UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE ENGENHARIA DE COMPUTAÇÃO

TIAGO COMASSETTO FRÓES

Exploring Image-to-Image Translation Techniques for Microscopy Images

Work presented in partial fulfillment
of the requirements for the degree of
Bachelor in Computer Engineering

Advisor: Prof. Dr. Claudio Rosito Jung

Porto Alegre
August 2024

”It is when your practice is rather greedy
that you become discouraged with it.”
— SHUNRYU SUZUKI

AGRADECIMENTOS

Primeiramente, agradeço aos meus pais, irmão e avós por terem me dado apoio e suporte durante toda a graduação. Sem ter todos ao meu lado, nada disso seria possível. Agradeço também a toda a restante família que, de alguma forma, participou dessa jornada.

Agradeço ao meu orientador, Prof. Claudio Jung, pela paciência, conhecimento e dedicação ao longo dos últimos dois semestres. Tudo isso tornou possível este trabalho. Também agradeço a todos os professores com quem tive aula ao longo do curso e à Universidade Federal do Rio Grande do Sul, que propiciaram a minha formação.

Agradeço a Angelo Angonezi por proporcionar os dados utilizados neste trabalho, assim como por compartilhar seu conhecimento. Sua ajuda foi essencial para a realização deste trabalho.

Por fim, um agradecimento especial aos colegas e amigos Athos Lagemann, Gabriel Conte, João Gubert e Rodrigo Wuerdig, pela parceria nas horas de estudo, nos almoços no restaurante universitário e nos inúmeros cafés com paçoca nos intervalos entre as aulas.

ABSTRACT

Fluorescent labeling plays a crucial role in the development of new drugs and the analysis and diagnosis of tumors. Biologists utilize these images to examine cellular morphology, structures, and phenotypes, which can be done manually or, more recently, automated by software providing a quantitative view. Although regarded as essential, fluorescent labeling has limitations such as being costly, time-consuming, and error-prone due to phototoxicity and photobleaching. Motivated by consolidated and recent advances in image generation techniques, we explore image-to-image translation methodologies to translate bright-field microscopy images, i.e., label-free, into fluorescence microscopy. After training established and modern models, our findings reveal that GANs deliver the highest-quality results, but with increased computational demands. In contrast, Latent Diffusion Models provide slightly lower quality outcomes but require significantly less computational power, suggesting promising results for future works, especially when working with larger datasets.

Keywords: Deep Learning. Image-to-Image Translation. In-Silico Labeling. Fluorescent Microscopy.

Explorando Técnicas de Translação de Imagem para Imagem para Imagens de Microscopia

RESUMO

A marcação fluorescente desempenha um papel crucial no desenvolvimento de novos medicamentos e na análise e diagnóstico de tumores. Biólogos utilizam essas imagens para examinar a morfologia celular, estruturas e fenótipos, o que pode ser feito manualmente ou, mais recentemente, de forma automatizada por software, proporcionando uma visão quantitativa. Embora seja considerada essencial, a marcação fluorescente apresenta limitações, como custo elevado, ser demorada e propensa a erros devido à fototoxicidade e fotodegradação. Motivados por avanços consolidados e recentes em técnicas de geração de imagens, exploramos metodologias de tradução de imagem para imagem para converter imagens de microscopia de campo claro, ou seja, sem marcação, em fluorescência de microscopia. Após treinar modelos estabelecidos e modernos, nossos resultados revelam que GANs entregam os melhores resultados em termos de qualidade, embora com maiores demandas computacionais. Em contraste, os modelos de Difusão Latente fornecem resultados de qualidade ligeiramente inferiores, mas com uma necessidade de poder computacional significativamente menor, sugerindo resultados promissores para trabalhos futuros, especialmente ao trabalhar com conjuntos de dados maiores.

Palavras-chave: Aprendizado Profundo. Translação de Imagem para Imagem. Marcação In-Silico. Microscopia de Fluorescência .

LIST OF ABBREVIATIONS AND ACRONYMS

BBDM	Brownian Bridge Diffusion Model
cGAN	Conditional Generative Adversarial Network
FBI	French BioImaging
FDI	Fréchet Inception Distance
GAN	Generative Adversarial Network
GFP	Green Fluorescent Protein
GPU	Graphics Processing Unity
ISBI	International Symposium on Biomedical Imaging
KL	Kullback-Leibler
LBBDM	Latent Brownian Bridge Diffusion Model
LDM	Latent Diffusion Model
MAE	Mean Average Error
MSE	Mean Squared Error
PCC	Pearson Correlation Coefficient
PI	Propidium Iodide
PSNP	Peak Signal-to-Noise Ratio
S/N	Signal-to-Noise Ratio
SSIM	Structural Similarity Index
UMAP	Uniform Manifold Approximation
VAE	Variational Autoencoder
VAE-GAN	Variational Autoencoder Generative Adversarial Network
VQGAN	Vector Quantized Generative Adversarial Network
wGAN	Wasserstein Generative Adversarial Network

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

The use of microscopes has enabled the study of cells in biology. As an indispensable tool for biologists, microscopy imagery enables the exploration of cellular structures and processes that are critical to the advancement of fields like microbiology, pharmacology, and neuroscience. Recent technological advances, including event-driven and super-resolution microscopy, have augmented the capabilities of this essential tool. The synergy between technological advancements and microscopy sets the stage for a groundbreaking alliance with deep learning (PYLVÄNÄINEN et al., 2023). This work delves into the convergence of deep generative models and microscopy, specifically exploring the potential of *in silico labeling* -— an artificial staining approach -— to reshape the landscape of biological research.

Cell staining is a crucial technique to enhance the microscopic observation of cells. This procedure involves the application of colored or fluorescent dyes to biological specimens, such as cells or tissues, to accentuate specific structures or molecules selectively. Fluorescent staining, in particular, is commonly used in high-throughput screening assays, where automated imaging systems rapidly analyze numerous samples. The capability to multiplex and measure fluorescence signals renders it well-suited for such applications, which are at the forefront of, for example, drug discovery. However, it is important to acknowledge that fluorescence labeling has limitations such as being costly, time-consuming, variability in its application, and phototoxicity; other problems can happen in live cell imaging, like the likelihood of some protocols killing cells, which can interfere with cell tracking.

This study centers on the approach proposed by Christiansen et al. (2018) called *in silico labeling*, which hypothesizes that unlabeled images (i.e., cell images deprived of fluorescent staining) contain more information about cell structures than initially apparent, especially in the brightfield microscopy modality. Brightfield images rely solely on the contrast of cells -— primarily composed of water —- and light, resulting in low-contrast images not propitious to the analysis of the human eye. However, they can generate high-definition images and are widely embraced as a standard imaging method within the scientific community. As also shown by Wieslander et al. (2021), Cross-Zamirski et al. (2022) and Cross-Zamirski et al. (2023), deep generative models can be employed to generate fluorescent-stained images from brightfield images to leverage this unnoticeable information.

Deep generative models are trained on a particular dataset to autonomously generate new, realistic data samples from an unseen input. If properly trained, the model will exhibit the capacity to capture and replicate complex patterns within the input data distribution, enabling the virtual staining of a brightfield image. Various architectures and methods of generative image models have been proposed in the past. Variational Autoencoders (VAEs) showed impressive results reconstructing images from its latent space representation, Kingma and Welling (2013). Generative Adversarial Networks (GANs) (GOODFELLOW et al., 2014) are highly praised and have numerous applications and variations. However, the use of GANs poses challenges, such as training instabilities and the potential to generate hallucinations, compromising its utility in biological downstream tasks (COHEN; LUCK; HONARI, 2018). In recent years, a new family of deep generative models has become the new *state-of-art* in several image generation tasks, the so-called Denoising Diffusion Probabilistic Models (DDPMs) (HO; JAIN; ABBEEL, 2020), also called diffusion models. They offer several advantages over GANs, such as stable training, easy scalability, and the ability to generate high-quality images.

The main goals of this work are to introduce the reader to basic concepts and recent advances in the *in-silico labeling* field, explore strategies of image-to-image translation, and evaluate its effectiveness on a particular dataset of microscopy images. The structure of this thesis is organized as follows. It begins with the *Theoretical Foundations*, which explores the concepts of live cell imaging, phototoxicity and photobleaching, generative image models, and in-silico labeling, providing a solid foundation for understanding the research. The *Related Work* chapter follows, delving into prior advancements in Image-to-Image Translation and In-silico Labeling, highlighting key developments in the field. Next, the *Proposed Methodology* outlines the dataset, tested models, and evaluation techniques employed in the study, serving as the core of the research approach. The *Experimental Results* chapter presents both quantitative and qualitative evaluations of the models, offering insight into the outcomes of the experiments. Finally, the *Conclusion* summarizes the findings and contributions of the work.

