Deep learning solutions for domain-specific image segmentation

Vadineanu, S.

Chapter 1

Introduction

Image segmentation is a computer vision task where an image is partitioned into multiple segments or regions. The goal is to assign each pixel in the image to a specific object or region, enabling the categorization of different parts of the image, such as distinguishing objects from the background. This can take the form of finding tumours in medical images [9], identifying traffic signs in self-driving cars [78], or environmental monitoring [75]. Initially, this operation would be performed manually by delineating the border of each segment within the image. However, given the tediousness of this process, there has been a significant push for automating it, with thresholding [76], i.e., the categorization of pixels based on their intensity values, being a first step in this direction (see Figure 1.1 for an overview of the main developments in segmentation techniques). Thereafter came methods such as region-based segmentation [164] and clustering [30] which based the categorization of one pixel on the intensities of the neighbouring pixels. With the advent of machine learning, better performance was achieved by hand-crafting image features and learning the segmentation from the data and its associated ground truth [162, 96, 134], commonly referred to as annotations. Currently, deep learning solutions based on convolutional neural networks [106] produce state-of-the-art results [131, 125, 28] in image segmentation, relinquishing the need for hand-crafted features. However, despite their impressive performance, machine learning, and especially deep learning, techniques require vast amounts of annotated images which are often created manually, making the annotation process a persistent bottleneck [8].

Deep learning is a subset of machine learning that uses artificial neural networks to learn patterns from the data. A network consists of multiple layers of neurons, where

