2023

Hard-Hearted Scrolls: A Noninvasive Method for Reading the Herculaneum Papyri

Stephen Parsons
University of Kentucky, stephen@srparsons.com
Author ORCID Identifier:
https://orcid.org/0000-0002-0428-9732
Digital Object Identifier: https://doi.org/10.13023/etd.2023.372

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Stephen Parsons, Student

Dr. W. Brent Seales, Major Professor

Dr. Simone Silvestri, Director of Graduate Studies

HARD-HEARTED SCROLLS: A NONINVASIVE METHOD FOR READING THE HERCULANEUM PAPYRI

DISSERTATION

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy in the
College of Engineering
at the University of Kentucky

By
Stephen Parsons
Lexington, Kentucky
Director: Dr. W. Brent Seales, Professor of Computer Science
Lexington, Kentucky
2023

https://orcid.org/0000-0002-0428-9732

ABSTRACT OF DISSERTATION

HARD-HEARTED SCROLLS: A NONINVASIVE METHOD FOR READING THE HERCULANEUM PAPYRI

The Herculaneum scrolls were buried and carbonized by the eruption of Mount Vesuvius in A.D. 79 and represent the only classical library discovered in situ. Charred by the heat of the eruption, the scrolls are extremely fragile. Since their discovery two centuries ago, some scrolls have been physically opened, leading to some textual recovery but also widespread damage. Many other scrolls remain in rolled form, with unknown contents. More recently, various noninvasive methods have been attempted to reveal the hidden contents of these scrolls using advanced imaging. Unfortunately, their complex internal structure and lack of clear ink contrast has prevented these efforts from successfully revealing their contents. This work presents a machine learning-based method to reveal the hidden contents of the Herculaneum scrolls, trained using a novel geometric framework linking 3D X-ray CT images with 2D surface imagery of scroll fragments. The method is verified against known ground truth using scroll fragments with exposed text. Some results are also presented of hidden characters revealed using this method, the first to be revealed noninvasively from this collection. Extensions to the method, generalizing the machine learning component to other multimodal transformations, are presented. These are capable not only of revealing the hidden ink, but also of generating rendered images of scroll interiors as if they were photographed in color prior to their damage two thousand years ago. The application of these methods to other domains is discussed, and an additional chapter discusses the Vesuvius Challenge, a $1,000,000+ open research contest based on the dataset built as a part of this work.

KEYWORDS: Computer Vision, Machine Learning, Heritage Science, Micro-Computed Tomography, Volumetric Imaging, Virtual Unwrapping

Stephen Parsons

08/03/2023
Date

HARD-HEARTED SCROLLS: A NONINVASIVE METHOD FOR READING THE HERCULANEUM PAPYRI

By
Stephen Parsons

Dr. W. Brent Seales
Director of Dissertation

Dr. Simone Silvestri
Director of Graduate Studies

08/03/2023
Date

Hey Dad,

Check this out.

ACKNOWLEDGEMENTS

I have been blessed with such wonderful and supportive people in my life and work that I can only describe the situation as an embarrassment of riches.

My advisor Brent Seales has given me space to figure myself out, enriched my life with innumerable once-in-a-lifetime experiences, and believed in me fully, something I am sure I made easy by being stubborn, changing my mind often, and procrastinating consistently. In my technical work I have stood on very many shoulders, but none as tall as those of Seth Parker, who's an even taller friend. And to Christy Chapman, to whom I am grateful for many things (for example enabling all of the work I get to do), thank you most of all for looking out for me.

Thank you also to Ankan Bhattacharyya, Mami Hayashida, Jack Bandy, Kristina Gessel, Bruno Athie Teruel, Sydney Chapman, and many other students for being great teammates, sharing in excitement, and always being nice to be around. Thank you to Beth Lutin and Amy Long for the same, and for supporting the logistics necessary to do this work. Thank you to Margaret Sumney for reading drafts of this dissertation and offering valuable feedback. For supporting me throughout this process I am grateful to my committee and outside examiner: Shawn Cheng, Samson Cheung, Nathan Jacobs, and Shuxia Wang.

This field requires large multidisciplinary teams, and somehow everyone involved is lovely. It has been a joy to work with Fabrizio Diozzi and Federica Nicolardi and their respective teams in Naples. Thank you to James Brusuelas, Marzia D'Angelo, Michael McOsker, Gianluca Del Mastro, Richard Janko, and others for your papryological insights. And thank you to Catherine Patterson and Douglas MacLennan at the Getty Conservation Institute for yet another project with great people.

I took a scary leap outside of my comfort zone with the Vesuvius Challenge, and at every step the team has been supportive and made this an exciting experience. To Nat Friedman, Daniel Gross, Daniel Havir, and JP Posma, thank you. I can't believe I have had the opportunity to work with you all and to see this project come to life. Thank you also to the members of the resulting community for pushing me, pushing the research, and reminding me daily why I am so excited about this work.

I am grateful for direct support from a National Science Foundation Graduate Research Fellowship under Grant No. 1839289. Field work and data acquisition were made possible by The Andrew W. Mellon Foundation, and by the support of the Arts and Humanities Research Council (AHRC) under Grant Reference no. AH/S005935/1. Complementary work has been made possible in part by the National Endowment for the Humanities: Democracy demands wisdom. Any opinions, findings, and conclusions or recommendations expressed in this dissertation are my own and do not necessarily reflect the views of these sponsors.

Above all, I will never understand how I have stumbled into such a wealth of friends and family. Truly it makes no sense. You all lift me up daily. I am especially grateful for the support of my mom Julie Parsons and my brother Michael Parsons. Chris, Elise, Terry, Granddad, and Dad: I carry you with me always. Kaitlyn: you inspire me, I adore you, and with you I can do anything.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

O ye, who patiently explore
The wreck of Herculanean lore,
What rapture! could ye seize
Some Theban fragment, or unroll
One precious, tender-hearted scroll
Of pure Simonides.

September, 1819
William Wordsworth

This dissertation is about using computer vision, and sometimes machine learning, to extract information from images. Specifically, I want you to see, as I have seen, the possibilities often exceed our expectations. In other words, there is more than meets the eye.

This is especially true with forms of imaging other than traditional photography. By traditional photography, I mean what we think of as "photos" in our daily lives. Photos resemble what our brains are wired to see and process. These images are two-dimensional, and captured in the visible range of the light spectrum.

Both of those constraints have been blown open by modern imaging technology. In addition to visible wavelengths, images can be captured in ultraviolet, infrared, X-ray, and more. Further, images can be acquired in 3D: they can follow a surface shape, or in the case of tomography can be fully volumetric. These powerful images exceed the built-in human ability to view or process them, and we are consistently forced to reduce them to lesser forms so we can make sense of them with our eyes. Algorithms are not restricted by the same constraints, and are in many cases able to extract more from images than would be possible with the naked eye.

To explore these ideas, this work will focus in on the world of damaged manuscripts: scrolls and books, damaged beyond repair, so fragile they cannot be opened physically as they would fall to pieces. The only way to recover their contents is to use non-invasive imaging (typically X-ray computed tomography, or CT) and process the

resulting images to reveal the text inside.

Like many others before me, I became captured by a particular collection of texts: the Herculaneum scrolls. In addition to having outsize historical value, they are an unmatched and compelling technical challenge. Reading these scrolls noninvasively is on the threshold of possibility in so many ways that it is comical (some days darkly so). Standing on very tall shoulders, I have chipped away at this technical problem, and this is the subject of the dissertation ahead.

I want to convey the sense of excitement I have developed while doing this work, both for the power of multidimensional imaging and for our ability to extract surprising results from it. This dissertation is a step in that direction, and the findings have implications for a number of fields where high-dimensional images are used. The primary motivation of the work, however, has been to find a method to read the unknown texts of the intact Herculaneum scrolls.

Alright, let's dig in.

1.1 The Herculaneum scrolls

Much of what we know about ancient texts comes from copies. Typically written on papyrus, these texts were subject to physical decay and had to be copied regularly in order to survive to medieval manuscripts and then modernity. As communities come and go and empires rise and fall, many of these texts were lost in the process. Surviving original texts from antiquity require fortuitous physical conditions: an arid climate, some burning (but not too much), or other similar circumstances that can preserve papyrus for thousands of years.

Original texts preserved in this way are exceedingly rare, and often fragmentary. From the entirety of the ancient world, there exists only a single library that has survived in situ: the library of the Villa dei Papiri in Herculaneum. [1]

The preservation of the library and its texts was caused by an event that also led

[1] For more background on the Villa and collection, I recommend David Sider's excellent book on the Herculaneum papyri [1].

to widespread destruction. The eruption of Mount Vesuvius in A.D. 79 buried the Vesuvian towns of Pompeii, Stabiae, and Herculaneum. In Herculaneum in particular, successive waves of pyroclastic flow covered the site. Incandescent rocks, ash, and hot gas swept over the town, destroying the upper levels of buildings and killing the town's inhabitants instantly. The flow simultaneously brought the buried materials of the town to combustion temperature and extinguished the oxygen supply, leading to a process called carbonization that is identical to the way charcoal is made.

Among the other buried ruins of Herculaneum sat a sumptuous seaside villa. Possibly belonging to Julius Caesar's father in law Lucius Calpurnius Piso Caesoninus, this villa was initially prized during excavation for its many bronze and marble statues, but is now known for its even more extraordinary contents. Today called the Villa dei Papiri, the villa housed a library of papyrus scrolls, preserved in place by carbonization during the eruption.

Though initially discarded or ignored due to their appearance resembling burnt sticks, these scrolls were quickly valued for what they are: the only surviving library from antiquity.

1.2 Physical unrolling attempts

Immediately following their discovery in 1752, scholars began efforts to unroll and read the charred papyri. As they could not have foreseen noninvasive imaging technology at that time, early methods were decidedly invasive. The first technique was to cut the scroll in half lengthwise with a knife. The innermost layers could be scraped away, revealing some text on the layers beneath. Once this text was transcribed, that layer was itself scraped away to reveal the next one. This method destroyed most of the scroll, and certainly the portions near the scroll center. This is particularly unfortunate, as these sections were often the least damaged, and may have been openable using gentler methods. These sections, near the end of the text, also often contained useful information such as the author, title, book number, or number of lines, all now

lost.

Other methods attempted included pouring mercury between the scroll layers, which crushed them, immersing the scrolls in mercury, which crushed them even more, and the use of other substances such as rose water, "vegetable gas," and others, none of which worked.

Eventually it was realized that the scroll need not be cut completely in two, but the vertical cut could stop midway to the scroll center, creating now three pieces. There was a narrower scroll center, still intact, in addition to two opposing outer pieces. These outer pieces were often then processed as before, scraping away successive layers to reveal and record some portion of the text.

Father Antonio Piaggio is responsible for the most successful physical unrolling. Using "Piaggio's machine," the narrow scroll centers were very gently and slowly unrolled. Animal membrane was attached to the outermost layer and suspended from above, allowing gravity to slowly pull the layer away. Though still somewhat destructive, far more caution was exercised using this technique than the alternatives, and much text has been recovered using this method.

Many other methods have been attempted since. Moist air, water immersion, smoking in sulfur gas, electric fields, and more have been attempted. None have improved measurably over Piaggio's method, and most have been quite destructive. The most recent somewhat successful efforts were led by Knut Kleve [2] as recent as the 1980s.

The purpose of outlining the variety of physical methods that have been attempted is to give some sense of the breadth and diversity of the collection as it exists today. Figure 1.1 shows a small sample of some of the forms now taken by the scrolls. Some scrolls, typically the ones that seemed the most challenging to open, have been left rolled and intact. Piaggio's and similar methods have yielded ~ 3500 trays of opened material, much of which has large regions of text visible, particularly when imaged

(a) Intact scroll.

(b) Larger unrolled section.

(c) Fragments.

(d) Individual fragment.