2 THEORETICAL FOUNDATIONS

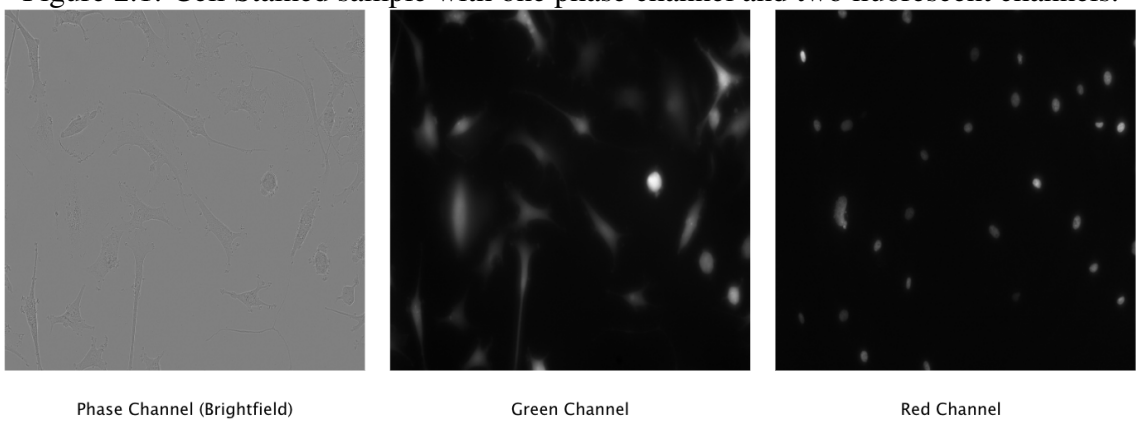
2.1 Microscopy

Microscopy has long been indispensable in unraveling the intricacies of cellular structures and functions. To better capture images of cellular elements, traditional microscopy often involves fixing samples, which is the process of "freezing" the sample in that particular state. After a sample is fixed, it can undergo further processes like staining or sectioning. However, the fixation process has notable drawbacks, including the loss of biological activity, potential alterations in cellular morphology, and mainly, it limits the ability to observe dynamic cellular processes in real-time. This static nature of traditional microscopy limits the insights that biologists can take from cell-cell interactions and other dynamic biological events, as discussed by Haraguchi (2002).

2.1.1 Live Cell Imaging

According to Cole (2014), live-cell imaging is a microscopy modality where samples are not fixed. Alive cells offer biologists the chance to get a better understanding of cell-cell interaction and dynamic biological events, such as cell division, migration, state, signaling, and responses to stimuli throughout time. Such imaging modality offers valuable insights into cellular behavior and molecular processes not just visually but, through image processing, quantitatively. To achieve such conditions, samples undergo a staining process with fluorescent markers, and still, to excite and make the stained structure visible, they must be exposed to strong light. An example of a Cell Stainined sample is shown in Fig. 2.1, and cell structures are clearly highlighted in the stained version. Despite the advantages of cell staining, the process of adding fluorescent markers and exposing the samples to light can harm the sample, and this damage is called phototoxicity. Optimizing imaging conditions to maximize signal-to-noise ratio (S/N) and sample health is an ongoing challenge many researchers face in every experiment Icha et al. (2017).

Figure 2.1: Cell Stained sample with one phase channel and two fluorescent channels.



2.1.2 Phototoxicity and Photobleach

To activate the applied fluorescent tags (fluorophores), the sample must be exposed to strong light throughout its life. Explained by Icha et al. (2017), this process can lead to phototoxicity, which is health damage, and photobleaching, which is the loss of fluorescence signaling, both due to excessive exposure to light over time. Both have significant negative side effects on the sample *Phototoxicity* can slow down its cycle, modify processes such as cellular respiration and protein synthesis, and induce cell death which is a valuable morphological characteristic in numerous studies, such as those investigating the efficacy of specific drugs in cancer treatments. Photobleaching occurs when repeated light exposure weakens or extinguishes a sample's fluorescent signal, making long sample observations have diminishing or even misleading results. This highlights how *Phototoxicity* and *Photobleaching* can potentially impact study outcomes, potentially leading to false negatives detrimentally.

Managing and understanding *phototoxicity* and *photobleaching* is essential to obtaining high-quality and healthy live cell imaging so that it can be transformed into quantitative data to make experiments reliable and reproducible. Many techniques to mitigate these outcomes such as lowering illumination intensity, controlling time exposure, and managing oxygen concentration in the cell culture, can preserve sample health but will result in lower imaging quality. Although well-established, these techniques continue to present significant challenges that researchers encounter in routine experiments.

2.1.3 Computer Vision and Microscopy

Attaching cameras to microscopes enables scientists to capture and store images of what is being observed. Since these images contain biological information, computer vision algorithms can be used for analysis, assistance, or even automation of tasks that can be time-consuming and prone to error when done manually by a biologist. These advances might enable scientists to carry novel analysis methodologies to advance their field of research.

One of the main computer vision-based tasks in microscopy is *Cell Segmentation*, which can be used to count the number of cells in a particular image, but also to get individual morphology features of each cell – this is called single-cell analyses. The base of many classical cell segmentation algorithms is *Intensity Thresholding* (QIAO et al., 2007), which assumes that the foreground (cell) and the background have notable intensity differences (contrast). This intensity difference may be global or local, and because of that, a fixed or histogram-based adaptative threshold must be used (PLISSITI; VRIGKAS; NIKOU, 2015).

Nowadays, with the rapid progress of deep learning, more methodologies and applications are being created, many of which are fit for microscopy images. Pylvänäinen et al. (2023) divide these techniques into two categories: *Image Analysis* includes tasks such as identifying cell state with image classification, single cell analysis, including morphology, and cell counting using instance segmentation, time-lapse analysis with object tracking and profiling, the act of extracting cell features for downstream analysis; and *Image Acquisition* that centers on enhancing the quality of the images to be analyzed, which include diminishing *phototoxicity* and *photobleaching* by using techniques to increase image resolution and diminish signal-to-image ratio.

2.2 Generative Image Models

Generative image models are a class of deep learning models designed to generate new images similar to a given dataset. Training these models involves learning the data distribution of a particular dataset, which consists of the distribution of pixels and structure patterns that constitute the images in the training set. Generative image models can be utilized in tasks such as style transfer, image restoration, and conditioned image generation. In recent years, generative image models conditioned by text gained popularity

due to their flexibility in generating high-quality images, and the work by Rombach et al. (2022) is a notable example. This work focuses on image-conditioned models for the task of *Image-to-Image Translation* applied to microscopy images.

2.2.1 Variational Autoencoders (VAEs)

Introduced by Kingma and Welling (2013), Variational Autoencoders (VAEs) are a class of probabilistic generative models known for their ability to learn a probabilistic distribution of the data resulting in efficient latent representations that can be manipulated (PRINCE, 2023). VAEs can be applied to many image-generation tasks such as image denoising, super-resolution, and image-to-image translation. VAEs are composed by an *encoder*, which maps the input x to the latent space z and the *decoder*, which maps the latent space representation z back to the pixel space, as illustrated in Fig. 2.2[1].

VAEs differ from regular autoencoders, proposed by Hinton and Salakhutdinov (2006). Autoencoders' *encoder* directly map the input to a point in the latent space, on the other hand, VAE's *encoder* output parameters (mean and standard deviation) for a probability distribution from which z is sampled by the decoder through

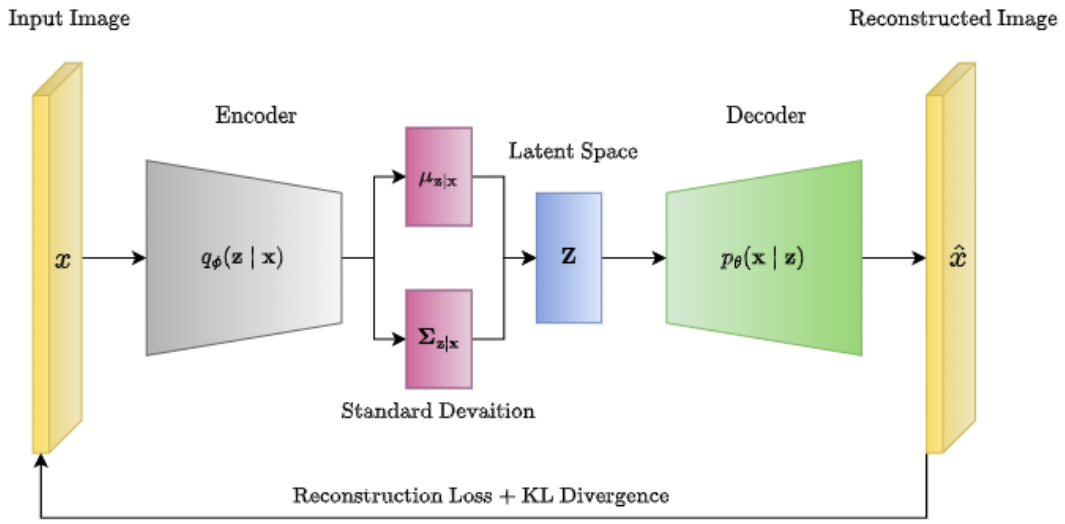
$$z \sim q(z|x) = \mathcal{N}(z; \mu(x), \sigma(x)). \tag{2.1}$$

The objective function of a VAE consists of two components: the reconstruction loss and the regularization term. The reconstruction loss guarantees that the decoded output y' closely approximates the ground truth y. The regularization term, represented by the *Kullback-Leibler (KL) divergence*, prevents overfitting and aligns the learned latent distribution $q(z|x)$ with the prior distribution $p(z)$. This alignment enhances the model's ability to handle imprecise data and unexpected outcomes. Mathematically, it is given as

$$\mathcal{L} = \mathbb{E}_{q(z|x)}[\log p(x|z)] - D_{KL}(q(z|x) \parallel p(z)). \tag{2.2}$$

[1]The referred pixel space may be a paired image y on a different domain for image-to-image translation or the original image x in the original domain when doing image reconstruction

Figure 2.2: Standard VAE architecture.