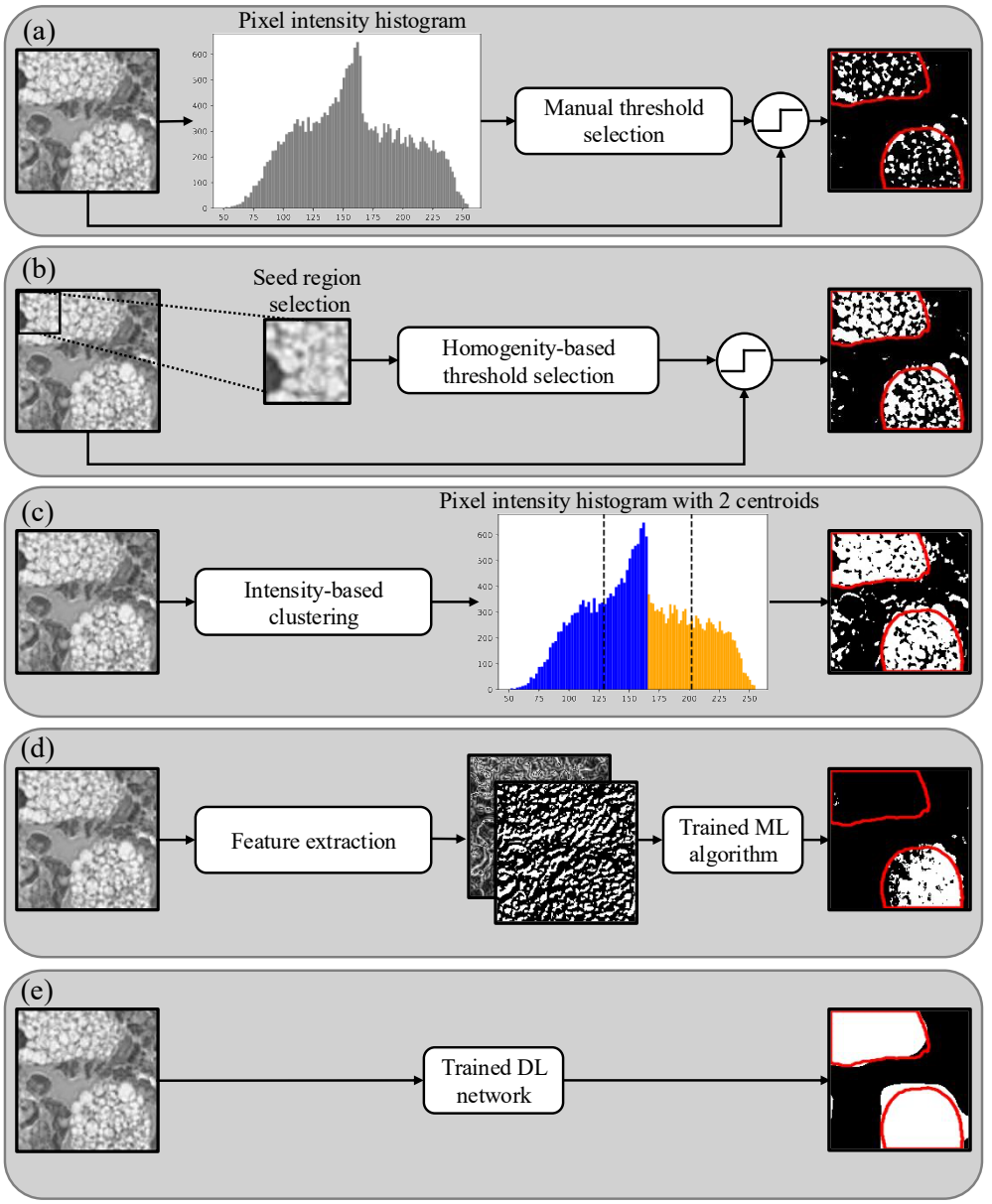

Figure 1.1: The evolution of image segmentation techniques showcased on the segmentation of a cell image. Figure (**a**) shows a segmentation pipeline for manual thresholding, Figure (**b**) corresponds to region-based thresholding, Figure (**c**) to clustering-based methods, while Figures (**d**) and (**e**) illustrate predictions of supervised machine learning and deep learning models, respectively. The red contour corresponds to the ground truth annotation. The input cell image is a crop from Lizard nuclear segmentation data set [59].

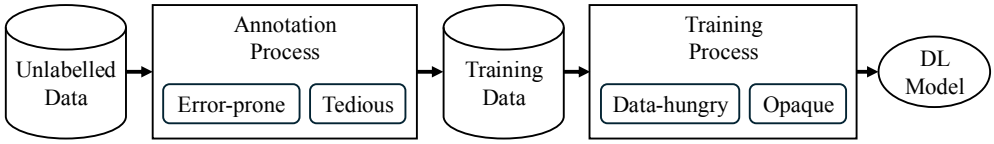

Figure 1.2: Deep learning pipeline from data to trained model and its associated challenges for the human operator.

each layer processes the input data and extracts increasingly complex features. Early layers capture general features such as edges in images, while deeper layers recognize more abstract patterns, for instance, objects. During training, the network requires various examples of input data and ground truth, adjusting its internal parameters to minimize errors. In this way, the network improves its ability to make accurate predictions. In the context of image segmentation, the input data requirement involves the procurement of large collections of images, whereas the annotations consist of pixel-level labels created for every image. A typical supervised deep learning pipeline from the data to the trained model, illustrated in Figure 1.2, consists of an annotation process and a training process.

For segmentation tasks relying on general knowledge, for instance, scene segmentation for self-driving cars [178], the annotation requirement is largely surmounted by the large number of available data sets and by the relative ease with which new data can be annotated, e.g., via crowdsourcing [35]. However, this requirement becomes significantly more demanding in scientific domains. Here, annotations must be created or verified by trained domain experts who are in limited supply. As a result, the adoption of deep learning in these specialized fields progresses more slowly than in general-knowledge domains. To overcome this, it becomes necessary to develop solutions that reduce the burden placed on expert annotators. Considering the deep learning pipeline from Figure 1.2, such solutions can target the annotation process, the training process, or both. On the annotation side, innovative methods are needed to increase the volume of data that can be annotated within a fixed time frame, while preserving their quality. Concurrently, on the training side, there is a demand for networks with transparent learning processes that require less annotated data while still delivering competitive results.

This thesis represents a collection of solutions focusing on streamlining the annotation and training processes for segmentation tasks in two scientific domains with expensive annotation processes, namely cellular imagery and archaeological remote sensing. We tailor methods leveraging the particularities of each domain to increase

the number of available annotations and to train networks at reduced annotation costs. In this chapter, we first describe in Section 1.1 the interplay between the versatility of deep learning solutions and the necessity of expert knowledge, we then provide an overview of the fields of cell imaging and archaeological remote sensing, presented in Section 1.2 and Section 1.3, respectively. In Section 1.4, we then introduce the fundamentals of deep learning methods for computational imaging, with an emphasis on segmentation. Lastly, in Section 1.5, we present the research questions that shape the scope of this thesis.

1.1 Learning Methods and Expert Knowledge

The past two decades have seen a paradigm shift in computational problem-solving, particularly in fields such as image segmentation. Traditionally, each application domain required the algorithms to be tailored specifically to the unique characteristics and challenges of the problem at hand. For instance, segmentation algorithms for medical imaging [5] would differ from those used in autonomous vehicles [50] or environmental monitoring [19]. These domain-specific solutions often demanded significant manual effort and expertise, limiting their adaptability to different fields.

The advent of data-driven techniques, particularly deep neural networks (DNNs), has significantly changed this landscape. Unlike traditional methods, DNNs provide a generic framework for learning features directly from data, without the need for handcrafted rules. This capability allows researchers to develop solutions that are broadly applicable, with minimal customization for specific domains. However, this flexibility comes with a caveat: the success of these methods depends on collaboration between computer scientists and domain experts. Domain experts play an important role in defining the key problems to be solved, curating high-quality data, and interpreting the results generated by the neural networks.