Figure 1.1: A small sample of the diversity of material presently included in the Herculaneum scroll collection.

under infrared light. These and the other more destructive methods have also resulted in thousands of smaller scroll fragments, in every imaginable condition.

The carbonization resulting from Mount Vesuvius's eruption has resulted in damaged, brittle papyrus rolls that are extremely challenging to open and read. These same forces, however, are responsible for preserving the library. Without carbonization, this library would have been lost to decay like the others from the ancient world.

1.3 Virtual unwrapping

During the last few decades, advancements in imaging technology such as X-ray CT have led to hopes that the scrolls may be decipherable noninvasively. It is now difficult to imagine that physical unrolling methods would ever be again permitted on the remaining scrolls. Despite this optimism, as well as successful results with other manuscripts, noninvasive methods have yet to deliver on the promise of revealing the hidden texts of the Herculaneum papyri. This work presents a set of methods that

move us much closer to that goal. To do so, it builds on a technique known as virtual unwrapping.

Virtual unwrapping [3, 4] has now been established as a prominent intersection of heritage science and computational methods, and there are a growing number of successful applications. Virtual unwrapping involves noninvasive volumetric imaging followed by a multi-step computational pipeline that traces the 3D geometry of the surfaces (segmentation), filters the volumetric image to make writing appear (texturing), and maps the result to a 2D image that is easily read (flattening). The details of these steps and their ordering can vary to adapt to different challenges.

Multiple works have applied virtual unwrapping as a functional proof of concept on lab-made proxy manuscripts, illustrating the potential across a range of imaging methods, writing substrates, inks, and manuscript forms such as scrolls, codices, and folded sheets. Early work focused on proxy materials with exaggerated features, validating the computational concept using much lower imaging resolution than is now available [5–7]. Since then, virtual unwrapping has worked for bamboo scrolls [8, 9], metallic inks on parchment [10, 11], rolled and folded papyri [12, 13], and even specifically carbon ink on papyrus in the right conditions [14, 15].

Building on the learnings from lab-made proxy manuscripts, other works have gone on to recover text from genuine heritage manuscripts. These also span a wide range of materials, including metallic inks on parchment, paper, and papyri [16–24], etched metal scrolls [25, 26], and lead amulets [27]. Some works [28] focus specifically on segmentation, leaving ink detection as future work. Ultimately, the goal of developing virtual unwrapping methods is to recover substantial texts leading to new material for scholarly publication, an objective achieved with the En-Gedi scroll [21, 29].

The above works vary significantly in the segmentation approach and in the imaging method. The segmentation approach varies in order to handle different materials and their respective geometries. Imaging techniques are typically varied in order to

generate high contrast between the ink and substrate, if possible. Relying on this strong contrast, texturing methods are consistently simple local filters that directly extract the image intensities in a local neighborhood. Recently, this has shifted with the introduction of machine learning-based texturing methods that detect subtle signals [14] and learn to map the volumetric input to other image domains [30].

Existing virtual unwrapping approaches encounter two primary roadblocks when applied to the Herculaneum scrolls. First, the carbon-based ink used in this collection seems to have no contrast in X-ray against the papyrus substrate, which is also largely made of carbon. Following segmentation and flattening, one is left with a virtual blank papyrus sheet. The primary focus of this work is on methods to extract more ink contrast from these images using machine learning. Another obstacle is the segmentation itself, which is challenging with the Herculaneum scrolls due to the complex structure of papyrus sheets, the large number of layers in each scroll, and the physical damage to the scrolls resulting in warped and tightly compressed layers. Though work remains, this work presents large steps forward on both of these fronts.

1.4 Imaging the Herculaneum scrolls

A growing body of work has pursued multiple avenues to chemically or otherwise characterize the Herculaneum papyri and their ink. These studies are oriented towards the discovery or development of imaging methods that are capable of highlighting the ink of the scrolls with visual contrast against the papyrus background. Both the scroll fragments with exposed, visible text, and the intact, rolled scrolls with hidden text have been the targets of these investigations.

For scroll fragments with exposed writing, strong visual contrast is possible through the use of spectral imaging in the infrared bands [31, 32]. For a rare Herculaneum fragment with text on the verso, or the back side of the writing surface, the use of shortwave-infrared (1000-2500 nm) imaging has even enabled the recovery of this writing [33]. While exciting for the few exposed fragments with verso text, this

method has limited penetration into the papyrus and would not extend to the hidden layers within a rolled scroll.

Considerable work has investigated the chemical composition of the ink of the Herculaneum fragments with the goal of informing future imaging methods that would be able to capture clear ink contrast from within an intact scroll. There have been mixed results. Early investigations [34] used at least nine different imaging technologies on three Herculaneum fragments, looking for signals that would differentiate ink from papyrus. Scanning electron microscopy and X-ray fluorescence found calcium in the ink but not papyrus, and particle-induced X-ray emission suggested the trace presence of lead and strontium in the ink only.

Other work with different Herculaneum fragments has discovered a stronger presence of lead in the ink, leading to clear imaging contrast using X-ray fluorescence [35, 36]. The reason for the lead presence is not yet known, but could arise from contamination or from deliberate use as a pigment, binding, or drying agent. It is also not known why the ink of some fragments exhibits a strong lead profile while others do not, though this is not too surprising, considering the Herculaneum papyri were authored by different scribes, using homemade inks, over a period of three centuries. A more recent and thorough X-ray fluorescence study of 38 Herculaneum scroll fragments validated the other studies, finding multiple elements that sometimes correlated strongly with ink: phosphorus, 5 fragments; iron, 3; copper, 3; and lead, 2 [37]. Carbon-based inks not specific to the Herculaneum papyri have also been studied thoroughly, with similar findings and implications [38–40].

When viewed with extremely high resolution, as is possible with scanning electron microscopy, a form of ink contrast other than an intensity shift becomes visible. To examine this, a papyrus sheet with carbon ink characters was made in the lab and imaged. Figure 1.2 shows a detail of a small region along the ink/no-ink boundary of one of the characters. In this imaging modality, the ink shows not only an intensity

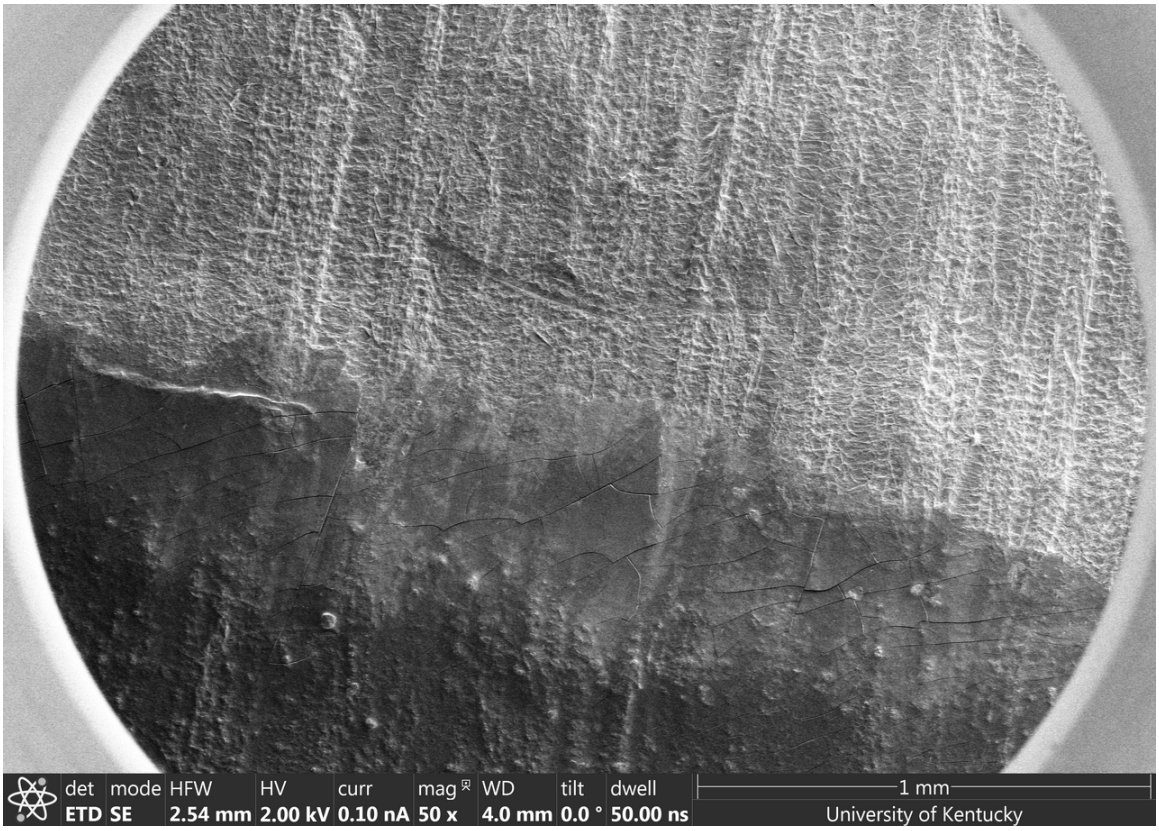


Figure 1.2: Scanning electron microscopy of exposed boundary between carbon ink (bottom) and bare papyrus (top), showing textural or morphological contrast.

difference but also a *textural* or *morphological* contrast. The surface of the ink is smooth, with some cracks, while the bare papyrus surface is more uneven. One can imagine if the scrolls could be imaged in X-ray CT at this resolution, these patterns would be visible to the naked eye even if there were no intensity contrast. This resolution is not feasible for CT, but I suggest there are similar morphological effects captured in the achievable CT images. Though too subtle to see by eye, it seems machine learning methods are capable of detecting them. This work refers to this idea as the *morphological hypothesis*, and this hypothesis seems supported by the experimental evidence.

The same lab-made papyrus sheet was also physically cut through the middle and imaged in SEM again, this time imaging the ink layer in cross section. The ink layer

was measured to be $\sim$ 5 µm in thickness. This method cannot be applied to the real Herculaneum scrolls due to its destructive nature, but this lab-made proxy likely approximates the Herculaneum ink thickness to the right order of magnitude. This information is used to inform the imaging protocol, and to evaluate the experimental results achieved.

Imaging of the intact Herculaneum scrolls has also matured, despite not yet achieving clear ink contrast. The first X-ray CT images of intact Herculaneum scrolls revealed the internal structure but no ink contrast [41, 42]. Phase contrast X-ray CT was also proposed [34] and then conducted [43–45] as a potential technique to achieve ink contrast inside a rolled scroll. Despite early claims of textual discovery, this technique alone has not yet led to further discoveries or ongoing scholarly work.

Unlike the prior work, which prioritized phase contrast effects above spatial resolution, the work presented in this dissertation focuses on X-ray micro-CT acquired at extremely high resolution, showing that a software pipeline is capable of recovering more ink contrast in this data than is immediately apparent. The research and development of methods for improved imaging contrast are still highly complementary to the work presented here. Though machine learning-based methods are now capable of detecting Herculaneum ink in X-ray CT images, there remains room for improvement. This will happen in part due solely to improved software pipelines using existing data, but any improvement in imaging contrast would greatly boost the ultimate accuracy. The additional information captured by phase contrast or other techniques can ultimately only help, as trained models can select the most useful features from the input images.

Figure 1.3 visually summarizes some imaging methods on the same scroll fragment. In color photography, text is barely visible, as the black ink shows little contrast against the carbonized papyrus. Infrared photography shows much clearer ink contrast. In X-ray, even less contrast is visible than in color photography, with es-

(a) Color photograph. (b) 1000nm infrared. (c) X-ray.

Figure 1.3: Different images of P.Herc.Paris. 2 fr. 47 (short for Papyrus Herculaneum Paris 2, fragment 47).

sentially no ink being readily visible in this view. Despite this initial lack of contrast, this work will show there is ink signal captured in the X-ray that can be captured computationally.