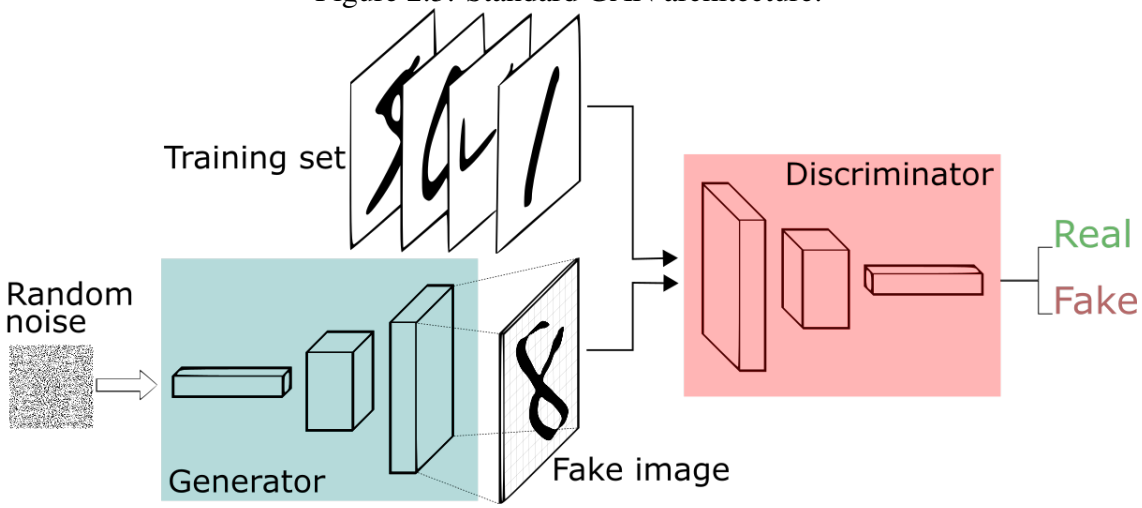


Source: <https://shorturl.at/9pzAJ>

2.2.2 Generative Adversarial Networks (GANs)

A *Generative Adversarial Network* (GAN) consists of two networks: the *Generator* and the *Discriminator*. The training of a GAN can be seen as a min-max game between both networks since the role of the *Discriminator* is to correctly classify whether it's input is the *ground truth* image or the image generated by the *generator*. A standard GAN architecture can be seen in Fig. 2.3. GANs are notoriously difficult to train as the non-cooperative nature of a min-max game does not guarantee convergence (ARJOVSKY; BOTTOU, 2017). The original loss function of the entire GAN model is exclusively based on the loss function of one of the networks, the *Discriminator*, which is derived from the *binary cross-entropy* loss function, which presents limitations to the model (GOODFELLOW et al., 2014). Alternative loss functions, such as the Wasserstein loss proposed by Arjovsky, Chintala and Bottou (2017), have been suggested. Despite these alternatives, the challenge persists without a mechanism allowing the flexibility to optimize the network for specific objectives. To enhance model performance, additional information associated with the input images, such as labels or segmentation maps, can be included as input to the *Generator* during the training phase. This inclusion facilitates convergence by introducing a supplementary loss function conditioned on these labels. The new loss value is combined with the standard one in a weighted form. This variant of the traditional GAN architecture is referred to as a *conditional GAN* (cGAN).

Figure 2.3: Standard GAN architecture.



Source: https://sthalles.github.io/intro-to-gans/

2.2.3 Denoising Diffusion Probabilistic Models (DDPMs)

Denoising Diffusion Probabilistic Models (DDPMs) (HO; JAIN; ABBEEL, 2020) are a class of deep-learning models typically used for image generation. DDPMs are characterized by the diffusion process, also referred to as the forward process or noising process, which consists of incrementally adding noise to the input image until it becomes random in a Gaussian distribution manner. The diffusion process does not involve any learning, the noise is added in a systematic and predictive manner, as defined by Eq. (2.3) and (2.4) where x_0 is the original data, x_t, is the data at timestep t, β_t is the variance schedule, which is bound to t, $\mathcal{N}$ represents a Gaussian distribution and T is the final timestep. The learning step happens in the denoising process: the model learns how to reverse the added noise to get an approximation of the original image. The training process is based on predicting the noise distribution and removing it so the original image is revealed. Since the forward process is done in a systemic time-stepped manner, for each step $t + 1$, a single amount of noise is added to the image, which gives room to create many training schedules, However, the most common approach involves the denoising network predicting the noise added to the image at time step $t - 1$. The reverse process is given by Eq. (2.5) here, $\mu_\theta(\mathbf{x}_t, t)$ is the predicted mean and, $\Sigma_\theta(\mathbf{x}_t, t)$ is the predicted variance learned by the model at timestep t. θ represents the model's parameters.

During inference, the model starts with a random noise image x_T. The trained denoising network iteratively takes the current noisy image x_t and predicts the image $x_t - 1$. The prediction is the estimation of how much noise was added at the time step t,

the final output is the image x_0, which is a sample from the learned data distribution. This image is a realistic-looking sample generated by the model.

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}). \tag{2.3}$$

$$q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t \mid \mathbf{x}_{t-1}). \tag{2.4}$$

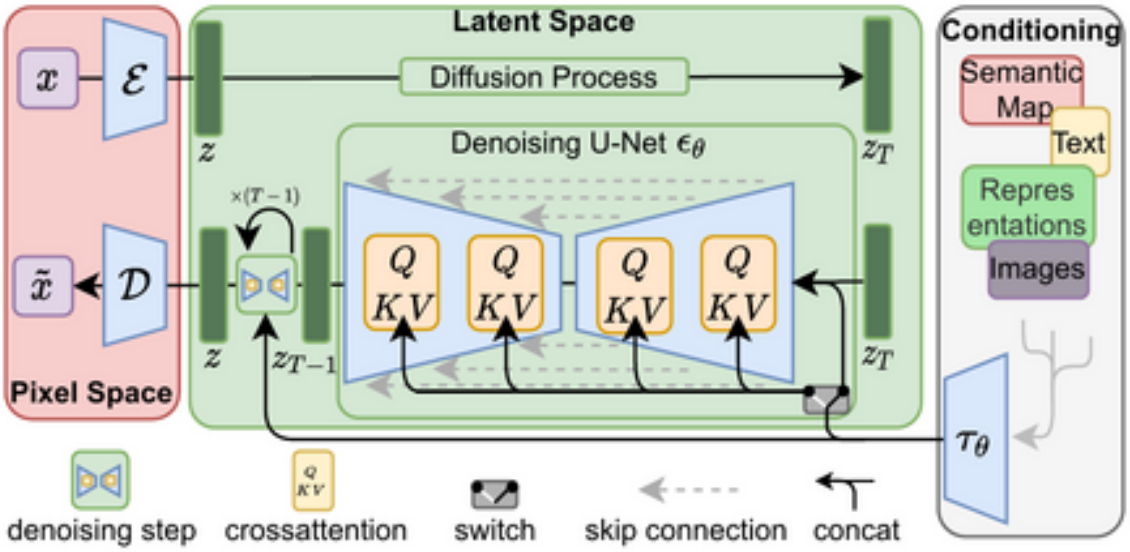
$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)). \tag{2.5}$$

2.2.4 Latent Diffusion Models (LDMs)

Latent Diffusion Models (LDMs) are a class of deep-learning models typically used for image generation (ROMBACH et al., 2022). LDMs work similarly to Denoising Diffusion Probabilistic Models (DDPMs), but instead of operating in the pixel space, LDMs work in the latent space. After mapping the input image to a lower-dimensional latent representation using an encoder, the diffusion process is performed in this compressed space. It involves the same steps – forward and reverse diffusion processes – as in regular DDPMs. The result is a new latent representation, which is then converted back to the pixel space by a decoder.

The ability to work in the latent space has given LDMs greater flexibility, as this allows the incorporation, called *conditioning* of encoded information in other formats, see Fig. 2.4. Examples are text or segmentation masks, which can condition the model to generate new, unseen data. Since LDMs operate on the latent space, which is a lower dimensional representation of the image on the pixel space, they are more computationally efficient than regular DDPMs. It is important to notice that the encoder and decoder used by LDMs are pre-trained, further reducing the computational load compared to DDPMs.

Figure 2.4: The architecture of a LDM



Source: Rombach et al. (2022)

2.3 In-Silico Labeling

Brightfield imaging is characterized by its low cost and accessibility. This imaging technique has proven particularly advantageous for live imaging due to its minimal invasiveness and reduced phototoxicity and photobleaching. The ability to observe biological specimens in their native state over extended periods makes brightfield imaging an invaluable tool for capturing dynamic cellular processes (GUPTA et al., 2022).

Introduced by Christiansen et al. (2018), *in silico labeling* is a non-invasive technique for enhancing brightfield images. The objective is to investigate whether computers can learn and predict features in unlabeled images that are typically only discernible through invasive labeling methods, such as fluorescence dyes. For this purpose, a deep generative model was trained in a supervised manner, where the inputs to the network were z-stack (with three pictures in different focal planes) brightfield unlabeled images and the corresponding stained image, used as the target ground truth. With the intent of building a model capable of generating several different labeled images with different fluorescences, the authors used training pairs from different experiments across various labs, samples, imaging modalities, and fluorescent labels. However, not every input had all the possible fluorescence (ground truth) available. The chosen model was a U-Net (RONNEBERGER; FISCHER; BROX, 2015) based on individual blocks inspired by the inception architecture (SZEGEDY et al., 2015) modified so that spatial information does

not degrade throughout the many layers of the model. The model exhibited accurate per-pixel predictions for some labels, specifically Hoechst and DAPI, which highlight cell nuclei; and *propidium iodide* (PI), which marks dead cells. The authors used pixel-wise cross entropy as the loss function, and the metric used to assess the quality of the synthesized images was the Pearson correlation; both methods consider image similarity in a pixel-wise manner.