This interplay between generic computational frameworks and domain-specific expertise is exemplified by initiatives such as the Society, Artificial Intelligence and Life Sciences (SAILS) of Leiden University. SAILS is a university-wide interdisciplinary program aiming to disseminate the usage of artificial intelligence throughout the various disciplines within Leiden University. Its projects bring together data, algorithms, and domain experts in collaborative research efforts. This thesis is part of the SAILS initiative, leveraging its multidisciplinary framework to address challenges in image segmentation for cellular imaging and archaeological remote sensing.

1.2 Cell Imaging

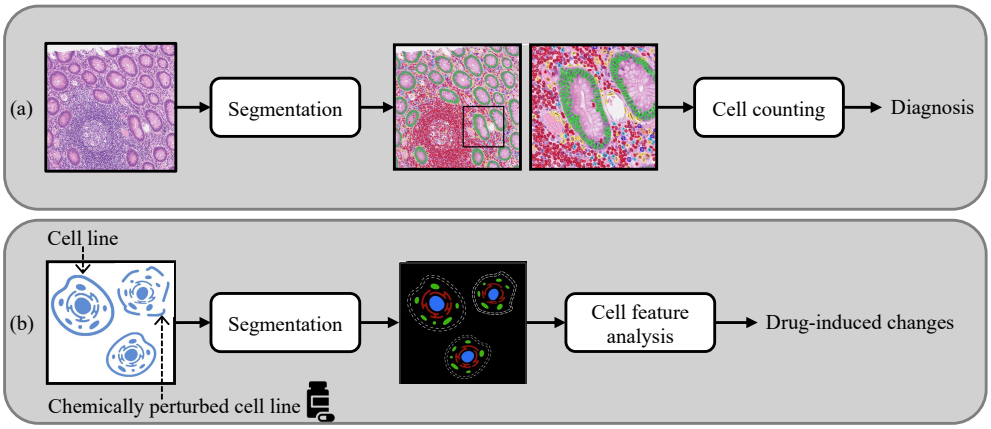

Figure 1.3: Applications of cell segmentation in biomedical research. Figure (a) illustrates the use of segmentation for cell counting, an important step in health diagnostics. Figure (b) demonstrates its application in drug discovery, where segmented cell structures are analysed to assess drug-induced changes. The cell boundary figure and crop in (a) are adapted from [59], and the cell shapes and their segmentations in (b) are from [86].

Cell imaging comprises a set of techniques that enable the visualization and analysis of cellular structures and their dynamics. By monitoring cells' behaviour over prolonged periods of time, researchers can understand how cells react to changes in the local environment or how they respond to various stimuli. This capability is important for advancing efforts in understanding disease pathology and in drug discovery [155] (see Figure 1.3 for an illustrative example). For instance, by tracking the complete blood count from a patient's sample, i.e., counting the white cells, red cells and platelets, the doctors can assess the overall health of that patient [143]. In addition, by tracking the changes appearing in targeted cells, experts can assess the effectiveness of new drugs [86]. For such tasks, the researchers require a good delineation of the cell structures of interest, i.e., cell segmentation, whose manual completion, however, implies tedious work from medical experts. Hence, oftentimes deep learning is perceived as a viable alternative [52]. Given the high annotation demands of deep learning and the challenges in obtaining these annotations—particularly for cell segmentation, where each cell must be individually identified and its shape precisely outlined—many off-the-shelf algorithms, which perform well in domains with abundant annotated data, are not directly applicable in this context.

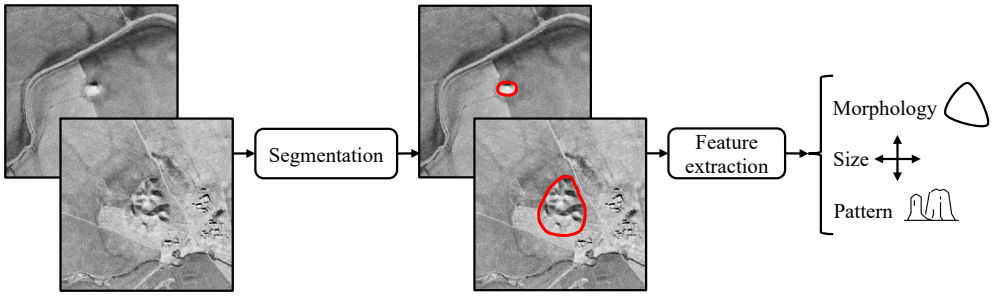

Figure 1.4: The segmentation of archaeological sites as a first step towards a more detailed characterization of the sites within a given area. The aerial images are obtained from CORONA Atlas & Referencing System [26].