In broad strokes, the overall situation of scroll collection and available imaging methods is summarized in Figure 1.4. Contrast is achieved easily for the scroll fragments with exposed text, though the imaging methods do not penetrate enough to reveal the contents of the rolled scrolls. Similarly, the interiors of the rolled scrolls can be imaged at high resolution, but without evident ink contrast. The objective is the best of both worlds: ink contrast from within a rolled scroll.

1.5 Definitions

Some helpful definitions are included here:

- **Scroll:** a rolled manuscript, typically formed from a single long sheet rolled upon itself. Due to physical unrolling efforts, what was once one Herculaneum scroll is often now many pieces.

- **Fragment:** A detached, smaller piece from a scroll, typically with exposed text that is evident in infrared if not in visible light.

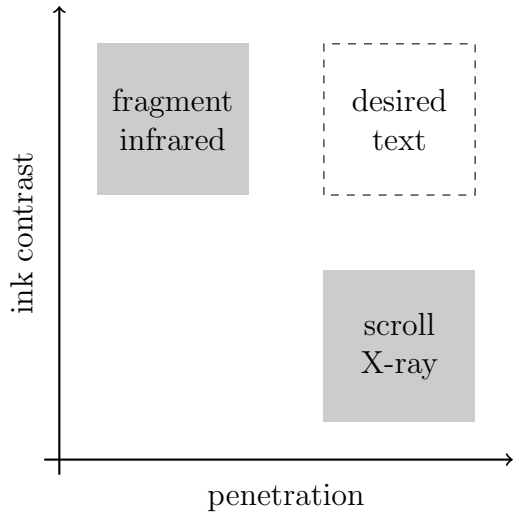


Figure 1.4: Infrared images reveal ink contrast, but do not penetrate to scroll interior. X-ray CT penetrates to scroll interior, but without clear ink contrast. To reveal the desired text, both contrast and penetration are necessary.

- **Intact:** A scroll that remains complete or partially complete in its rolled form. Textual contents unknown.
- **Layer:** In reference to a manuscript, indicates one thickness of writing material, often positioned over or under other layers. A scroll is a single sheet that, when rolled, forms many layers.
- **Tray:** A tray with large, relatively flat scroll pieces, often the result of Piaggio's unrolling method. Most active scholarship of the Herculaneum scrolls uses these trays.
- **Recto:** The "front" face or side of a writing surface.
- **Verso:** The "back" or "reverse" face or side of a writing surface.
- **Papyrus:** A writing surface made from the fibers of a reedy plant of the same name and used for the Herculaneum scrolls. Two layers of fibers are pressed and naturally adhere to form one sheet. The resulting sheet displays a grid-like pattern of plant **fibers**.

- **Papyrologist:** A classicist specialized in the study of texts on papyrus.
- **Papyri:** Plural of papyrus, but typically used to refer to the scrolls collectively, e.g. "the Herculaneum papyri."
- **Parchment:** A writing surface made from thinned animal skin with a much simpler structure than the fibers of a papyrus sheet.
- **Substrate:** General term for a writing surface upon which ink is applied.
- **Volumetric:** A dense 3D image such as that produced by a CT scan that contains pixel (or "voxel") information for each point in a dense 3D grid. Often visualized using 2D image slices.
- **3D:** Three-dimensional. Based on context, can refer either to dense 3D (volumetric) images, or to sparser representations such as geometric meshes.
- **Computed tomography (CT):** A noninvasive imaging method, often but not always using X-ray, that offers a volumetric view of object interiors. Individual X-ray projection images are transformed via the reconstruction algorithm to their volumetric representation.

Further clarification is also offered here regarding what the methods in this work aim to recover. **Ink** is a physical substance applied to a writing substrate, ideally resulting in visual contrast. The technical methods in this work are primarily concerned specifically with ink detection, and use highly localized information to detect ink presence. Detecting ink may or may not be sufficient for textual scholarship, based on completeness and whether there is enough detected to reliably form written **characters** a scholar can interpret. Scholars rely not only on characters, but on their combinations into **words**, referred to as **text** or **writing**, requiring that the characters follow expected contextual conventions of size, orientation, spacing, vocabulary, and so on.

The methods in this dissertation have no knowledge of characters, and look only for highly localized ink presence. In other words, the presented methods are not optical character recognition (OCR) or related techniques. Their ultimate objective, however, is to detect ink reliably enough that the resulting images do in fact display characters forming text or writing.

1.6 Thesis statement

The thesis of this dissertation can be summarized as follows:

> High-dimensional images such as volumetric CT exceed our visual capacity, capturing some patterns humans are unable to see directly. By leveraging other image modalities that more clearly capture these patterns, it is possible to train machine learning-based models that surpass human expert detection, extracting more from the data than we can observe. This approach is capable of detecting the presence of carbon ink in X-ray CT, a task previously considered impossible. When applied to the Herculaneum papyri, this enables the successful recovery of hidden writing, unseen for almost 2,000 years. This revelation can be achieved noninvasively, without damage or physical disturbance to the fragile scroll.

1.7 Research contributions

This work is centered around the following research contributions, also serving as a roadmap of the dissertation:

- **Ink detection framework:** Chapter 2 presents a novel geometric framework and machine learning-based software pipeline capable of verifiably recovering carbon ink from X-ray CT.
- **Ink detection results:** Chapter 3 shows experimental results demonstrating the successful recovery of carbon ink from the Herculaneum papyri using X-ray

CT. Evaluation techniques are presented that validate the results using visual, quantitative, and papyrological methods. An exploratory data analysis also deepens our understanding of Herculaneum papyri, and its carbon ink, in X-ray CT.

- **Multimodal transformations:** Chapter 4 introduces a multimodal paradigm for the software pipeline, generalizing the approach to learn mappings between image domains, for example simulating a color photograph from X-ray CT.

- **Hidden texts:** Chapter 5 discusses the application of these methods to hidden, internal layers. Requisite tools are developed, and the domain shift between training and inference data is examined. Characters from the hidden layers of Herculaneum papyri are revealed noninvasively for the first time.

- **Ablation studies:** Chapter 6 demonstrates ablation and related experiments that probe the limits of the various stages of the ink detection pipeline.

- **Vesuvius Challenge:** Chapter 7 discusses the Vesuvius Challenge, an open research contest launched based on the seminal contributions of this dissertation in order to accelerate and scale up the restoration of text from the massive Herculaneum collection. The associated code and data release, a first of its kind dataset, opens the research problem to the global community.

Chapter 8 then discusses the status and outlook of this particular research problem, and places the work in broader context.

CHAPTER 2. INK DETECTION: FRAMEWORK

2.1 Introduction

This chapter presents a method capable of detecting the presence of carbon ink on Herculaneum papyri from inside X-ray CT images. [1] This method closes the gap left by the previous discussion, in which there was no technique capable of combining ink contrast with the penetration required to image the interior of the intact scrolls. This method was developed in tandem with the datasets and experimental results it enables, which are discussed in Chapter 3.

This chapter focuses on the method itself, centering on the following contributions:

- **Geometric framework:** an enhanced virtual unwrapping pipeline linking 3D and 2D images, enabling the labeling of subtle signals such as carbon ink in X-ray CT.

- **ink-ID**: a reference implementation of the machine learning-based ink detection enabled by the above geometric framework. ink-ID has three primary functions: dynamic data sampling from large volumes, machine learning training and inference, and dataset management and visualization.

2.2 Background

This section discusses the conceptual framing of machine learning-based ink detection. The method relies on two optimistic assumptions at the beginning, visualized in Figure 2.1. Images are considered of varying "resolution," here a simplification of various real-world factors that can also be described as the ability to capture the presence of carbon ink. As seen in Figure 1.2, there is a resolution threshold beyond which the ink is detectable in the image by the human eye. This threshold is exceeded

[1] The core ideas of this chapter were introduced in [14] and extended in [46]. This chapter consolidates these ideas and provides additional details and discussion. Some text overlaps with the above publications.

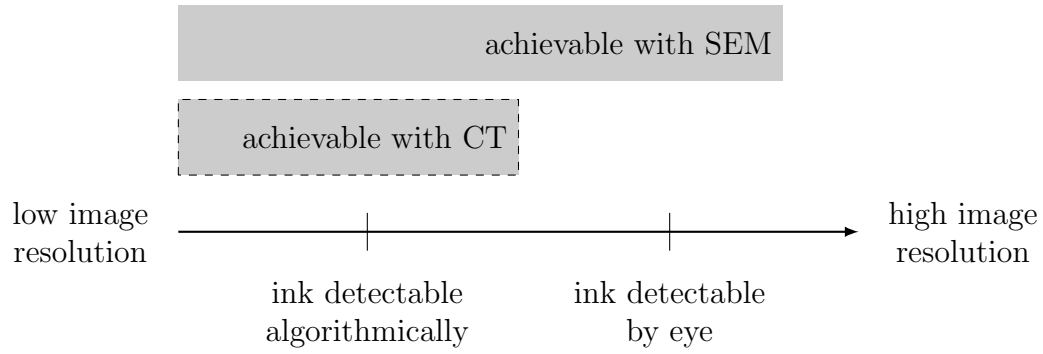


Figure 2.1: Two optimistic assumptions behind machine learning-based ink detection, with "resolution" simplifying multiple real-world factors.. First: algorithms can detect carbon ink at a lower resolution threshold than human eyes. Second: X-ray micro-CT is capable of meeting this threshold.

by scanning electron microscopy, which captures enough textural detail that ink is easily visually distinguished from bare papyrus.

The first assumption required for machine learning-based ink detection is that there is a separate threshold for algorithmic ink detection, and that it is lower than that for visible ink detection. This creates a window of opportunity, in which imaging methods could capture enough of the ink presence that it can be detected computationally, but not enough that it can be detected visually.

The second assumption is that X-ray micro-CT falls within this window, surpassing the imaging threshold for ink detectability. Perhaps the carbon ink of Herculaneum papyri is actually captured in existing CT images, even if it is so subtle that it eludes human observation. If this is true, an algorithm could conceivably detect the ink presence even where eyes cannot.

The challenge in developing this algorithm is that it is unclear at the outset what pattern the algorithm should be designed to detect. It is established that there is no clear ink pattern discernable to human eyes in the CT images. If there were some form of visible ink contrast, an algorithm could be tailored to automatically detect it; however, this visible contrast would obviate the need for the algorithm in the first place.

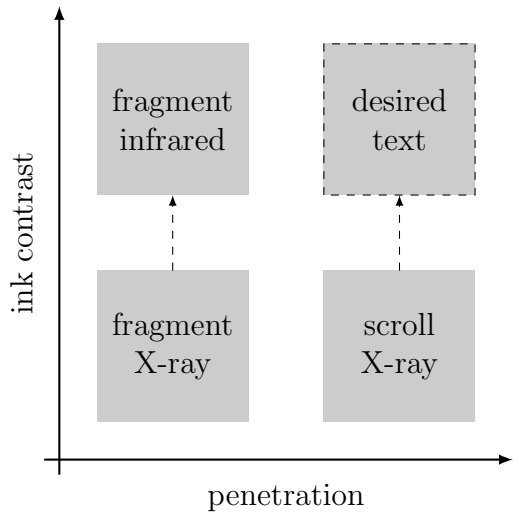


Figure 2.2: Using fragments with known ground truth, models can be trained to learn the relationship between X-ray CT and infrared appearance (or ink presence). Models can then be applied to intact scrolls, revealing their contents.

The opened scroll fragments of the Herculaneum collection offer the possibility of circumventing this problem using supervised learning. As the fragments have exposed, visible text, the known locations of ink can be associated with the same locations in the CT image. A model can then be trained to predict the presence of ink given a CT input. If indeed there is more signal present in the image than the human eye is capable of detecting, a model trained in this manner should be able to find it.

The ultimate goal of such a model, trained on the exposed scroll fragments with known ground truth, is to reveal the unknown writing of the intact scrolls. This happens following training, when during inference the model is applied to a CT scan of a rolled scroll. Figure 2.2 illustrates this idea, showing how the use of scroll fragments as labeled training data augments the situation shown earlier in Figure 1.4.