3 RELATED WORK

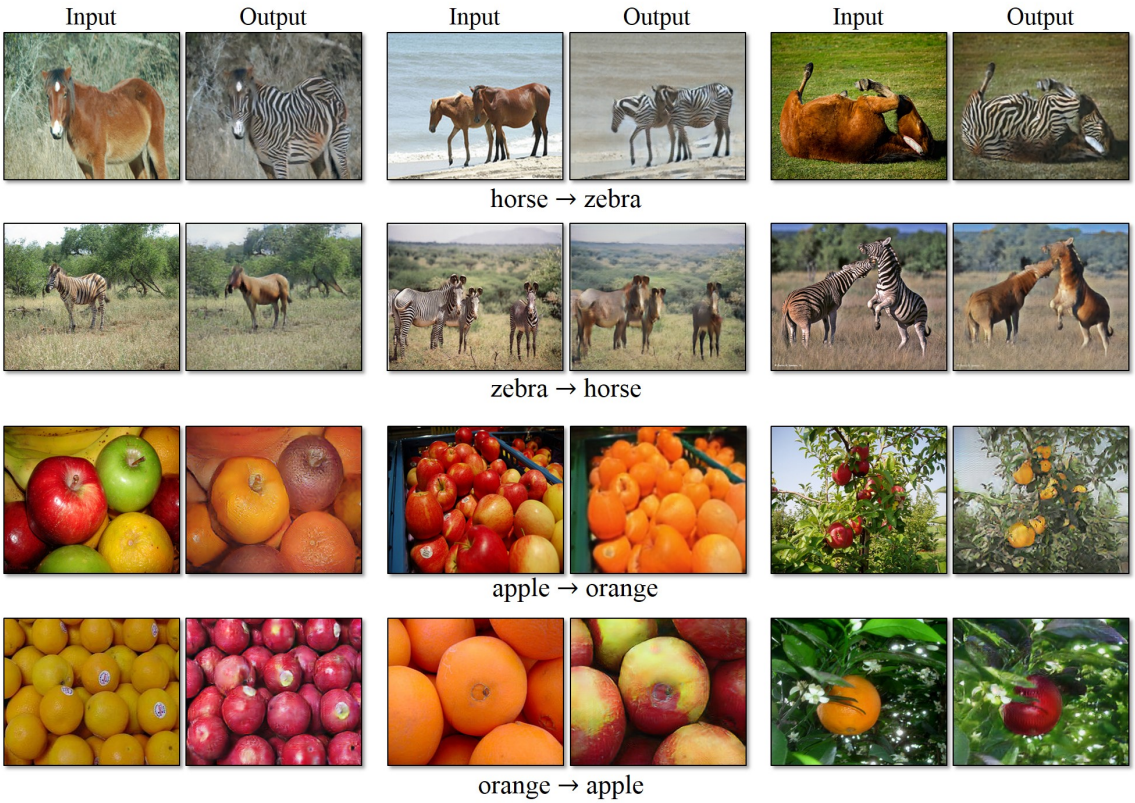
In this section, the reader will be introduced to the main topics and methods used throughout the experiments: *image-to-image translation* and *in-silico labeling*. First, we introduce an overview of generic *image-to-image translation* techniques. We discuss applications and notable models such as VAEs, GANs, (PANG et al., 2021), and the emerging DDPMs and LDMs. Following this, we present a section on *in-silico labeling*, focusing on how *image-to-image translation* is applied to the biomedical field.

3.1 Image-to-Image Translation

Image-to-image translation is a problem where generative models convert an image from one *domain* to another, where the images in different domains may be paired or unpaired. In a dataset with paired images, we have the exact image in two distinct domains, one example of this is colored and *black and white* images, which can be seen as supervised learning. In the unpaired scenario, we have several samples of the two domains (e.g., images during the day and images at night), but not necessarily from the same scene. CycleGAN (ZHU et al., 2017) is a classic example of unpaired domain translation, and one example is shown in Fig. 3.1. Unpaired image-to-image translation can be seen as unsupervised learning. For this purpose, the model should be conditioned with input from the original domain. Examples of common and intuitive applications of *image-to-image translation* are the colorization of black-and-white photos, the translation of sketches into realistic images, and artist-style transfer.

Many image-generation architectures are suited for this task. VAEs, for example, have seen significant success in image-to-image translation by learning shared latent representations that enable smooth domain transformations. Their ability to handle unpaired data makes them versatile for applications such as style transfer. However, VAEs also excel in tasks involving paired images, leveraging their stable training and easily manipulated latent spaces to achieve high-quality translations in tasks such as image colorization (DESHPANDE et al., 2017), super-resolution (LIU; SIU; CHAN, 2020), and domain adaptation.

Figure 3.1: Examples of domain-translation.

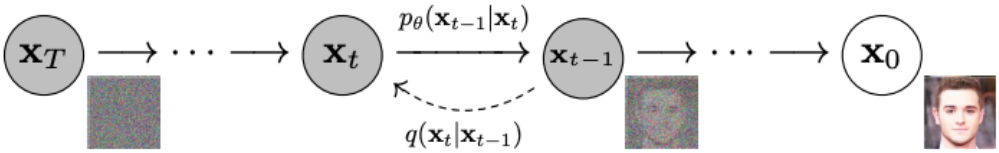


Source: Zhu et al. (2017)

GANs were widely explored due to their two independent stage training process and especially because of the *discriminator* part that can provide visually and contextually consistent results. Among these methods, the *pix2pix* model proposed by Isola et al. (2017) is particularly notable for its ability to generate images based on paired datasets conditionally, ensuring high-quality outputs in domain-translation tasks. Examples of image translation in different domains are shown in Fig. 3.1. Despite the good results achieved by GANs, there are known problems such as training stability and the difficulty of hyper-parameter tuning.

Denoising Diffusion Probabilistic Models (DDPMs) emerged as an alternative to GANs on image generation, and Saharia et al. (2022) showed that this class of methods can be used for *image-to-image translation*. The training of DDPMs does not struggle with training stability and hyper-parameter tuning difficulty since it offers the possibility of controlling the rate of noise injection and denoising. The gradual refinement of the image, illustrated in Fig.3.2, also results in better quality of the generated images. However, it is important to consider that DDPMs require more computational power, not only in training the models but also when sampling new images (inference).

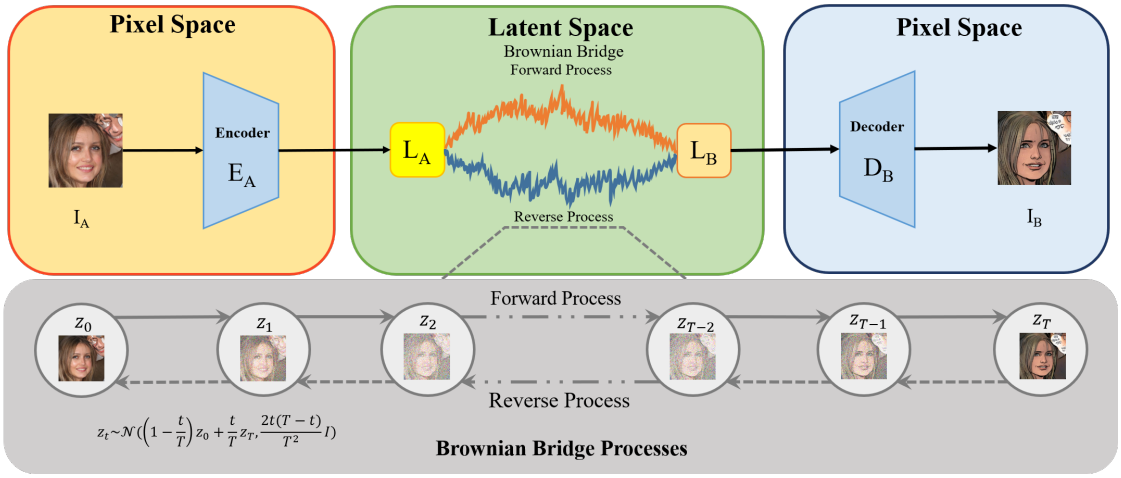
Figure 3.2: The forward process q and the denoising process p_θ on a DDPM



Source: Ho, Jain and Abbeel (2020)

More recently, Latent Diffusion Models (LDMs) have achieved *state-of-the-art* results in image generation manipulating image and text on the latent space. Formulating *image-to-image translation* as a *Brownian Bridge* process (see Fig.3.3), Li et al. (2023) shows encouraging results on tasks such as style transfer on both the pixel and latent space, same result can be seen in Fig 3.4.

Figure 3.3: The architecture and the forward and denoising process of the *Brownian Bridge Diffusion Model* on the latent space.