1.3 Archaeological Remote Sensing

Archaeological remote sensing is a suite of non-invasive techniques used to detect, map, and analyse archaeological sites and features from a distance, without disturbing the ground. These techniques involve the detection of physical and chemical properties of the Earth's surface, which can indicate the presence of archaeological materials or features such as buried settlements, roads or changes in vegetation patterns caused by human activity [20]. One such task is the identification of settlements from aerial or satellite imagery, which involves the careful analysis of the surrounding landscape before assessing whether a certain image feature represents a settlement. By analysing the distribution patterns of these settlements as well as the local variations in their morphology, the archaeological researchers can gain insights into "the emergence, development, and organization of the first complex human societies." [104]. Here, similarly to cells, the segmentation of the sites can help in extracting morphological patterns which can then be further used to categorize the sites (see Figure 1.4 for an example). Many times, especially in the same geographical area, these settlements share similar visual characteristics, but their widespread distribution makes their manual identification tedious. Here too deep learning can bring considerable advantages by performing the detection of sites in a (semi-)automatic way. Although crowdsourcing is increasingly being used in annotating large archaeological data sets, the involvement of non-experts often leads to issues with data quality [110]. Thus, similar limitations to cell imaging apply here as well with the addition that archaeological research typically receives less funding than the research of medical sciences [151].

1.4 Deep Learning

1.4.1 Machine Learning for Imaging Tasks

Within a machine learning pipeline, an image is typically represented as a real-valued array of pixels $x \in \mathcal{X} \subset \mathbb{R}^{N \times M \times C}$, where $\mathcal{X}$ is the input space, N and M represent the number of rows and columns, respectively, and C corresponds to the number of channels. Coloured images have $C = 3$, corresponding to the red, green and blue channels, whereas grayscale images contain only one channel. The goal is to find the mapping $f : \mathcal{X} \to \mathcal{Y}$ from the input space $\mathcal{X}$ to the output space $\mathcal{Y}$. Machine learning algorithms approximate this mapping with a parametrized function f_δ, whose parameters δ are learned from the set of training images $X \subset \mathcal{X}$ paired with the corresponding expected output $Y \subset \mathcal{Y}$. Unlike the input space $\mathcal{X}$, whose shape is generally fixed, the shape of the output varies depending on the imaging task. For instance, for image classification, where the goal is to categorize images in k different classes, the output space becomes $\mathcal{Y} = \mathbb{R}^k$. Here, the model learns to predict a vector of k probabilities with the highest one providing the predicted class. When it comes to image segmentation with k classes (segments), $\mathcal{Y} = \mathbb{R}^{N \times M \times k}$. Similarly to classification, a probability vector is produced with probabilities for each of the k classes. However, in this case, a classification vector is generated for each pixel of the input image.

1.4.2 Convolutional Neural Networks

Convolutional neural networks (CNNs) have experienced widespread adoption in recent years, owing to their state-of-the-art performance and remarkable versatility across various imaging tasks [91, 93]. This type of neural network is comprised of filters organized in layers which are applied to the input image generally in a sequential manner, i.e., the output of a set of filters (layer) becomes the input to the next. Each layer i creates an intermediate image $z_i \in \mathbb{R}^{N_i \times M_i \times C_i}$, called a feature map, that reflects the importance of certain image features. The shape of z_i can change depending on the desired spatial dimensions and the number of filters (C_i) that the layer contains. The feature map z_i^j from filter j of layer i is obtained as

$$z_i^j = \sigma\left(\sum_{l=0}^{c_{i-1}} (z_{i-1}^l \circledast h_{ij}^l) + b_{ij}\right), \tag{1.1}$$