Designed around this principle, this chapter introduces an enhanced virtual unwrapping pipeline that combines machine learning with a novel geometric framework linking 3D and 2D images. The geometric framework provides a bidirectional map-

ping between 2D and 3D image spaces, which allows the use of 2D images as labels for the 3D CT data. Models can then be trained using supervised learning to detect the presence of ink from CT images alone.

One central aspect of this approach is that it is inherently verifiable. Before being applied to scans of intact scrolls with unknown contents, the models can be validated on the labeled scroll fragments using cross validation or other train/test splits. This validation is used throughout the development process, and builds confidence that the model is accurately detecting the actual ink presence, not merely hallucinating plausible shapes that resemble characters.

2.3 Geometric framework

The creation of paired set of input and label images relies on a shared image space to which they can each be aligned. To make this space, the surface is first segmented, or traced, from the CT image. This 3D surface is then flattened to a plane, onto which the 2D images of the same surface can be registered, or aligned. This section provides an overview of the different pipeline components, and then discusses their creation and alignment in more detail.

2.3.1 Overview

Figure 2.3 shows an overview of the geometric framework for a single scroll fragment, P.Herc.Paris. 2 fr. 47. All images, both inputs and labels, come from this same object, shown in a color photograph in Figure 2.3a. The fragment is scanned using X-ray CT, creating a volumetric representation of the surface and the structure below the surface (Figure 2.3b).

From the CT image, the 3D surface of the fragment is first segmented, yielding a surface mesh that closely follows the writing surface (Figure 2.3c). This mesh is then flattened to a plane, defining the 2D image space to which the label images will later be aligned. Following flattening, the UV map of the mesh defines the $\mathbb{R}^3 \to \mathbb{R}^2$ mapping that translates coordinates from the 3D mesh surface to the corresponding

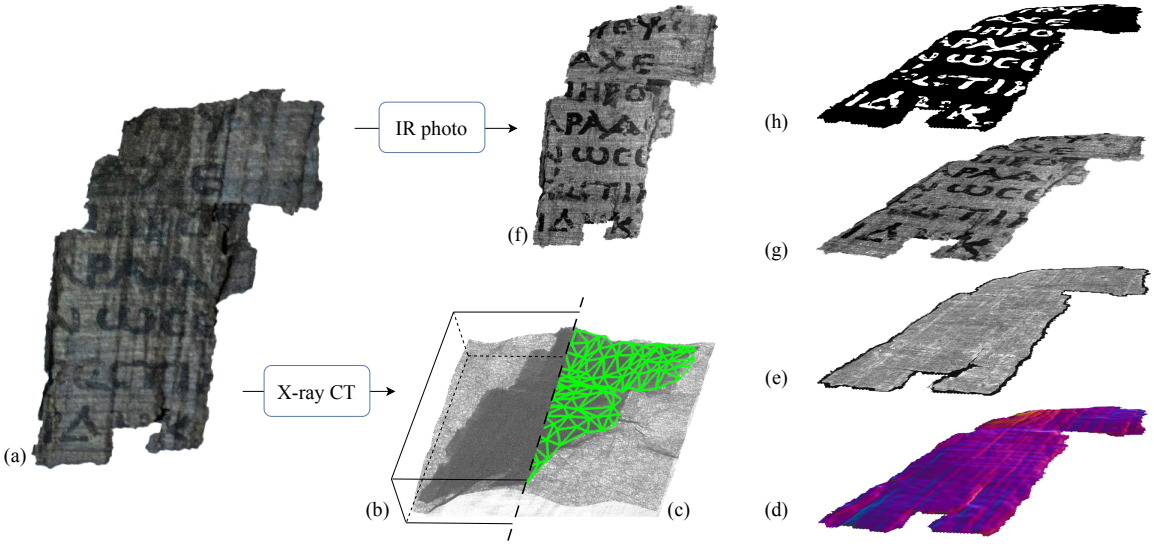


Figure 2.3: Geometric framework, visualized with P.Herc.Paris. 2 fr. 47. (a) Scroll fragment, RGB photograph. (b) Volumetric X-ray CT image. (c) 3D surface segmentation. (d) Per-pixel map (PPM). (e) Texture image. (f) Infrared photograph. (g) Infrared photograph aligned to texture image. (h) Aligned binary ink labels.

coordinates in the new 2D image.

To make the mapping bidirectional, a per-pixel map (PPM) is also generated (Figure 2.3d). The PPM is a 6-channel 2D image of the flattened surface mesh. The PPM defines the $\mathbb{R}^2 \rightarrow \mathbb{R}^3$ mapping by storing at each 2D pixel the 3D coordinate and 3D surface normal vector of the corresponding position on the original surface mesh. As the PPM and label images are aligned, this means that each label pixel can be mapped to a position and orientation in the 3D volume. The following notation is used to define the coordinate frames of these images:

(u, v) or $\vec{p}$: 2D coordinates in the PPM image P or, by definition, also the label images.

(x, y, z) or $\vec{v}$: corresponding 3D positions in the CT volume.

(n_x, n_y, n_z) or $\vec{n}$: corresponding 3D surface normal vectors in the CT volume.

In principle, the spatial resolution of the PPM P (and therefore the label images) is

decoupled from that of the CT volume. In practice, the implementations in this work use a pixel size in the PPM P that is identical to the voxel size in the CT volume.

Using the PPM, a texture image (Figure 2.3e) is generated, which plots the intensities from the original CT scan onto the 2D image plane. Though the low contrast ink does not readily appear, this image reveals the structure of the papyrus fibers on the fragment surface. It can therefore be used as a fixed image to which an infrared image of the fragment surface (Figure 2.3f) can be registered, creating an aligned label image (Figure 2.3g). This label image can be further processed, for example by manually tracing the text characters to create a binary ink mask (Figure 2.3h).

The product of this process is an aligned set of images, mapping labels on the fragment surface to their corresponding positions in the 3D CT volume. As a result, (input, label) pairs can be generated that relate small neighborhoods in the CT image to the presence of ink at that location. These inputs are used to train models using supervised learning, discussed in Section 2.4. The remainder of this section will discuss the design and implementation of the above pipeline steps, continuing to use P.Herc.Paris. 2 fr. 47 as an example case. This same process is applied to other fragments to build the training set.

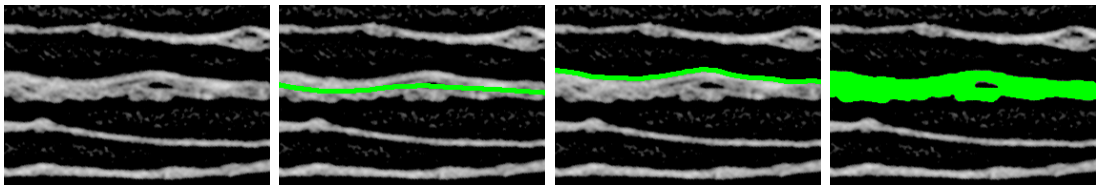
2.3.2 Acquisition

The details of the acquisition process vary based on the object itself and on the imaging instrumentation. This is true for both the micro-CT (Figure 2.3b) and the surface (Figure 2.3f) images. This stage will be explored in more detail in later sections, with respect to the particular datasets discussed.

2.3.3 Segmentation

Segmentation is the important step of tracing the detailed 3D structure of the surface of interest. Capturing this surface shape is the first step in defining the bidirectional mapping that links 3D to 2D and enables paired image labels.

There are multiple options for how the desired segmentation may be defined. In all



(a) Papyrus layer. (b) Mesh, midpoint. (c) Mesh, surface. (d) Volumetric.

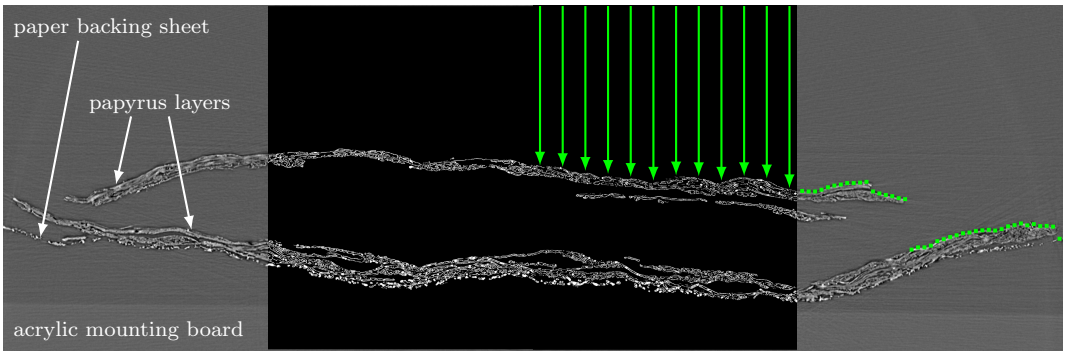
Figure 2.4: Possible segmentation objectives for the papyrus layer shown in (a).

cases, some 3D representation of the layer shape is sought. As shown in Figure 2.4, the specific form of this shape can vary. For the central papyrus layer captured in the CT cross section in Figure 2.4a, a triangular mesh could be fit to either the midpoint of the layer thickness (Figure 2.4b) or to either face of the layer, here the topmost surface being shown (Figure 2.4c). Instead of a mesh, in some cases a volumetric segmentation may be desired (Figure 2.4d), capturing not only the surface shape on both faces but also its thickness and other more subtle patterns. Unless otherwise specified, this work assumes that the desired segmentation is a mesh following one face of the papyrus sheet (Figure 2.4c), typically the face on which writing is known or expected to exist.

The remainder of Section 2.3.3 outlines the segmentation method used for the exposed fragment surfaces with known ground truth. This is the most common segmentation method used in this work, forming the labeled dataset that is the basis of the ink identification method presented. For hidden, subsurface layers, and for some other case studies beyond the Herculaneum scrolls, slight variants are sometimes used, which are described as they arise.

Initial segmentation with Canny edge detector

An initial segmentation is first constructed using Algorithm 1, which is visualized in Figure 2.5. The algorithm iterates over the set of volume slices. For each slice, a Canny edge detector [47] is first applied to the image, generating a binary mask labeling the detected edges (Figure 2.5b). A set of rays is then projected from one of the slice image boundaries into the image center, each ray proceeding until it



(a) Slice image. (b) Detected edges. (c) Projected rays. (d) Resulting points.

Figure 2.5: Visualization of Algorithm 1 applied to a slice of P.Herc.Paris. 2 fr. 47.

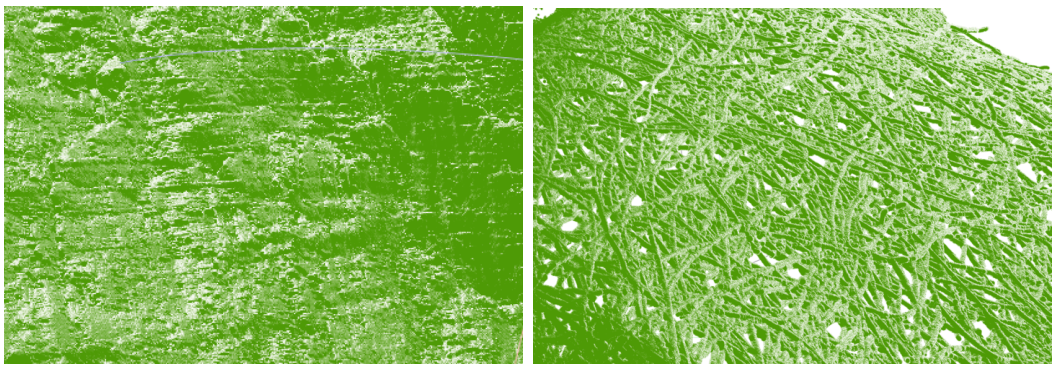
intersects a detected edge (Figure 2.5c). The intersections create a set of (x, y) points in the slice image defining the papyrus surface (Figure 2.5d).