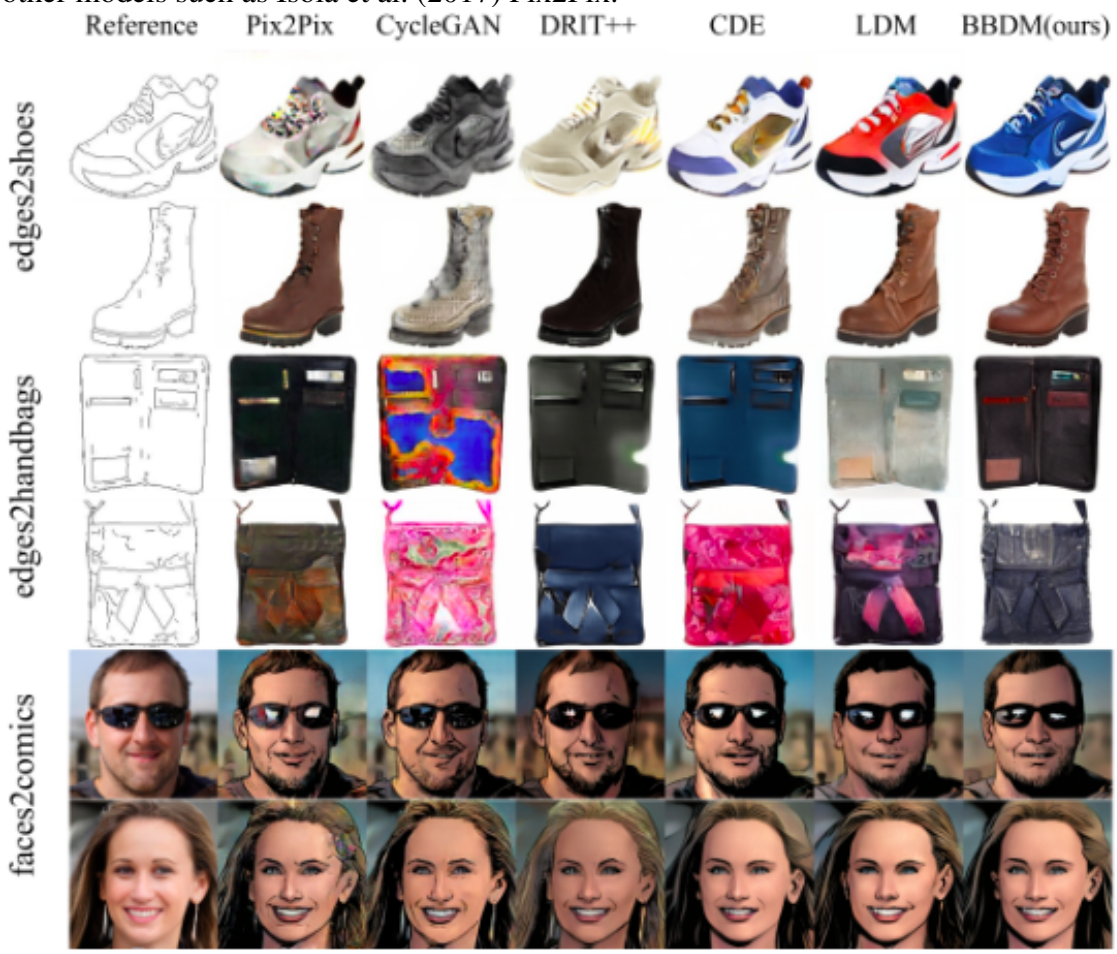


Source: Li et al. (2023)

3.2 In-silico labeling

Since the seminal work by Christiansen et al. (2018), the field of *in silico labeling*, has advanced on different fronts. The JUMP-Cell Painting Consortium (JUMP-Cell Painting Consortium, The Broad Institute, 2022) is a dataset of brightfield-to-fluorescence paired images that have become available and can be used to train *in silico labeling* techniques. Unlike the dataset used in Christiansen et al. (2018), JUMP uses Cell Painting Bray et al. (2016), a particular fluorescence protocol that aims to be a standard for image-based computer vision methods. This makes it a reliable resource for result comparison.

Figure 3.4: Comparison of the results of *image-to-image translation* using BBDM with other models such as Isola et al. (2017) Pix2Pix.



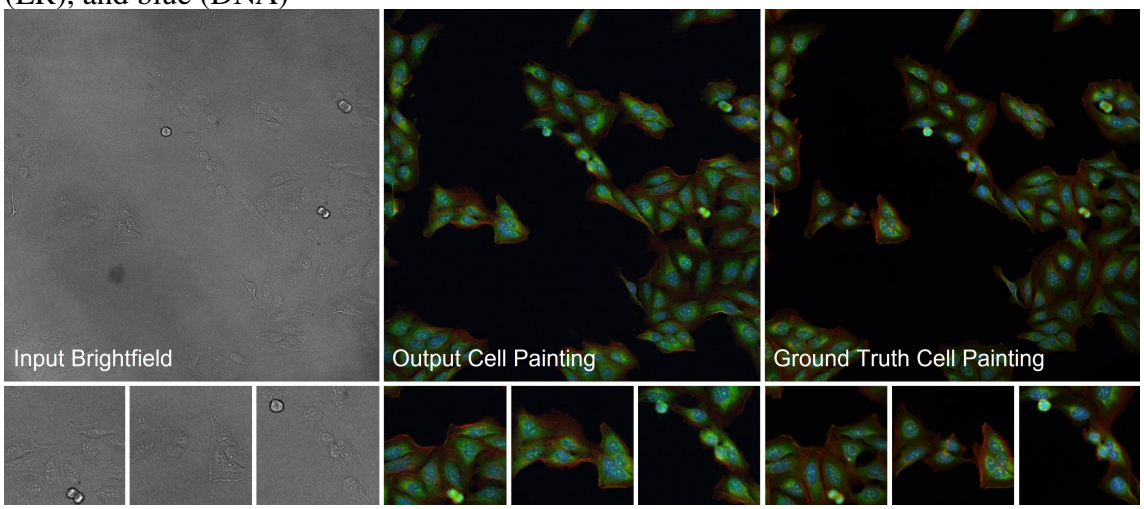
Source: Li et al. (2023)

Developing new generative image models offered an opportunity to improve the results of Christiansen et al. (2018). Moreover, the increasing interest and potential for adopting this method demanded more rigorous evaluation processes for the generated images. Wieslander et al. (2021) propose the use of GANs (GOODFELLOW et al., 2014), see Fig. 3.6, to generate the fluorescence images and also the Learning Using Privileged Information (LUPI) (VAPNIK; IZMAILOV et al., 2015) paradigm, which is a technique that uses information that would not be available for the model on inference when training – in this particular paper, the segmentation masks for the cells' nuclei. For the evaluation phase, a CellProfiller pipeline (CARPENTER et al., 2006) is used to extract features (profiling), such as morphology, intensity, and count, from the generated images and the original fluorescence images. The comparison is based on the extracted features rather than the raw image pixels. Three models were generated, one for each fluorescence channel: nuclei, lipids, and the last for the cytoplasm. Each model was tailored for the task

and only the model for predicting the lipid channel used conditional GAN. Nuclei and cytoplasm channels were predicted by models that were based on U-Net. Note that this study does not predict Cell Painting fluoresce.

Cross-Zamirski et al. (2022) proposed predicting the five channels of the Cell Painting protocol using a proprietary dataset. For this task, two models were used: the first one is based on a U-Net, and the second is built on the Wasserstein GAN (WGAN) paradigm with penalty gradient (ARJOVSKY; CHINTALA; BOTTOU, 2017). The evaluation used image-based similarity metrics and comparing features extracted via profiling. In their study, the novelty was the use of Uniform Manifold Approximation (UMAPs) (MCINNES; HEALY; MELVILLE, 2018), a method popular among biologists that provides a visual component to the analysis of clusters based on extracted features. In their study, UMAP was used to separate instances based on treatments. In all metrics, the WGAN model surpassed the U-Net.

Figure 3.5: An example of *state-of-the-art in-silico labeling*, here referred as *Cell Painting*. Three colored channels are exhibited in output and ground truth, red (AGP), green (ER), and blue (DNA)



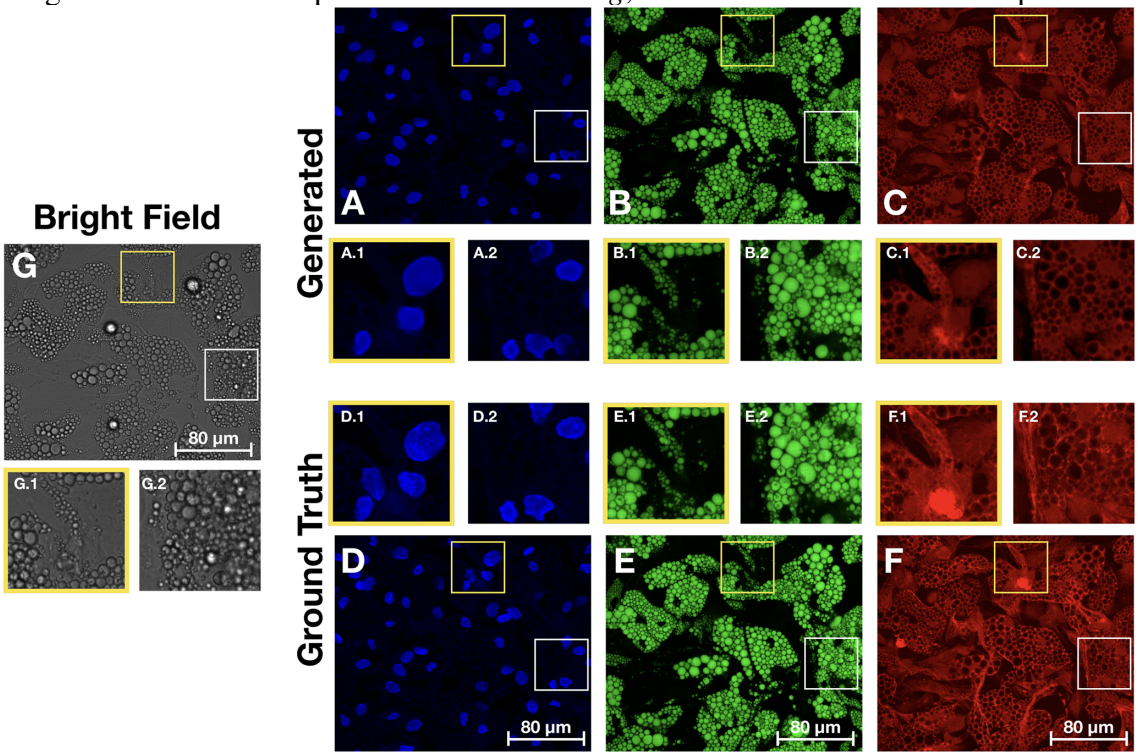
Source: Cross-Zamirski et al. (2023)

At last, Cross-Zamirski et al. (2023) still pushes to advance the usage of more complex methods. They used a subset of the JUMP-Target dataset (JUMP-Cell Painting Consortium, The Broad Institute, 2022) for predicting the five fluorescence Cell Painting channels. For this purpose, a variant of DDPM called *Class-Guided Diffusion Model* was used – *Class-Guided* means that extra information related to each instance is used to fit the probability distribution of the data better. In Cross-Zamirski et al. (2023), the extra information is the meta-data of each well[1], such as the type of cell or which chemical

[1]In a multi-well plate, a single well is a container which holds a sample inside the microscope

was used to produce perturbations to the sample. The authors found that the metadata did improve overall metrics, reaching *state of the art* results, which can be seen in Fig.3.5. One particular finding of this study is that when using the compound target as the label for class guidance, the model would produce more background noise, undermining the image level metrics but improving feature extraction metrics.