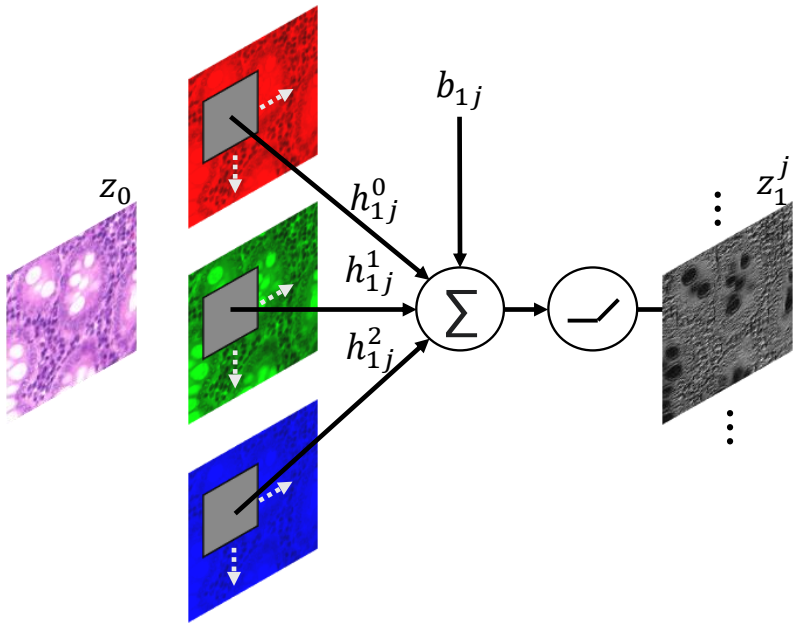

Figure 1.5: Example of how a feature map is obtained for a filter j in the first layer of a CNN with a coloured image as input. Each of the three channels of the input is convolved with a corresponding set of weights of the filter, the results are summed, a bias term is added and a non-linear function is applied to the summation result. This process is repeated for every filter of the layer. The input cell image is a crop from Lizard nuclear segmentation data set [59].

where each channel of the previous feature map z^l_{i-1} is convolved with a corresponding set of weights of the filter $h^l_{ij} \in \delta$ and the results are summed across all channels of the previous layer. After adding the bias term $b_{ij} \in \delta$, an activation function σ, oftentimes the rectified linear unit (ReLU) [2], is applied to introduce non-linearity, allowing the network to capture more complex patterns. A schematic representation of this process is presented in Figure 1.5.

In a supervised setup, in order to estimate the CNN parameters $\widehat{\delta}$, a training step is performed wherein the CNN's predictions of a set of input images, called the training set X, are compared against the ground truth output Y via a loss function $L : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. The goal is to find the optimal parameters

$$\widehat{\delta} = \underset{\delta}{\operatorname{argmin}} \sum_{(x,y)\in X\times Y} L(f_\delta(x), y) \tag{1.2}$$

that minimize the loss. Due to the high number of parameters and the non-linearity

introduced by the activation functions, analytical solutions are difficult to apply to CNNs. Hence, the optimization step usually involves an optimization algorithm, such as ADAM [82], which iteratively adjusts the parameters by using the partial derivatives of L with respect to them. Given that the training step is performed on a finite, and often small, subset of $\mathcal{X}$, a good fit on the training data does not guarantee the same performance on unseen images. Thus, to avoid such discrepancy, referred to as overfitting, a separate set, called validation set, can be employed. In this way, the optimization is still performed on the training set, while the performance of the CNN will be evaluated on the validation set, with the parameters being updated only if this results in a lower error on the validation set.

1.4.3 Challenges for the Human Operator

Although, when successfully trained, deep learning networks offer substantial benefits in automating image segmentation, obtaining these advantages comes with challenges for the human operators both during the annotation and training processes. These challenges are visually represented in Figure 1.2.

Firstly, not only is the large-scale annotation of images for segmentation a time-consuming task but it can also result in inconsistencies being introduced due to, for instance, fatigue, low image quality, or the ambiguity of the segmented structures. In many specialized domains, there can be a lack of consensus between experts when it comes to defining the category of an object, e.g., whether it is an archaeological site or a hill. Moreover, disagreements can also appear when defining the boundary of structures, for instance, when delineating cell nuclei from the surrounding cytoplasm. Thus, the effort required in annotating a data set and the potential errors it may contain pose significant apriori challenges to the deployment of deep learning by the human operator.

Secondly, training deep learning networks is a process that typically is susceptible to overfitting which can be caused by insufficient training data. Thus, an already annotated set of images does not guarantee a successful training process. Additionally, even after being successfully trained, the adoption of deep learning is challenged also by the opacity of its decision process. Due to the large number of parameters (generally, in the order of millions) involved in generating the output of deep learning networks, the steps undergone for producing it cannot be traced in a way comprehensible to humans. This can hinder trust and limit the integration of deep learning in less technical fields such as archaeology.

1.5 Research Questions

In this thesis, we address these challenges by focusing on segmentation tasks where the integration of standard deep-learning networks is suboptimal due to the limited availability of annotated samples. We propose solutions targeting both the annotation and the training processes with applications in cellular imaging and archaeological remote sensing.

An overview of the chapters of this thesis together with the research questions that they answer is presented below.

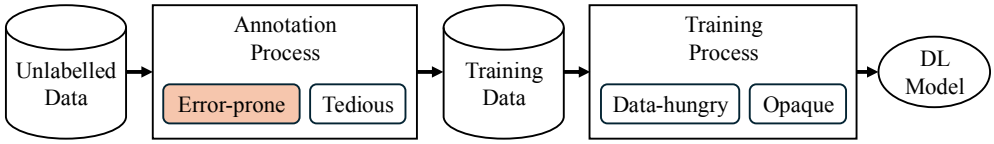

Figure 1.6: Our first research question examines the consequences of errors in the annotation process.

