" used as small test tubes in various laboratory procedures

plasmid A small, circular piece of DNA that can replicate independently of chromosomal DNA

stably transfected Refers to cells that have permanently incorporated foreign DNA into their genome [90]

Erklärung zur selbständigen Bearbeitung

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbständig verfasst und nur die angegebenen Hilfsmittel benutzt habe. Wörtlich oder dem Sinn nach aus anderen Werken entnommene Stellen sind unter Angabe der Quellen kenntlich gemacht.

Ort Datum Unterschrift im Original