Figure 3.6: More examples of *in-silico labeling*, this time with the channels separated.



Source: Wieslander et al. (2021)

After understanding the relevance and potential of in-silico labeling to advancing biological research, especially in drug development and disease diagnosis, a review of consolidated and the latest works in in-silico labeling and image-to-image translation was provided in this chapter. This review showed that there are still opportunities to explore, especially regarding the advancement of image models. This is the motivation behind the experiments that will be described in the next chapter.

4 THE PROPOSED METHODOLOGY

This chapter presents the methodology used in this study. This includes the selection of a proper dataset, followed by why the selected methodologies were chosen and how each model was trained, along with the choices of the main parameters.

4.1 Dataset

The dataset was collected and made available for this work by University colleagues from the *Laboratório de Sinalização e Plasticidade Celular* (LabSinal), which is associated with the *Instituto de Biofísica* and the *Centro de Biotecnologia* from the *Universidade Federal do Rio Grande do Sul* (UFRGS). All images depict cells from the U87 line (brain cancer), and they are collected from 6 different wells of the same plate. Two of the wells are control samples, two wells are treated with *dasatinib*, and two wells are treated with *dasatinib* and *temozolomide*, with the concentration of the administered drug being also varied. The provided dataset has three channels: the phase, which is a brightfield image consisting of one z-plane image; the green channel labeled with **Green Fluorescent Protein-LC3** *(GFP-LC3)* and the red channel labeled with *mAple-53BP1*. Both labels are protein-based, where GFP and mAple are fused to the *LC3* and *53BP1* proteins, respectively, causing fluorescence when excited by light. *GFP-LC3* is indicative of cellular autophagy, while *mAple-53BP1* is typically involved in the DNA damage response. It is important to note that previous studies, such as (CROSS-ZAMIRSKI et al., 2022) and (CROSS-ZAMIRSKI et al., 2023), used three channels for the phase (brightfield) images, each on a different z-plane (focal plane). Due to computational limitations, the study centers on only two channels, phase and red.

The dataset consists of 4,536 images[1], divided into training (3,174 images), validation (454 images), and test (908 images) splits. All images were pre-processed before training the models. The pixel values of the images were normalized between -1 and 1, and its resolution was downscaled from its original 1408 by 1040 pixels to 512 by 512 pixels. The original dataset contained 4,608 images, but some images were out-of-focus or were out of bounds of the well. Hence, they were manually removed from the dataset.

[1]Number of images after defected ones were removed from the originally provided dataset

4.2 The tested models

Building on recent advances in image generation, three different models were selected for this work. Two of these models are diffusion-based and, although different from the diffusion model used in Cross-Zamirski et al. (2023), they are well-suited for the *image-to-image translation* task. In one model, the diffusion process occurs in pixel space, while the other leverages latent space representation. The third model is based on the well-established VAE-GAN architecture. This selection provides a comprehensive overview of recent advancements and potential future improvements.

4.2.1 Brownian Bridge Diffusion Models (BBDM)

Li et al. (2023) propose a new method for *image-to-image translation*. Rather than starting from random noise and conditioning each forward or denoising step with the image on the original domain, as done in *image-to-image translation* task using the original DDPM process, BBDM uses a stochastic Brownian bridge process to translate an image from its original domain to the new one without any conditioning, making it a bidirectional process that fits the task of *in-silico labeling*.

The code provided by the authors of the original paper was used to run the experiments[2], following the configurations of the *aligned dataset*. The network was trained from scratch on a single RTX A6000 GPU, with batch size four and a learning rate of 5×10^{-6}. The chosen optimizer and loss function were *Adam* and *Mean Squared Error*, respectively. The model was trained for 200 epochs, which took 52 hours. Each image is generated in 20 seconds. The number of diffusion steps was maintained from the original implementation: 1000 during training, and 200 during image generation.

4.2.2 Latent Brownian Bridge Diffusion Models (LBBDM)

LBBDM follows the same idea of using Brownian Bridges to make the diffusion process bidirectional as in the original BBDM. However, instead of occurring in the pixel space, the forward and denoising processes now takes place in the latent space.

For this, the methodology used in the original paper by Li et al. (2023) was fol-

[2] <https://github.com/xuekt98/BBDM>

lowed. In the original LDM paper (ROMBACH et al., 2022), the VQGAN architecture, (ESSER; ROMBACH; OMMER, 2021), is used to encode and decode the images. Although VQGAN is a consolidated architecture for this, we opted to use a VAE model, also adopted in later iterations of LDMs. For this purpose, the model is trained to reconstruct the original image, i.e. brightfield channel images are reconstructed as brightfield channel images, the same goes for red channel images.

The VAE was trained so the diffusion process could take place on the latent space where the 512 by 512 image is encoded to a $4 \times 64 \times 64$ latent representation. The model was trained on an RTX A6000 GPU, with a batch size of 16 and a learning rate of 1×10^{-4}. The chosen optimizer and loss function were *Adam* and *Mean Squared Error* (MSE), calculated on the latent representation, respectively. The model was trained for 100 epochs, which took 8 hours. Each image is generated in 3 seconds. The number of diffusion steps was again, maintained from the original implementation: 1000 during training, and 200 on image generation.

4.2.3 Variational Autoencoder Generative Adversarial Network

Combining Variational Autoencoders (VAEs) with Generative Adversarial Networks (GANs) takes advantage of the strengths of both models, enabling high-quality sample generation while maintaining a regularized latent space. Using the VAE architecture as the generator on a GAN model has already been explored in Rombach et al. (2022), where this architecture is responsible for transforming the image into the latent space for the diffusion process to be carried out. The original code[3] and model weights provided by the authors of Rombach et al. (2022) were used in this experiment. The code was adapted from its original task of image recreation to perform domain translation, after that, the original model weights were fine-tuned for our dataset and task. To minimize instability during the fine-tuning process, the discriminator component was introduced after 20 epochs, this ensured that the discriminator loss did not affect the generator loss during the initial epochs. The model was fine-tuned on a total of 60 epochs. The model was fine-tuned on two RTX A6000 GPU, with a batch size of 4 and a learning rate of 5×10^{-5}. *Adam* was the chosen optimizer. The loss is a combination of the generator loss, described in Section 2.2.1, and the discriminator loss, which used a pre-trained *VGG16*, (SIMONYAN; ZISSERMAN, 2014), to classify the generated and ground-truth

[3] <https://github.com/CompVis/stable-diffusion>

images as the real one or not. With the fine-tuning process complete, the model takes 5 seconds to generate each image, on a single RTX A6000 GPU.

Due to errors stalling the training process, it was not possible to accurately determine the duration it took to fine-tune this model. However, an analysis of the logs suggests that the fine-tuning process took 28 hours given that each of the 56 epochs took roughly 30 minutes.

4.3 Evaluation

Following the example of previous work on *in-silico labeling*, the idea is to evaluate the model performance on two fronts: using standard image-to-image translation metrics, which evaluate the generated images on a pixel level, and with methods to assess the biological information contained in the generated image.

4.3.1 Pixel-Level Metrics

As in the evaluation process done in previous works, various metrics are used to evaluate the generated images at the pixel level. They aim to compare how close or distant the generated is from the ground truth. The chosen metrics were:

Mean Average Error *(MAE)*, calculates the average absolute differences between the predicted and actual pixel values, measuring the average magnitude of errors without considering their direction. Similar to *MAE*, **Mean Squared Error** *MSE* measures the difference between the predicted and actual pixel values, but here the difference is squared, focusing on penalizing larger errors.

Structural Similarity Index *(SSIM)* evaluates the structural similarity between two images, considering luminance, contrast, and structure, and is often more aligned with human visual perception than pixel-wise measures, (WANG et al., 2004).

Fréchet Inception Distance *(FID)* measures the similarity, by computing its Euclidean distance, between the distributions of generated and real images by comparing feature vectors obtained from a pre-trained *Inception V3* model introduced by Szegedy et al. (2016), where lower values indicate higher similarity.

Pearson Correlation Coefficient *(PCC)* measures the linear correlation between the predicted and actual images, providing a value between -1 and 1, where 1 indicates a

perfect positive correlation.

Peak Signal-to-Noise Ratio *(PSNR)* quantifies the quality of the generated images compared to the originals, with higher values indicating better image quality and lower noise levels.