Research question 1. *How do different types of annotation errors impact the performance and robustness of deep learning models in the task of cell segmentation?*

It is commonly assumed that, when training deep learning networks, errors in annotations can severely degrade their performance, given the sensitivity of these networks to input quality (emphasized in orange in the deep learning pipeline in Figure 1.6). The annotations for cell segmentation are susceptible to errors which can be difficult to find and correct. Thus, understanding the specific impact of various annotation errors would allow us to develop more robust deep-learning networks or decide which inconsistencies require the most care to be prevented.

In Chapter 2, we analyse the impact of annotation inconsistencies on deep-learning-based cell segmentation. We introduce perturbations (see Figure 1.7) to emulate errors typical to the annotation of cells and we measure how they affect the performance of different network architectures designed for segmentation.

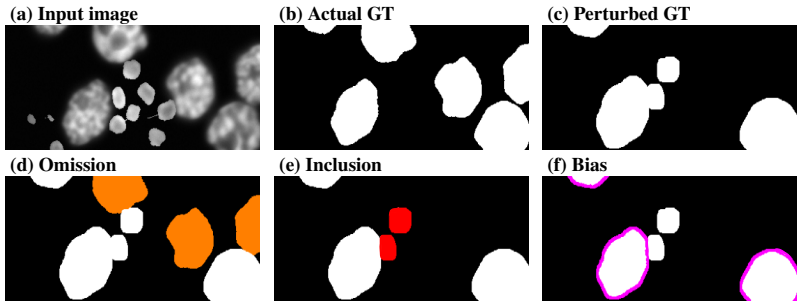

Figure 1.7: Example of the perturbations we perform. Figures (**a**, **b**) were generated with a virtual microscope from [166].

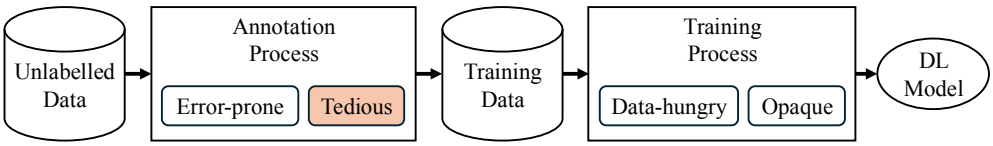

Figure 1.8: Our second research question focuses on reducing the human effort required for producing annotations.

Research question 2. *To what extent can we reduce the human effort required for cell segmentation annotation by using a convolutional neural network to improve lower-quality annotations?*

Given the labour-intensive nature of cell segmentation, reducing manual annotation effort is crucial for expanding deep-learning applications in biomedical fields (emphasized in orange in the deep learning pipeline in Figure 1.8). One promising approach is to relax the strict quality standards traditionally applied to annotations, thereby enabling a greater volume of annotated samples within a fixed time frame. However, these lower-quality annotations may not be immediately suitable for direct training of segmentation networks and may require refinement to be fully effective during training.

In Chapter 3, we propose a solution to enhance the annotation process by reducing the human effort in training deep learning algorithms. We achieve this by automatically enhancing the quality of noisy annotations, produced with low effort. We propose a learning pipeline in which a CNN is trained to upgrade low-quality annotations. For

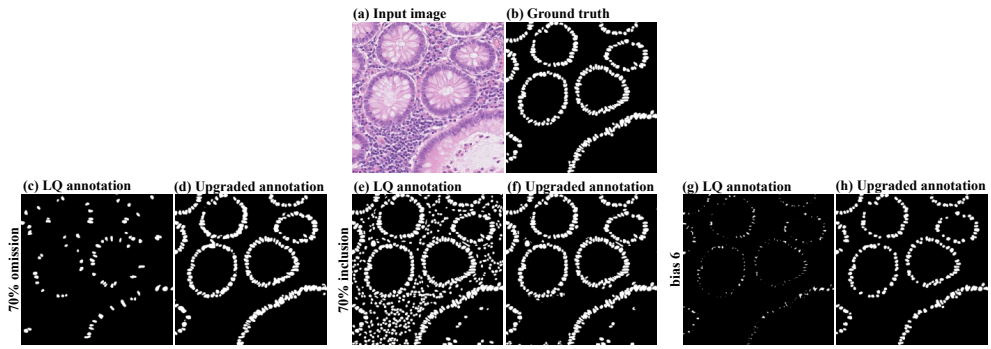

Figure 1.9: Example of perturbed annotations and their corresponding upgraded versions. Figures (**a, b**) are from [59].

this, we employ a small set of well-annotated samples whose annotations we perturb, similarly to Chapter 2, such that the CNN can learn a mapping from different versions of low-quality annotations to high-quality ones. We then use the initial set with high-quality annotations together with the upgraded noisy annotations to train segmentation networks on this larger combined set. In Figure 1.9, we show different types of perturbations applied to annotations together with the results after applying the upgrading CNN.