Algorithm 1 Surface segmentation using Canny edge detection. Given a volume V and Canny edge detection parameters C, returns a dense point cloud P with points defining the fragment surface.

```
1:  procedure CANNY-SEGMENT(V, C)
2:      P = ∅                                          ▷ empty point cloud
3:      for z_i ∈ [0, V.slices − 1] do
4:          I_i = V[z_i]                               ▷ extract slice image
5:          E_i = CANNY(I_i, C)                        ▷ detect edges
6:          for x_j ∈ [0, E_i.cols − 1] do
7:              for y_k ∈ [0, E_i.rows − 1] do
8:                  if EDGE(E_i[y_k, x_j]) then
9:                      P = P ∪ (x_j, y_k, z_i)        ▷ add point
10:                     break
11:     return P
```

The slice index is used as the z coordinate for all points in the slice image, converting the set of points to 3D. The set of all points across all slices defines a dense 3D point cloud, with points defining the fragment surface in the coordinate space of the volumetric CT scan.

The raw point clouds generated using this method contain a high level of detail. Visualization reveals not only the general shape of the surface, but also textural details. Figure 2.6 shows two examples of these details, where the grid-like structure of the papyrus fibers is visible on the fragment, and the paper fibers are visible on



(a) Papyrus fragment surface. (b) Paper backing sheet.

Figure 2.6: Details of raw point cloud from P.Herc.Paris. 2 fr. 47, showing surface textures recovered using Canny segmentation. Visualized using Meshlab.

the backing sheet to which the fragment is mounted.

The success of segmentation using Canny edge detection depends highly on the Canny parameters chosen. The optimal parameters vary by dataset, and require some exploration before the segmentation quality is adequate. For this parameter exploration, a graphical user interface was created to interactively experiment with the edge detection parameters (Figure 2.7). For larger datasets, it may become desirable or necessary to automate this optimization process. For the datasets used in this work, it was most time effective to create the graphical interface and spend a few minutes per CT scan choosing the optimal parameters.

Selecting the optimal Canny edge detection parameters is a balancing act between false positives and false negatives. Figure 2.8 demonstrates the failure cases to be avoided. When tuning the edge detector to a particular dataset, the user adjusts the parameters while observing the visual results, and tries to find a sweet spot that performs well across all slices of the scan. This can be difficult for some datasets, particularly those with CT noise in the air regions, as removing the noisy detected edges from those regions often leads to false negatives on the actual papyrus surface, resulting in "holes" in the final segmentation. Where a compromise is forced between false positives or false negatives, false negatives are given slight preference, as they are

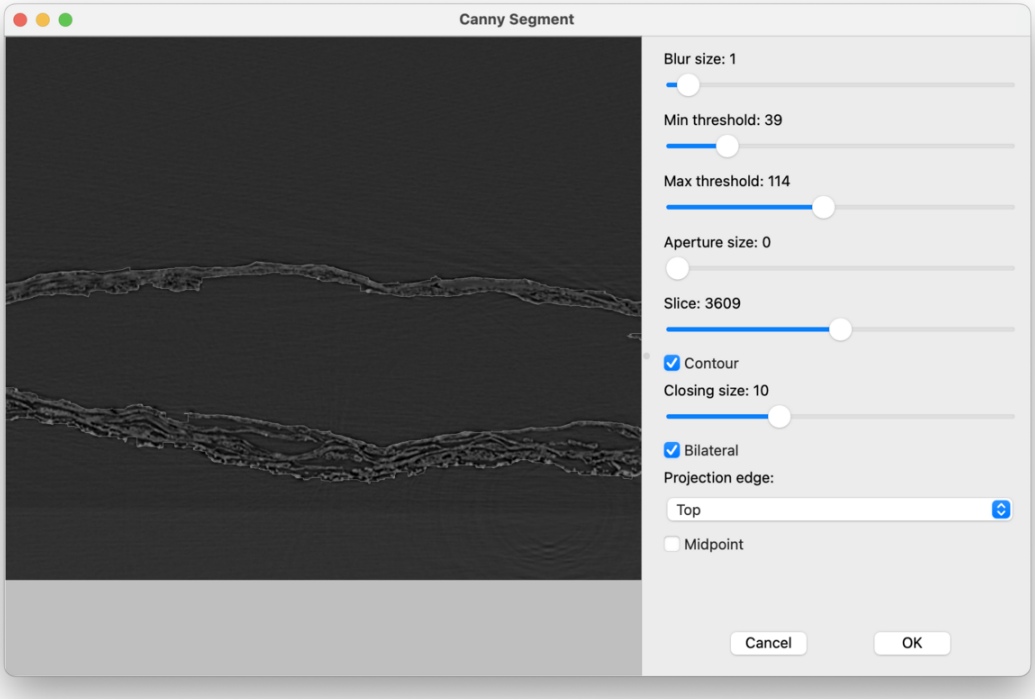
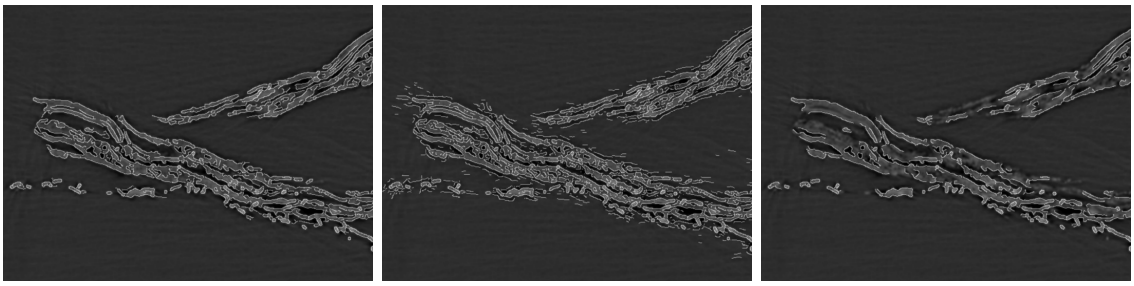


Figure 2.7: Graphical user interface for determining optimal Canny edge detection parameters, then input to Algorithm 1.



(a) Optimal. (b) False positives. (c) False negatives.

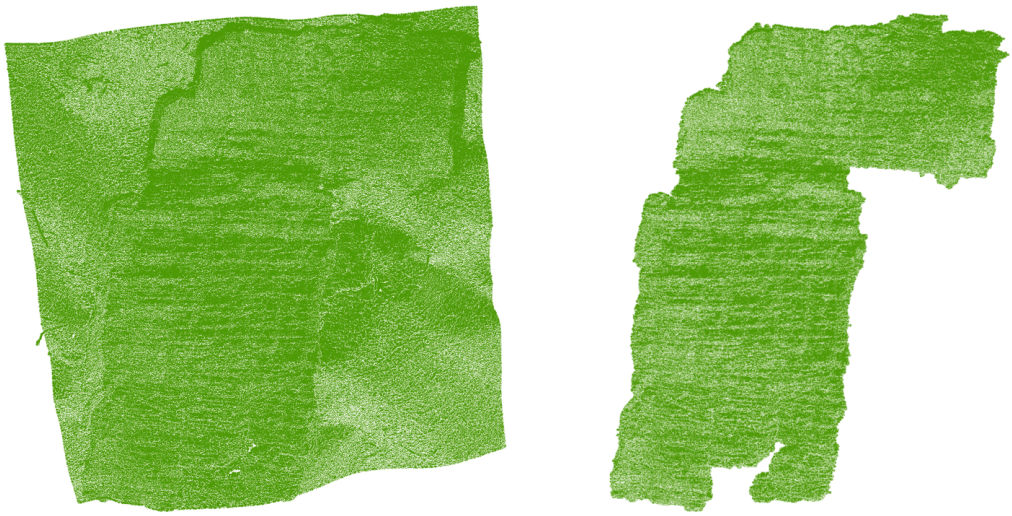
Figure 2.8: Optimal edge detection parameters for a slice of P.Herc.Paris. 2 fr. 47, alongside two failure cases to be avoided. False positive edges generate fuzzy, inaccurate segmentation that follows air above fragment instead of surface itself. False negatives create "holes" in resulting segmented surface.

more easily overcome in downstream steps. It was possible to find adequate tunings for all datasets used in this work, though locally adaptive methods could perhaps make this easier for future datasets.

Mesh processing

The point cloud generated using the Canny segmentation method is dense, and can contain on the order of 10-100M points. This number of points is difficult to visualize interactively or process effectively, and contains more detail than is necessary for a precise surface segmentation. Therefore, the first step in mesh processing is to simplify the raw point cloud. For the fragment datasets used in this work, simplifying the point cloud to ∼100,000 points preserves enough detail while easing downstream steps. Point cloud simplification is performed using the open source Meshlab interface [48], as are the remainder of the steps in this section.

The point cloud also captures features from the CT volume that do not lie along the fragment surface. Typically these are other physical structures present in the scan volume, such as the backing paper upon which the papyrus fragment is mounted. Sometimes other small features appear due to artifacts in the CT images. These are removed manually, by selecting and deleting the extraneous points. Figure 2.9 shows the point cloud before and after this step.

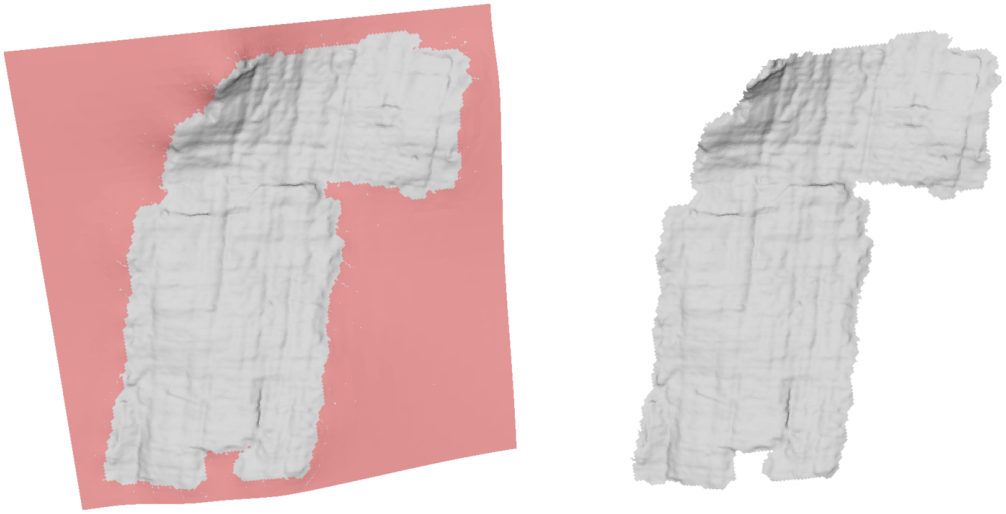


(a) Simplified point cloud. (b) Cleaned point cloud.

Figure 2.9: P.Herc.Paris. 2 fr. 47 point cloud before and after the manual removal of extraneous points. Visualized using Meshlab.

At this stage, depending on the noise in the original CT scan, the surface represented by the point cloud can have varying levels of noise itself. This noise can bias the subsequent meshing step, pulling the segmentation into or away from the the papyrus surface instead of the desired behavior which precisely follows the boundary between papyrus and air. To remove this noise, geometric outlier points are selected and deleted. Interactive visualization during the outlier selection allows one to tune the outlier detection until the intended points are selected. This selection and deletion is often performed two or three times. Sometimes, excellent performance of the initial Canny segmentation can render outlier removal unnecessary. This is typically the result of enabling the contouring option in Canny segmentation, for those datasets which are cooperative.

The processed point cloud is now meshed using screened Poisson surface reconstruction [49]. This reconstruction produces a full sheet that exceeds the fragment bounds, so extraneous faces are selected and removed (Figure 2.10). This selection is either performed by selecting faces with a thresholded Hausdorff distance [50] between the mesh and source point cloud, or by selecting faces with an edge length above some



(a) Initial reconstructed surface. Extraneous faces selected for deletion.