4.3.2 Biological Information

Although pixel-level metrics evaluate the visual quality of the synthesized images, the main goal of *in-silico* labeling is to extract biological information from microscopy images. Based on the characteristics of the dataset described in Section 4.1, we can categorize the samples into six groups based on the combination and concentration of drugs. Due to visible phenotype characterization on the red channel, we can asses how good the preservation of such phenotypes, i.e., biological information, are on the generated images.

Using the same train, validation, and test sets as in the *image-to-image translation* tasks previously described in Section 4.2, we trained an image classifier using the ground truth stained images (red channel). More precisely, we used a ResNet-50 backbone (HE et al., 2016) with weights pre-trained on ImageNet, and fine-tuned it with our dataset. The accuracy of the model on the test set provides an idea of how well the six categories can be separated using ground truth images.

To evaluate how well the virtually stained images maintain the biological characteristics of the six classes, we repeated the evaluation protocol for the images produced by the image generation models. The results can be compared to validate whether the biological information contained in the original images was preserved in the images generated using the in-silico labeling method.

5 EXPERIMENTAL RESULTS

This chapter presents the experimental results obtained from our study on image generation. The results are divided into two sections: quantitative and qualitative. After that, a discussion on the results and limitations of each model concludes the chapter.

5.1 Quantitative Analysis

Table 5.1 provides an overview and comparison of each model performance on the metrics mentioned in section 4.3.1. As observed, the VAE-GAN model outperformed all others across all metrics, followed by the LBBDM model, with the BBDM model ranking last. Notably, all metrics are correlated, especially when comparing image quality metrics to the classification metric. Additionally, it is important to remember that the FID metric is calculated by comparing feature vectors extracted from the generated and real images.

Table 5.1: Comparison of different models based on various metrics

Model	**SSIM** ↑	**FID** ↓	**PSNR**↑	**PCC** ↑	**MAE** ↓	**MSE** ↓	**Accuracy**[1] ↑
VAE-GAN	**0.978**	**0.014**	**34.22**	**0.719**	**0.011**	**0.001**	**62.33**
LBBDM	0.969	0.052	31.25	0.421	0.019	0.002	51.10
BBDM	0.963	1.296	25.06	0.448	0.077	0.011	16.41

Focusing on the comparison between the LBBDM and BBDM models, it is evident that performing the diffusion process in the latent space significantly improves the quality metrics of the generated images in the proposed dataset. This suggests that the information condensed in the latent representation remains highly representative of the original data while enhancing computational efficiency.

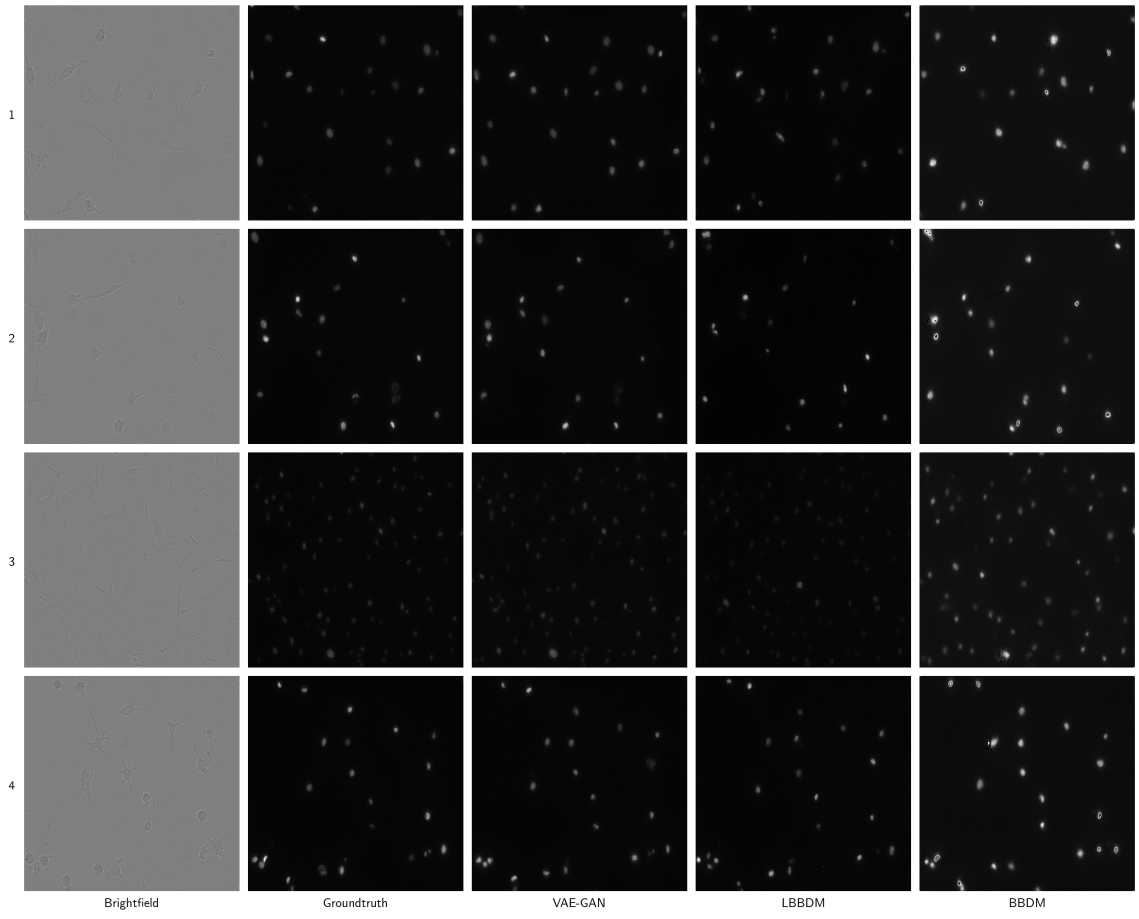
5.2 Qualitative Analysis

This section is focused on the qualitative results by presenting visual examples of the generated images, providing a comparative assessment of the visual fidelity and diversity achieved by the models. Fig. 5.1 presents some of the generated images using the test set (zoom in on the images to notice the details). In images generated by the BBDM model, a noticeable *locality* error is observed, with cells either being introduced or subtracted from the image. In tasks such as cell segmentation, this would result in sig-

[1]The result on the original images, which serves as a baseline for comparison, is 84.6%

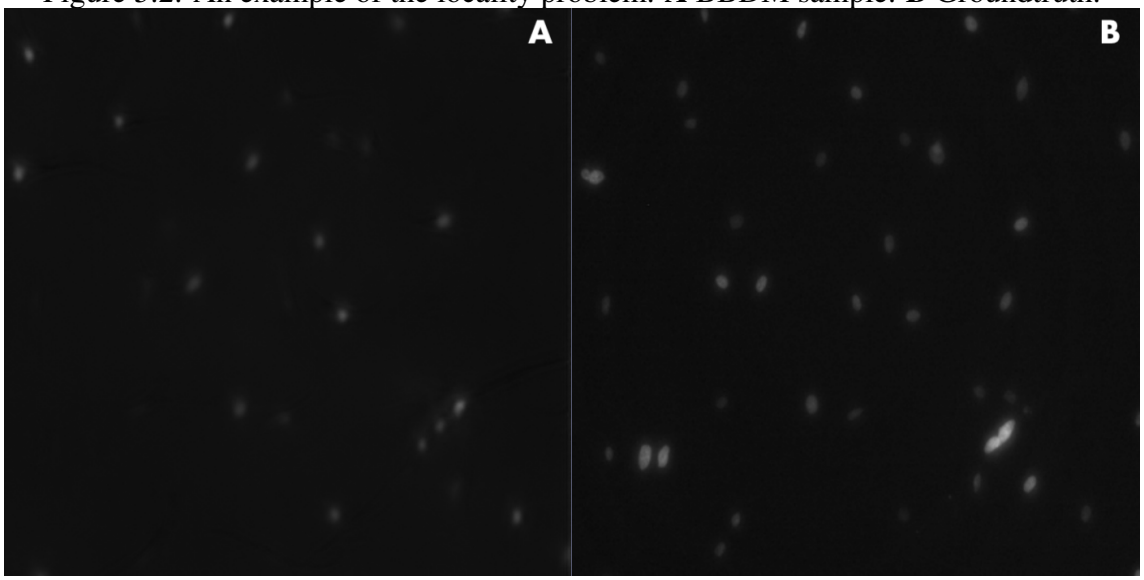
nificantly misleading analysis, characterizing the images generated by the BBDM model as low quality, an example can be seen in zoom-in in Fig 5.2. This *locality* issue also occurs in images generated by the LBBDM model, though less frequently. This can be observed in the images on the third row, which are characterized by a high number of cells. The VAE-GAN model, on the other hand, does not exhibit observable occurrences of the *locality* problem.

Figure 5.1: Example of samples generated by each model and their respective input image (brightfield) and ground truth.



Following the qualitative analyses, it is evident that the most common issue is the variation in fluorescence intensity within each cell, which is observable in images generated by all models. This variation poses a limitation since fluorescence intensity can be a key characteristic in identifying cell phenotypes.

Figure 5.2: An example of the locality problem. **A** BBDM sample. **B** Groundtruth.