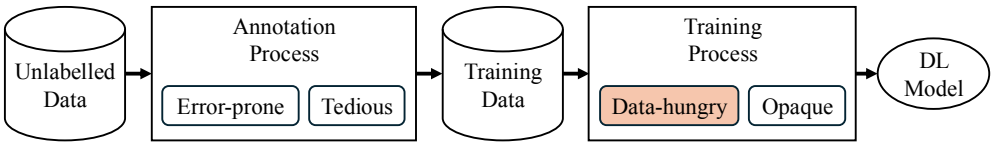

Figure 1.10: Our third research question is aimed at reducing the data requirements for training deep learning models.

Research question 3. *To what extent can the few-shot learning paradigm be effectively applied to cell microscopy image segmentation?*

An alternative approach to reducing the human effort in applying deep learning to specialized domains is to design algorithms that require less data than traditional CNNs (emphasized in orange in the deep learning pipeline in Figure 1.10). Few-shot learning, which aims to perform tasks with minimal labelled data, shows promise in addressing data scarcity in cell microscopy. However, its applicability to cell segmentation is uncertain due to particularities inherent to cell imagery compared to natural images where few-shot learning is more commonly applied.

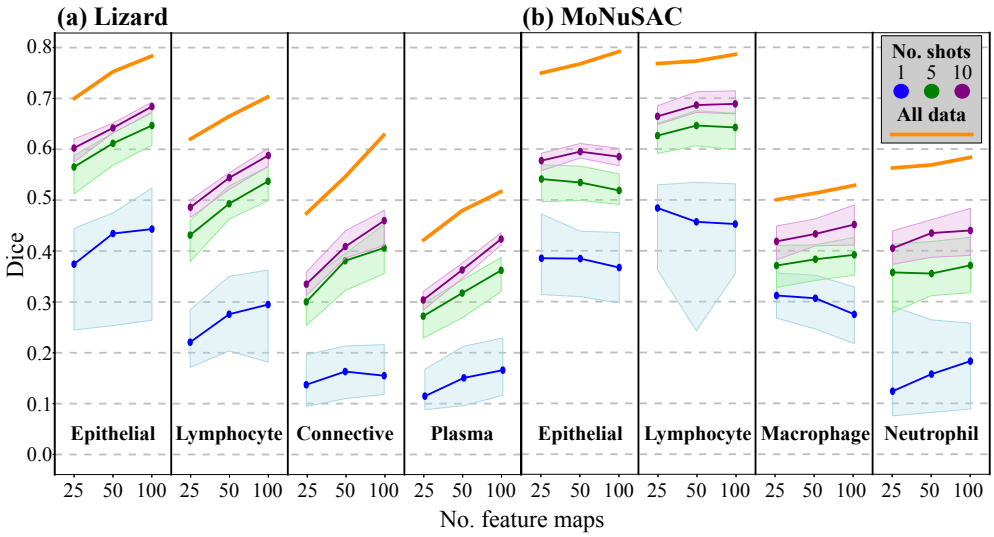

Figure 1.11: The quality of the segmentation against the number of shots and the number of features used. The orange line shows the average Dice score on the test set of the models trained on all labelled data. The points represent the median value across 50 experiments, while the shaded area is defined by the first and third quartiles.

In Chapter 4, we target the training process by proposing a new few-shot technique specifically tailored for cell segmentation. We leverage existing annotations for certain classes of cells to train feature extractor networks, which we then use to segment new cell classes using low amounts ($\leq$ 10) of annotations of the new class, called shots. While designing our few-shot algorithm, we consider requirements specific to cell segmentation such as the need for precise delineation between the cell instances and the relative similarity between the structures that are present in cell images. We study the performance of our algorithm on two data sets, each with four cell types, with the results shown in Figure 1.11. The graph shows the segmentation performance against the complexity of the feature extractors, with promising results for as little as 5 annotated images.

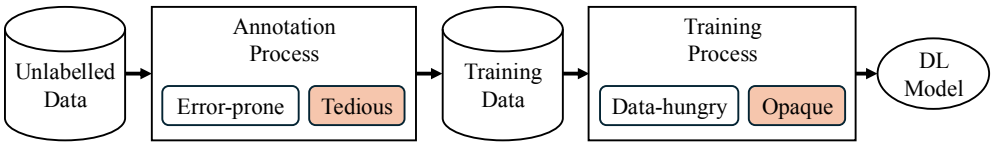

Figure 1.12: The final two research questions respectively tackle the tediousness of the annotation process and the opacity involved in training deep learning networks.