(b) Final mesh after cleaning steps.

Figure 2.10: P.Herc.Paris. 2 fr. 47 segmented mesh before and after final cleaning. Visualized using Meshlab.

threshold. A final cleaning is then performed, consolidating the mesh to a single connected component, closing small holes, and ensuring the mesh is two-manifold.

The resulting mesh precisely follows the exposed surface of the fragment. The 3D surface now resembles the general shape of the fragment as seen in a photograph, and the surface still captures details such as the papyrus fiber structure. The segmentation can be validated by viewing the intersection of the final mesh with the slice images of the original volume. A graphical user interface allows one to quickly validate the segmentation throughout the CT volume by scrubbing through the scan slices (Figure 2.11).

2.3.4 Flattening

Recall that the purpose of this multi-step process is to create an image space to which both the volumetric CT data and the 2D surface images can be aligned. This space is created in the flattening step, by mapping the segmented surface mesh to a 2D image plane. The resulting mesh parameterization defines a bijective mapping, in which points on the mesh surface can be translated between volumetric CT

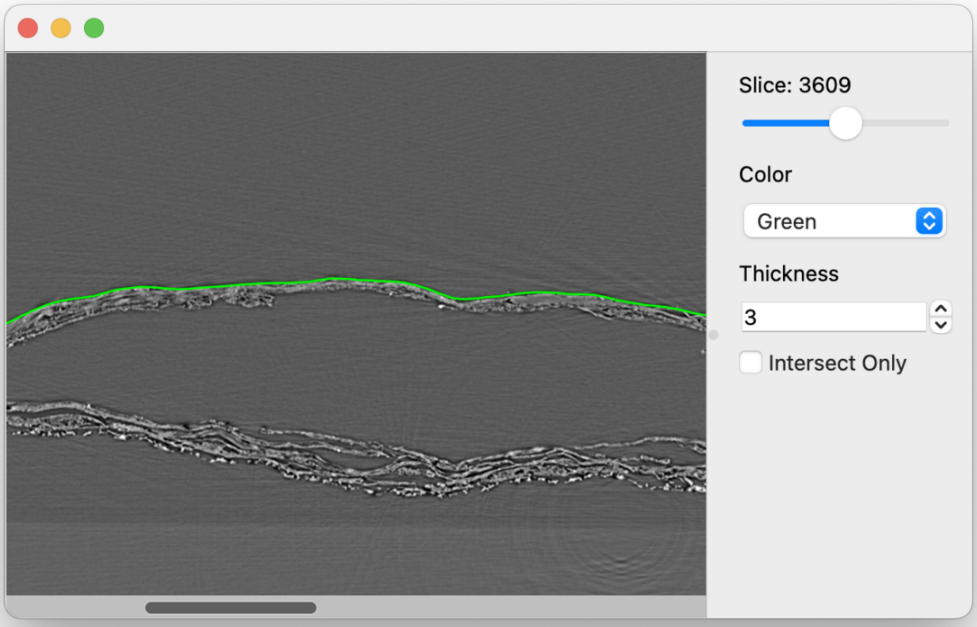


Figure 2.11: Validating the segmentation by viewing the intersection of the final surface mesh with a slice from the original CT volume.

coordinates and 2D image coordinates.

There are various methods to define this mapping between 3D mesh and image plane. Traditionally, the objective of virtual unwrapping is to "unwrap" or "flatten" a convoluted surface so that the text can be more easily read. In that context, the parameterization prioritizes angle preservation in order to avoid distortion. Approaches such as angle based flattening (ABF) [51] and least squares conformal mapping (LSCM) [52] are based on this principle.

For the scroll fragments to be used as training data, the objective is slightly different. The mapping will generate a 2D image to which the spectral photographs will later be aligned. To ease the alignment and minimize the deformations necessary, an orthographic projection is used instead of an angle preserving method.

To compute the orthographic projection, an oriented bounding box is first computed around the 3D mesh. The longest and second longest axes of the bounding box are chosen to define the 2D image plane. The mesh points are projected along the third bounding box axis onto this plane, defining the mapping. Algorithm 2 summarizes this projection method.

The result of the projection is a mesh parametrization mapping the 3D mesh to the 2D image plane. For P.Herc.Paris. 2 fr. 47, Figure 2.12 illustrates the mesh triangles UV mapped to this plane.

2.3.5 Per-pixel map

The mapping between image spaces is bidirectional; that is, it should also be possible to map mesh points in the 2D flattened image space to the 3D volume space. This mapping is implicitly defined as the inverse of the UV map, but it is helpful in later steps to store this explicitly as well. A per-pixel map (PPM) represents this $\mathbb{R}^2 \to \mathbb{R}^3$ mapping. The PPM P maps 2D coordinates to volume space with a mapping of the form:

$$P(u, v) \to (x, y, z, n_x, n_y, n_z) \tag{2.1}$$

Figure 2.12: P.Herc.Paris. 2 fr. 47 mesh triangles mapped to 2D image plane using orthographic flattening (Algorithm 2).

Algorithm 2 The orthographic projection method for mesh flattening. Given a mesh M, computes a UV map U which contains (u_p, v_p) for each vertex $\mathbf{p} \in M$, mapping the 3D point to the 2D image plane.

procedure ORTHOGRAPHIC-PROJECTION(M)
  B = ORIENTED-BOUNDING-BOX(M)
  l_u = NORM(B.axes[0]) ▷ longest axis
  l_v = NORM(B.axes[1]) ▷ second longest axis
  $\mathbf{u} = B$.axes[0]$/l_u$
  $\mathbf{v} = B$.axes[1]$/l_v$
  $\mathbf{o} = B$.origin
  U = [] ▷ empty UV map
  for $\mathbf{p} \in M$ **do** ▷ for each mesh point
    $u_p = (\mathbf{p} - \mathbf{o}) \cdot \mathbf{u}$ ▷ project to image plane
    $v_p = (\mathbf{p} - \mathbf{o}) \cdot \mathbf{v}$
    $u_p = u_p / l_u$ ▷ normalize UV to $[0, 1]$
    $v_p = v_p / l_v$
    U.append((u_p, v_p)) ▷ add to UV map
  return U

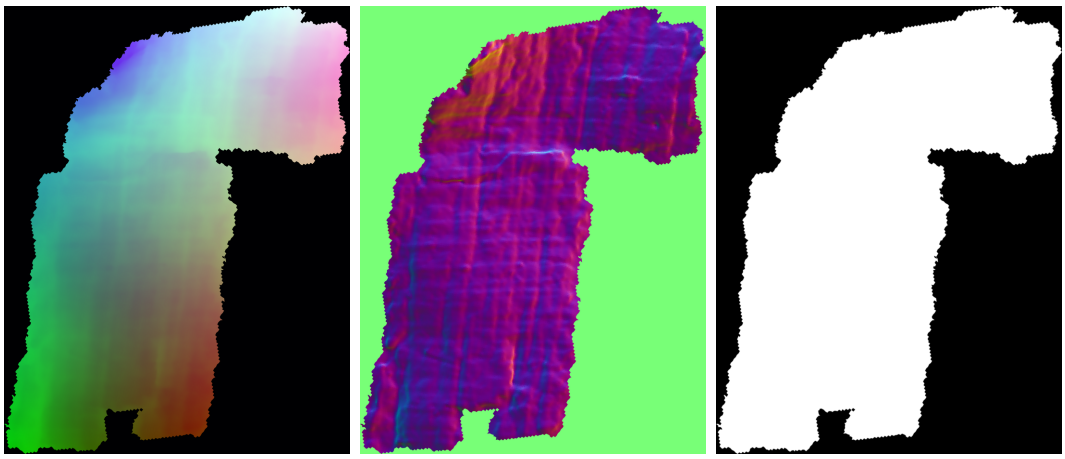
also notated:

$$P(\vec{p}) \rightarrow (\vec{v}, \vec{n}) \tag{2.2}$$

Where $\vec{p}$ is the 2D pixel position in P, $\vec{v}$ is the corresponding 3D coordinate in the CT volume, and $\vec{n}$ is a normal vector indicating the local orientation of the surface at that point in 3D.

The PPM is computed by iterating over the 2D pixels, mapping each one to its mesh triangle (if one exists) based on the UV map, and using the barycentric triangle coordinates to interpolate the surface position and normal vector from the values at the triangle vertices. The PPM is stored as effectively a 6-channel image file sharing the (x, y) dimensions of the flattened image space.

Figure 2.13 visualizes the PPM data for P.Herc.Paris. 2 fr. 47 by mapping the 3D coordinates and the surface normal vector values to RGB channels in a color image. Notably, the visualization of surface normals reveals some texture on the surface, showing the grid-like structure of papyrus fibers.



(a) 3D position. (b) Normal vector. (c) Surface mask.

Figure 2.13: Visualization of P.Herc.Paris. 2 fr. 47 PPM data. (a) 3D position (x, y, z) mapped to RGB channels. (b) Normalized surface normal vector (n_x, n_y, n_z) mapped to RGB. (c) Binary mask indicating surface bounds.

2.3.6 Texture image

A mapping has now been defined that connects 3D and 2D image spaces. To complete the dataset necessary to train a model to learn the presence of ink, the final step is to align label images to this image space. Before the actual alignment of the label images, a fixed image is first generated to which the label images can be registered.

By design, this image occupies precisely the same space as the UV map generated for the segmented mesh. For this reason, it can be used directly as a texture for mesh rendering, and so it is called a "texture image". The texture image plots the CT intensity on the PPM image space, providing a visualization of the segmented surface as seen in X-ray.

The texture image is generated using the existing PPM. The texturing algorithm iterates across the PPM surface, reads the local 3D position and normal vector, and then samples a local neighborhood from the original CT volume. The local neighborhood is filtered to produce a single intensity value which is then plotted on the texture image.

The texture images used in this work all sample a bidirectional linear neighborhood, centered on the point coordinates and oriented along the normal vector as read from the PPM. The length of this neighborhood is set to equal the estimated material thickness, which unless otherwise specified is set to 150µm for all texture images in this work. The actual thickness of papyrus sheets varies considerably, but this has been an effective estimate for this purpose.

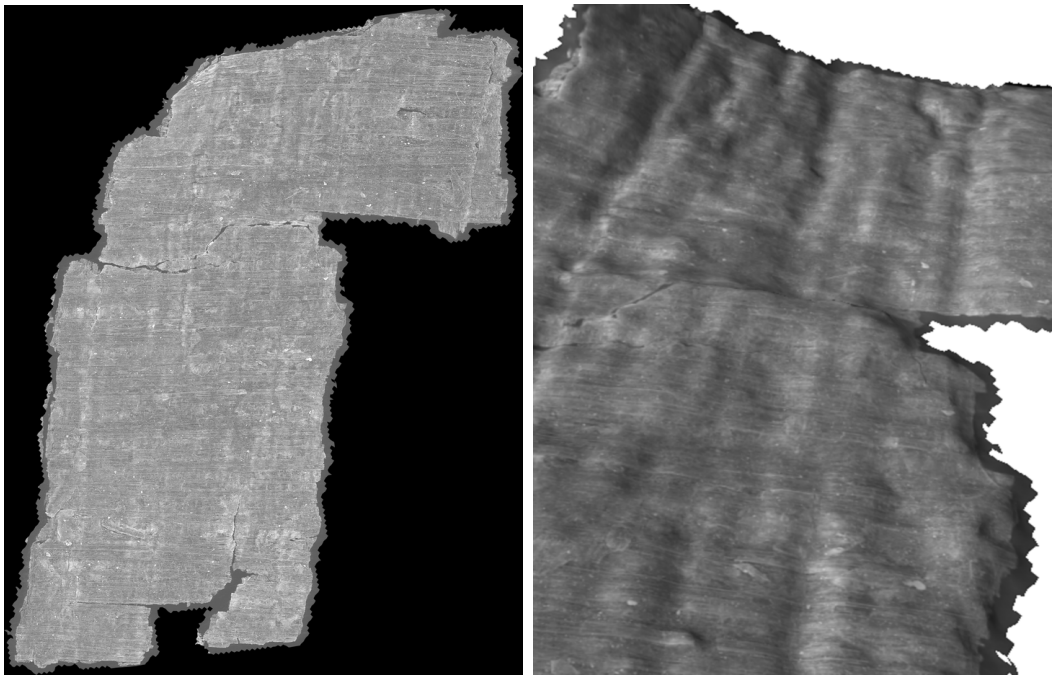
A max filter is then used, which simply takes the maximum value from this neighborhood and plots it. The texturing methods used are consistent with prior work [21].