5.3 Discussion

As we can see in Table 5.1, the VAE-GAN model has achieved the best results, both in metrics related to image reconstruction and in the classification task, which aims to evaluate the representation of biological information contained in the generated images. When evaluating the results of all models, we can observe that the metrics for image reconstruction and the classification task are correlated, indicating that indeed maintaining image similarity results in preserving its biological information. This is an important result since in a real laboratory situation, for example, drug research, the generated images will undergo further image analyses to extract phenotypes that are characterized by the drug concentration and variability that was used to label each class. The preservation of image similarity and structure can be qualitatively assessed by analyzing the generated images presented in Fig. 5.1. It is noticeable that the images generated by the VAE-GAN and the LBBDM models, both with similar quantitative results, are visually closer to the ground truth than those generated by the one with the worst quantitative results, the BBDM model.

It is important to emphasize that while the VAE-GAN model delivers the best results, it also demands the most computational power when training, even as a fine-tuned model rather than one trained from scratch, as described in 4.2.3. In contrast,

the LBBDM model requires significantly less computational power, despite necessitating the pre-training of a similar but smaller encoder-decoder model. When generating the samples, on a single RTX A6000 GPU, VAE-GAN takes 5 seconds per image, a slight edge over LBBDM, which takes 6 seconds per image, and, at last, the BBDM model which takes 20 seconds per image.

It is also important to note that when training with a larger U-Net architecture, responsible for the denoising process in LBBDM-based models, the model tends to overfit. This overfitting can likely be attributed to our dataset being smaller than those used in previous works. Notably, Cross-Zamirski et al. (2023), the only prior work that explored diffusion models, used 10 plates containing around $2,000$ images each, totaling roughly $20,000$ images, in contrast with a total of $4,536$ images used in our work. Additionally, in their dataset, the brightfield images consisted of three channels, representing three different z-focal planes, in our dataset the brightfield images consist of only one channel for a single z-focal plane, thus providing less information. Given these observations, the LBBDM model appears promising for future research.

6 CONCLUSION

This work introduced the reader to the importance and limitations of microscopy in biomedical research and how computer vision can enhance this field. In particular, we focused on the *in-silico labeling* application, which is a methodology for artificially labeling microscopy images without the drawbacks of fluorescent labels. Following this introduction, the key concepts of *image-to-image translation*, its applications, methodologies, and both established and promising models were reviewed. With this foundation, experiments using the exposed methodologies were proposed and executed. The results were quantitatively and qualitatively analyzed and discussed, suggesting that image quality and a classification method to analyze if biological information was preserved in the generated images were correlated, this could also be observed by qualitatively analyzing the generated images. Regarding model performance, although the VAE-GAN model achieved the best metrics and visual results, the LBBDM model seems more promising for future works, due to its lower computational requirements. The prominence of the LBBDM model extends when comparing it to the BBDM model, suggesting that performing the diffusion and denoising process is more efficient and results in higher-quality image generation when done on the latent space. The judgment of these results is meaningful, providing direction and insights for future research.

The research field of bioimaging is in rapid progress and the task of *in-silico labeling* is advancing with it. A key indicator of this progress is the inaugural *Light My Cell*[1] challenge, promoted by *French BioImaging* (FBI) at the *International Symposium on Biomedical Imaging* (ISBI) in 2024. This challenge introduced a new dataset comprising paired images from not just brightfield but three different microscopy modalities to four different fluorescent channels (labels). Moreover, the dataset is not fully annotated, aiming for a generalized model, a key obstacle in the application of computer vision in biomedical imaging. The knowledge acquired and presented in this work can hopefully serve as a foundation for new studies that can further advance the field.

[1] <https://lightmycells.grand-challenge.org/>

REFERENCES

ARJOVSKY, M.; BOTTOU, L. Towards principled methods for training generative adversarial networks. **arXiv preprint arXiv:1701.04862**, 2017.

ARJOVSKY, M.; CHINTALA, S.; BOTTOU, L. Wasserstein generative adversarial networks. In: PMLR. **International conference on machine learning**. [S.l.], 2017. p. 214–223.

BRAY, M.-A. et al. Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. **Nature protocols**, Nature Publishing Group UK London, v. 11, n. 9, p. 1757–1774, 2016.

CARPENTER, A. E. et al. Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. **Genome biology**, Springer, v. 7, p. 1–11, 2006.

CHRISTIANSEN, E. M. et al. In silico labeling: predicting fluorescent labels in unlabeled images. **Cell**, Elsevier, v. 173, n. 3, p. 792–803, 2018.

COHEN, J. P.; LUCK, M.; HONARI, S. Distribution matching losses can hallucinate features in medical image translation. In: SPRINGER. **Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I**. [S.l.], 2018. p. 529–536.

COLE, R. Live-cell imaging: The cell's perspective. **Cell adhesion & migration**, Taylor & Francis, v. 8, n. 5, p. 452–459, 2014.

CROSS-ZAMIRSKI, J. O. et al. Class-guided image-to-image diffusion: Cell painting from brightfield images with class labels. **arXiv preprint arXiv:2303.08863**, 2023.

CROSS-ZAMIRSKI, J. O. et al. Label-free prediction of cell painting from brightfield images. **Scientific reports**, Nature Publishing Group UK London, v. 12, n. 1, p. 10001, 2022.

DESHPANDE, A. et al. Learning diverse image colorization. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2017. p. 6837–6845.

ESSER, P.; ROMBACH, R.; OMMER, B. Taming transformers for high-resolution image synthesis. In: **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**. [S.l.: s.n.], 2021. p. 12873–12883.

GOODFELLOW, I. et al. Generative adversarial nets. **Advances in neural information processing systems**, v. 27, 2014.

GUPTA, A. et al. Is brightfield all you need for mechanism of action prediction? **bioRxiv**, Cold Spring Harbor Laboratory, p. 2022–10, 2022.

HARAGUCHI, T. Live cell imaging: approaches for studying protein dynamics in living cells. **Cell structure and function**, Japan Society for Cell Biology, v. 27, n. 5, p. 333–334, 2002.

HE, K. et al. Deep residual learning for image recognition. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2016. p. 770–778.

HINTON, G. E.; SALAKHUTDINOV, R. R. Reducing the dimensionality of data with neural networks. **science**, American Association for the Advancement of Science, v. 313, n. 5786, p. 504–507, 2006.

HO, J.; JAIN, A.; ABBEEL, P. Denoising diffusion probabilistic models. **Advances in neural information processing systems**, v. 33, p. 6840–6851, 2020.

ICHA, J. et al. Phototoxicity in live fluorescence microscopy, and how to avoid it. **BioEssays**, Wiley Online Library, v. 39, n. 8, p. 1700003, 2017.

ISOLA, P. et al. Image-to-image translation with conditional adversarial networks. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2017. p. 1125–1134.

JUMP-Cell Painting Consortium, The Broad Institute. **JUMP-Target**. 2022. <https://github.com/jump-cellpainting/JUMP-Target>.

KINGMA, D. P.; WELLING, M. Auto-encoding variational bayes. **arXiv preprint arXiv:1312.6114**, 2013.

LI, B. et al. Bbdm: Image-to-image translation with brownian bridge diffusion models. In: **Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition**. [S.l.: s.n.], 2023. p. 1952–1961.

LIU, Z.-S.; SIU, W.-C.; CHAN, Y.-L. Photo-realistic image super-resolution via variational autoencoders. **IEEE Transactions on Circuits and Systems for video Technology**, IEEE, v. 31, n. 4, p. 1351–1365, 2020.

MCINNES, L.; HEALY, J.; MELVILLE, J. Umap: Uniform manifold approximation and projection for dimension reduction. **arXiv preprint arXiv:1802.03426**, 2018.

PANG, Y. et al. Image-to-image translation: Methods and applications. **IEEE Transactions on Multimedia**, IEEE, v. 24, p. 3859–3881, 2021.

PLISSITI, M. E.; VRIGKAS, M.; NIKOU, C. Segmentation of cell clusters in pap smear images using intensity variation between superpixels. In: IEEE. **2015 International Conference on Systems, Signals and Image Processing (IWSSIP)**. [S.l.], 2015. p. 184–187.

PRINCE, S. J. **Understanding Deep Learning**. The MIT Press, 2023. Available from Internet: <http://udlbook.com>.

PYLVÄNÄINEN, J. W. et al. Live-cell imaging in the deep learning era. **Current Opinion in Cell Biology**, Elsevier, v. 85, p. 102271, 2023.

QIAO, Y. et al. Thresholding based on variance and intensity contrast. **Pattern Recognition**, Elsevier, v. 40, n. 2, p. 596–608, 2007.

ROMBACH, R. et al. High-resolution image synthesis with latent diffusion models. In: **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**. [S.l.: s.n.], 2022. p. 10684–10695.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: SPRINGER. **Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18**. [S.l.], 2015. p. 234–241.

SAHARIA, C. et al. Image super-resolution via iterative refinement. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 45, n. 4, p. 4713–4726, 2022.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. **arXiv preprint arXiv:1409.1556**, 2014.

SZEGEDY, C. et al. Going deeper with convolutions. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2015. p. 1–9.

SZEGEDY, C. et al. Rethinking the inception architecture for computer vision. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2016. p. 2818–2826.

VAPNIK, V.; IZMAILOV, R. et al. Learning using privileged information: similarity control and knowledge transfer. **J. Mach. Learn. Res.**, v. 16, n. 1, p. 2023–2049, 2015.

WANG, Z. et al. Image quality assessment: from error visibility to structural similarity. **IEEE transactions on image processing**, IEEE, v. 13, n. 4, p. 600–612, 2004.

WIESLANDER, H. et al. Learning to see colors: Biologically relevant virtual staining for adipocyte cell images. **Plos one**, Public Library of Science San Francisco, CA USA, v. 16, n. 10, p. e0258546, 2021.

ZHU, J.-Y. et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: **Proceedings of the IEEE international conference on computer vision**. [S.l.: s.n.], 2017. p. 2223–2232.