Research question 4A. *What insights into the learning process of different network architectures can be obtained from analysing activation maps in the context of archaeological site classification in Upper Mesopotamia?*

Research question 4B. *To what extent can activation maps be used as sources of annotation for site segmentation?*

In addition to the difficulty of creating annotated data sets, another challenge associated with deep learning in scientific domains is the relative reluctance with which these algorithms are perceived (emphasized in orange in the deep learning pipeline in

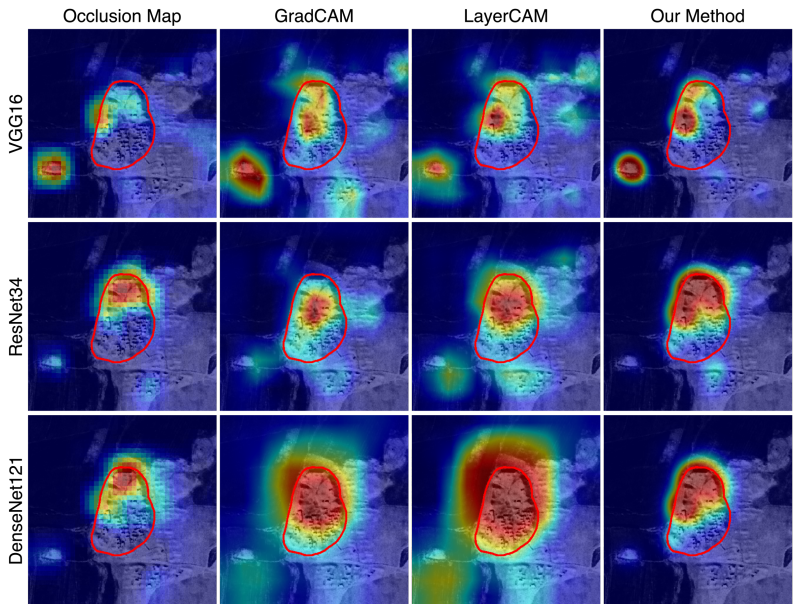

Figure 1.13: The activation maps of 3 CNNs (rows) given by 3 explainability techniques and our method (columns). The red line shows the expert delineation of the site. The site image is from [26].

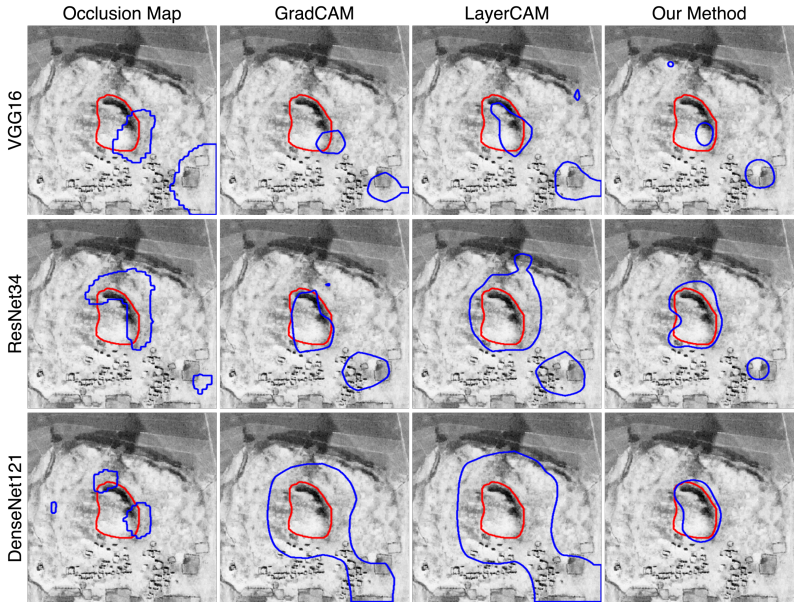

Figure 1.14: Site boundaries derived from the output of the explainability techniques (in blue) compared to the expert annotations (in red). The site image is from [26].

Figure 1.12). Due to their lack of traceability, deep learning predictions may not be trusted, particularly when the operator is unfamiliar with the neural network architecture and its operating principles. This reluctance is especially present in fields such as archaeology, which, as part of the non-exact sciences, relies heavily on interpretation and contextual understanding.

In Chapter 5, we tackle in tandem both problems associated with the usage of deep learning in archaeological research: the difficulty in explaining the results a CNN produces and the high cost of creating annotations for segmentation tasks. We explore how explainability techniques can enhance model interpretability and reduce annotation costs by applying these methods to three deep-learning architectures. We train the networks to classify whether an archaeological site is present in an image, a task for which the annotations are relatively cheap to produce. We then employ the activation maps, i.e., the output of the explainability techniques shown in Figure 1.13, both as sources of annotations (see Figure 1.14) for site segmentation and to analyze the visual cues that influence the networks' predictions. In addition, we develop a new method for creating activation maps specifically designed to produce accurate boundaries for this type of site.