For those objects with clear X-ray contrast between ink and substrate, the texture image reveals the text and the pipeline is complete. Of course, the lack of said contrast is the motivation behind the pipeline presented here. Multiple simple filters have been tried in an attempt to extract visible ink contrast, so far without success. In this case, rather than revealing text, the function of the texture image is to serve as the fixed image for label image registration.

Figure 2.14 shows the texture image generated for P.Herc.Paris. 2 fr. 47. As mentioned, this can be used directly as a mesh texture for 3D rendering, also shown. Any of the images subsequently registered to the texture image can therefore also be utilized in 3D rendering using the existing UV map.

2.3.7 Label alignment

With the texture image generated, it is now possible to align the label images, pairing labels with the CT data. As mentioned, infrared photographs are selected as the label photographs for their increased ink contrast. This problem of alignment, known as image registration, involves identifying one image as the "fixed" image and another as the "moving" image. In this case, the texture image is fixed, as it is tied to the generated UV map, and the infrared image is moved. A deformable warp (as opposed to a simpler transformation such as affine or projective) is necessary to



(a) Texture image. (b) Textured mesh (detail).

Figure 2.14: Texture image for P.Herc.Paris. 2 fr. 47.

accommodate the subtle differences in image geometry, as one was acquired using a lens and the other was generated by an orthographic projection from CT.

Image registration is an established area of research, with many successful algorithms using feature- or intensity-based matching, in addition to other approaches [53]. Feature- and intensity-based methods both struggle with the multimodal nature of this particular image registration problem. The ink appearing in one modality and not the other is a significant challenge. In addition, those features that do appear in both images (typically papyrus fibers) have a significantly different appearance in each. Due to these challenges, an automated method that is capable of successfully aligning the texture and infrared images has not yet been discovered.

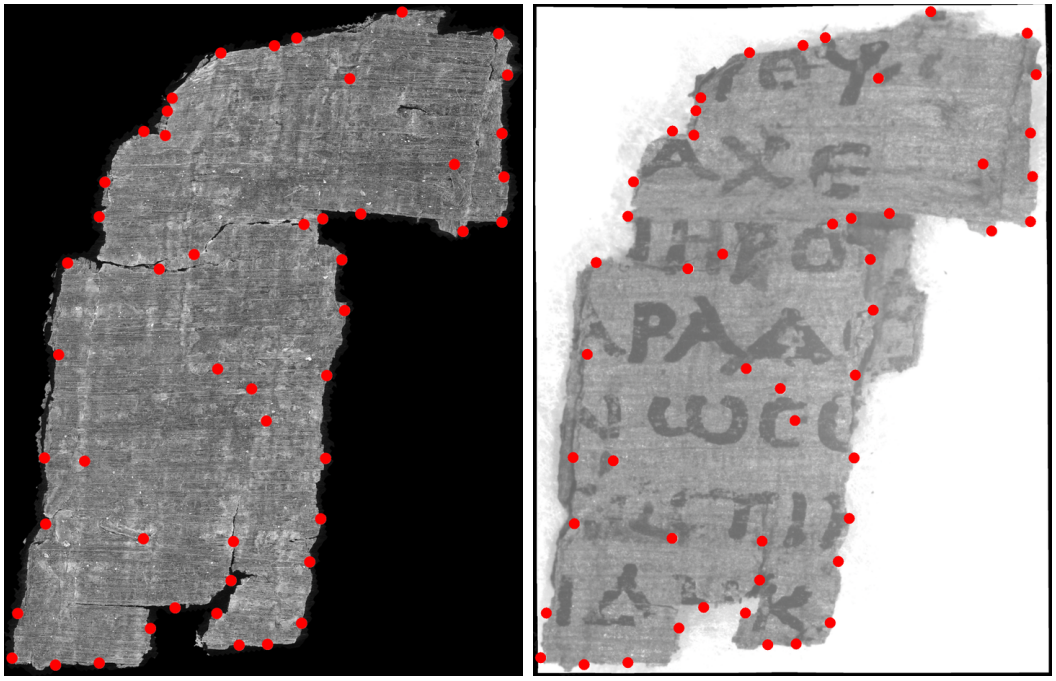
Instead, these images are aligned manually by identifying visual feature points that are common to both images. Photoshop's Puppet Warp allows the user to specify an arbitrary number of points and their desired locations, warping the image using a deformable mesh in order to align all reference points. Figure 2.16 shows the reference

Figure 2.15: Infrared image before registration (alignment).

points used when aligning P.Herc.Paris. 2 fr. 47. Figure 2.17 shows details of some of these points, illustrating the sorts of image features that are commonly used.

Manual feature point matching is feasible for a dataset of this scale, but is time consuming and introduces human error. An automated method would be preferable. Though the difference in modalities is significant and an automated method has not yet been developed for this registration problem, learned methods may be able to overcome the challenges. Some related work has used the small crack structure on painting surfaces for registration [54], and a similar approach leveraging the papyrus fiber structure could be a promising direction.

It is also possible that the registration process could be improved by using a different filter when generating the texture image. The max filter used in this work is adequate, but other filters may generate texture images with features more resembling what is visible in infrared.



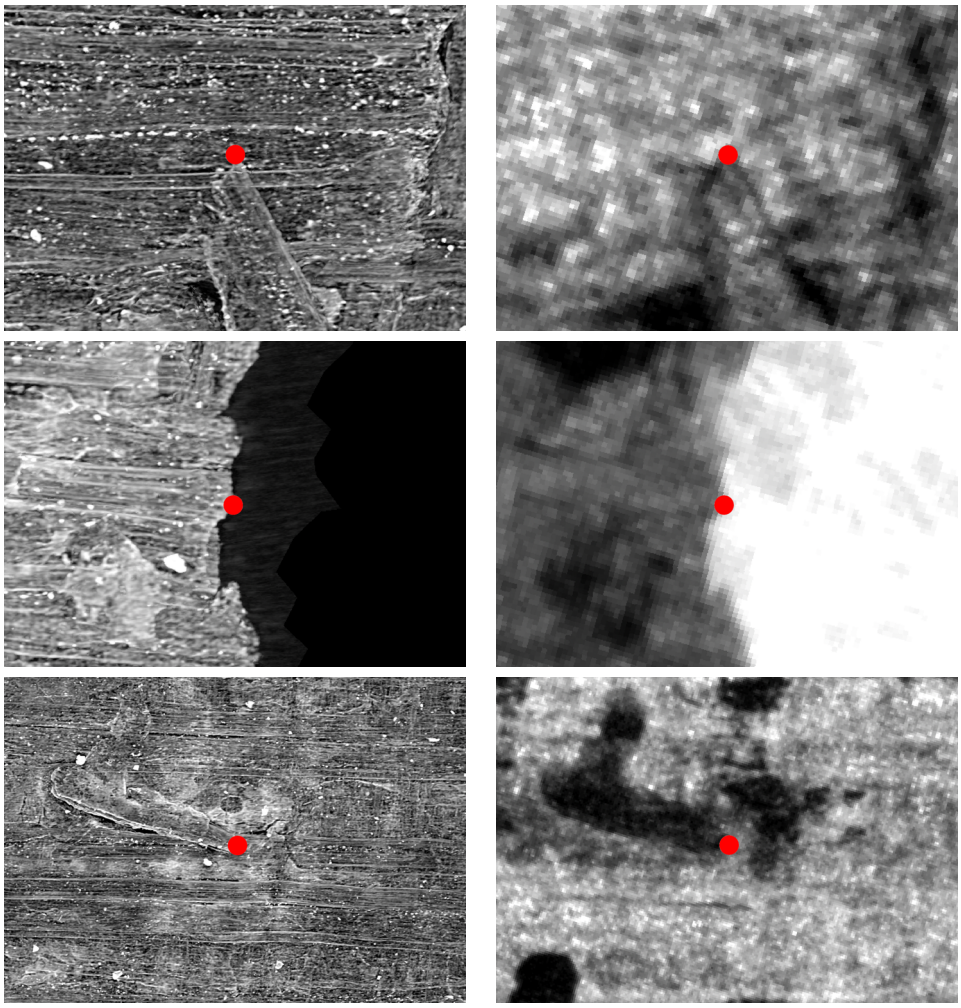
(a) Texture image. (b) Infrared image.

Figure 2.16: Manually identified alignment points used in the registration process for P.Herc.Paris. 2 fr. 47.

2.3.8 Binary ink labels

Following alignment, the CT and infrared photograph now comprise a dataset suitable for supervised machine learning. For a given pixel in the PPM space, a model can take a neighborhood of CT as input and learn to predict the appearance of that pixel in infrared. Sometimes other tasks beyond the direct prediction of the infrared image are desired.

Binary ink classification is one such task that is sometimes preferable. Classification is helpful when training models, as it is a well established task with stable conventions for loss functions and other parameters. For this task, an additional labeling step is included, where a binary segmentation mask is created by tracing the ink regions in the infrared image. This is performed manually using the Quick Selection tool in Photoshop, under the supervision of a papyrologist to help disambiguate difficult spots based on papyrological context. The binary label image resulting from this



(a) Texture image. (b) Infrared.

Figure 2.17: Details of some example alignment points. Contrast stretched to enhance details.

Figure 2.18: Binary ink labels for P.Herc.Paris. 2 fr. 47.

process is shown in Figure 2.18.

For particularly difficult spots, additional tools can be utilized to try to determine whether ink is present. Viewing the corresponding CT slice alongside the point of interest in the surface photograph can help to disambiguate some of these cases, by providing cues as to whether a dark region in the image comes from ink on the surface or is instead a shadow or a hole. Shown in Figure 2.19, a graphical interface was developed to view these images side by side, highlighting where a slice image intersects the PPM. This can also visualize any of the images aligned to the PPM, in this case the infrared image. By changing the slice index until the intersection crosses

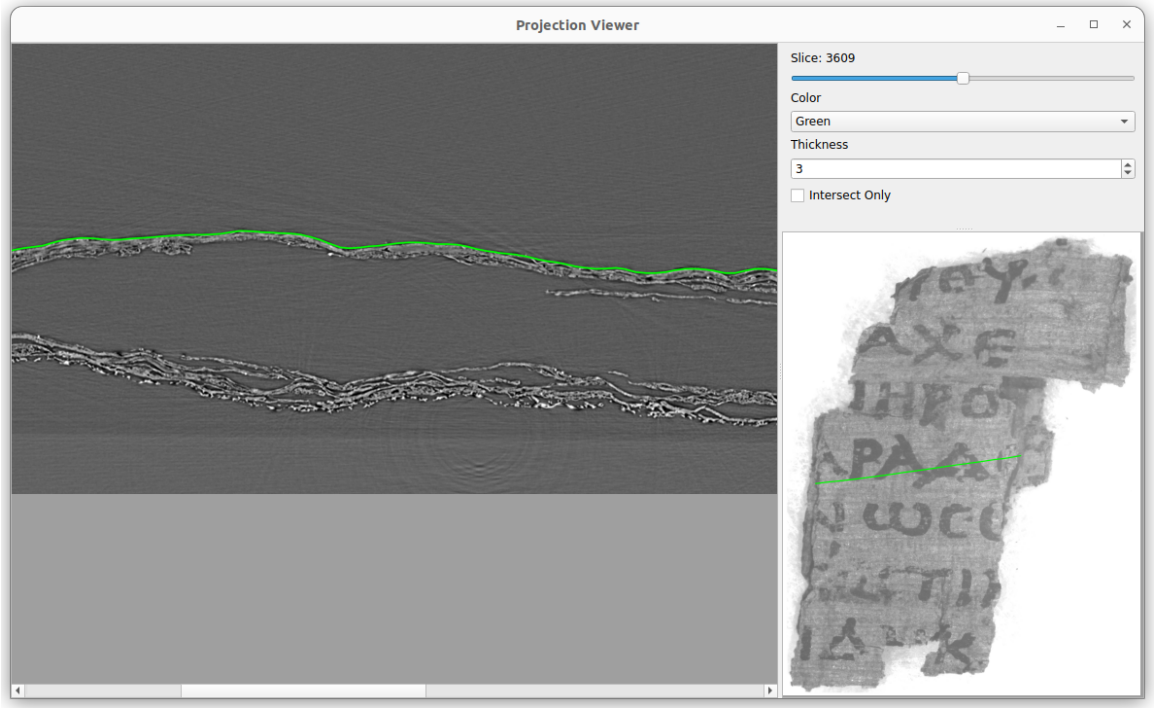


Figure 2.19: User interface developed to examine points of interest on the surface alongside their corresponding CT slices. Common use case is to determine whether a spot on the surface is a hole or could be ink.

the point of interest in the infrared, one can examine the corresponding slice image to see if a dark spot is a hole in the surface, suggesting it is not ink, or otherwise.

Despite these steps, the creation of a binary mask does introduce some labeling error. The complex distribution of ink on the surface is reduced in this step to a binary signal. In some places, where the presence of ink in the photograph is particularly difficult to discern, these ambiguous spots may lead to mislabeling even with expert labeling. This is explored further in Section 3.4. For these and other reasons to be discussed, binary ink classification is not likely to be the primary task used in the future. That said, the aforementioned benefits are enough to make binary classification a useful tool for the development of methods, and as such it is used frequently in this work.

2.4 ink-ID

Section 2.3 outlined the theoretical geometric framework that is the basis of the ink detection method in this work. The framework provides a dataset that serves as the starting point for supervised learning, with which models can be trained to detect the ink presence. This section discusses ink-ID, the reference implementation of the above framework, including its design choices and baseline methods for training and inference.

ink-ID is a software framework developed to test whether a learned model can outperform the standard texturing filters in the detection of subtle signals such as carbon ink. ink-ID can train a neural network to detect the presence of ink, and can also use the trained model during inference to generate "prediction images," displaying the model's ability to reveal ink or the desired signal.

ink-ID performs three primary functions:

- Section 2.4.1: Dynamic data generation (inputs and labels)
- Section 2.4.2: Machine learning (training and prediction)
- Section 2.4.3: Dataset management (cross validation splits, visualization, and evaluation)

The remainder of this section will discuss the design of these components in more detail.

2.4.1 Dynamic data generation

The geometric framework outlined above enables the generation of (input, label) pairs for machine learning, where the inputs are portions of the CT volume and the labels come from the aligned surface imagery. This section discusses how ink-ID leverages the PPM mapping to perform this data generation.

ink-ID navigates two image spaces: the 2D image space generated from virtual unwrapping, and the 3D image space of the volumetric CT scan. As mentioned in

Section 2.3, these spaces are linked with a bidirectional mapping associating the 3D surface mesh with a 2D parametrization. One direction of the mapping is defined by the UV map of the surface mesh, and the other is defined by the per-pixel map (PPM). In addition to defining the mapping back to 3D, the PPM itself also occupies the 2D flattened image space itself. The PPM can therefore refer both to the mapping, and to this 2D image space. It is this PPM P (Equation 2.1) that serves as the primary image space that ink-ID navigates. Training and prediction routines iterate over the pixels of P, using these as starting points from which inputs and labels are sampled.

Inputs

ink-ID generates 3D neighborhoods of CT data to be used as inputs to the neural network. These 3D arrays, referred to as subvolumes, are sampled dynamically during training and inference using the following process.

For a given (u, v) pixel $\vec{p}$ on the PPM image space, the PPM P is first sampled directly:

$$P(\vec{p}) \rightarrow (\vec{v}, \vec{n}) \tag{2.3}$$

This sampling yields the CT volume coordinate $\vec{v}$ and the surface normal vector at that position $\vec{n}$. A subvolume is then sampled from the CT volume centered on $\vec{v}$ and oriented with respect to $\vec{n}$, visualized in Figure 2.20.

There are multiple options which together specify the details of how a subvolume is sampled. The size and shape of the subvolume, in both physical units and in voxels, are among the most important (Figure 2.21).

The size in physical units (Figure 2.21a) can be thought of as the local neighborhood that is required in order to determine the presence of ink. If the carbon ink is not represented in CT by an intensity change, but instead as a morphological pattern, as appears to be the case, then a single voxel or very small neighborhood will not be sufficient to detect the ink. Instead, a larger neighborhood or region of support is necessary. The optimal size of this neighborhood is not known a priori, instead,

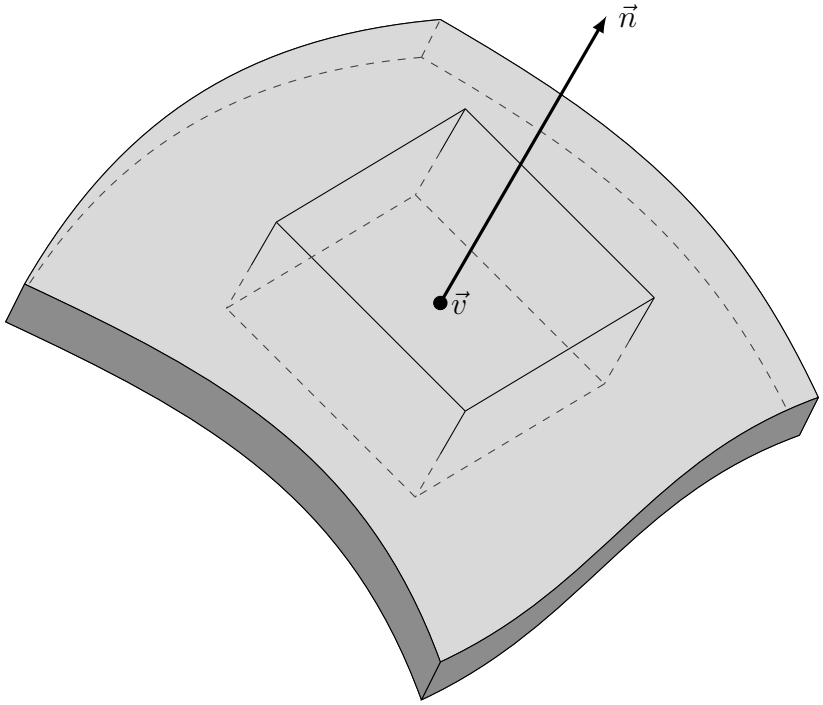
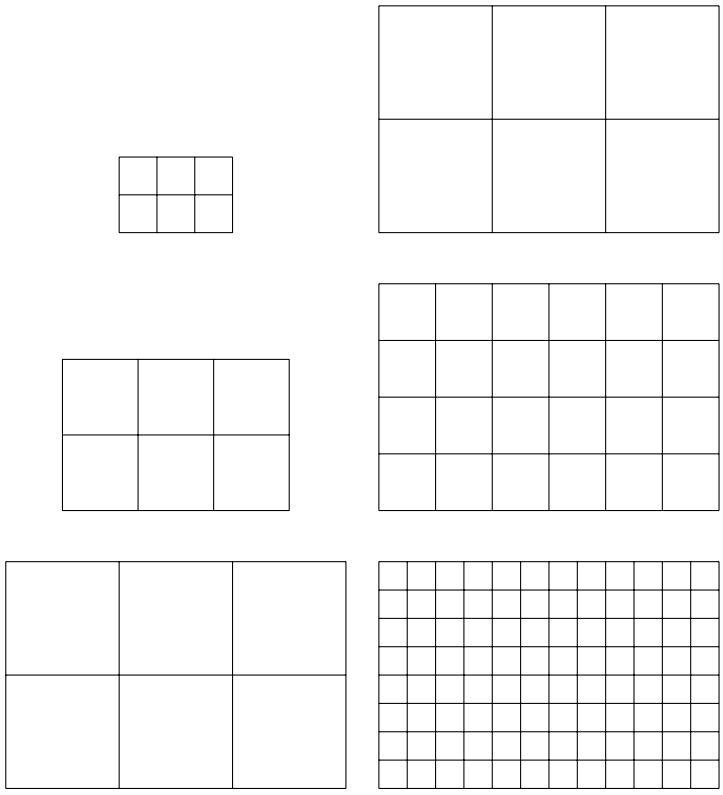


Figure 2.20: Dynamically sampling a subvolume from the CT volume, centered at 3D position $\vec{v}$ and oriented to align with the surface normal vector $\vec{n}$. Idealized papyrus surface shown in gray.

different sizes are tried and the results are measured empirically. The subvolume also does not need to have the same dimensions in each axis: flatter subvolumes penetrate less into the papyrus surface, but capture more lateral context, while deeper subvolumes do the opposite. The best settings often depend on the context of the physical properties of the scanned material as well as the characteristics of the CT scan. These settings are treated as hyperparameters and are explored further in Section 6.5.

In addition to the spatial extents of the subvolume, the sampling rate is configurable (Figure 2.21b). This determines the size of the array that ends up being sampled. Typically, this sampling is set automatically to be 1:1 with respect to the CT volume resolution, such that it is not oversampling or undersampling the original CT image. There are however occasions when it is helpful to alter this sampling, for example when working simultaneously with multiple CT volumes of different resolutions and



(a) Spatial extents. (b) Sampling rates.

Figure 2.21: 2D generalization of the difference between altering subvolume sampling by spatial extent (a) or by sampling rate (b).

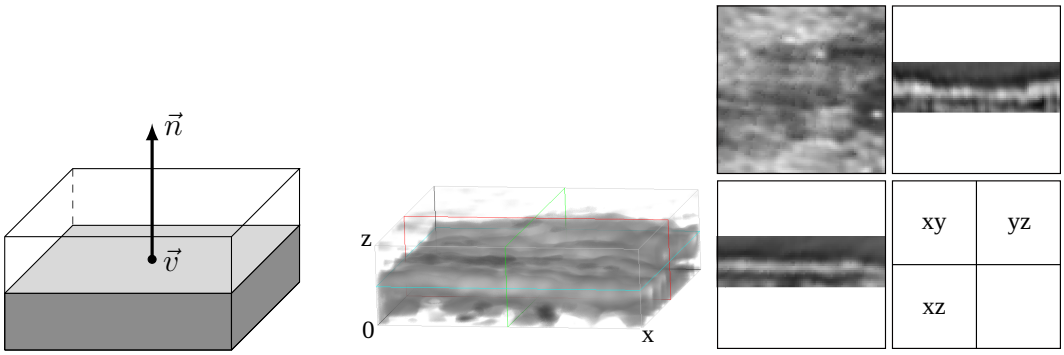
wishing to sample subvolumes with consistent spatial extents.

Another option when generating a subvolume is to first move the subvolume origin $\vec{v}$ along the normal vector $\vec{n}$ by some amount prior to sampling. This can be set to a constant value applied to all subvolumes, useful in cases where the segmented mesh does not follow the exact surface (for example as shown in Figure 2.4b).

More commonly, rather than a set value, the movement along $\vec{n}$ is uniformly sampled from a configurable range. This option, called "jitter" in ink-ID, is used as a form of data augmentation to artificially move the subvolumes with respect to the segmented surface prior to sampling. By effectively worsening the quality of the segmentation, the ink detection model learns to be more robust to segmentation error, and can detect ink presence even when the papyrus surface is not perfectly aligned through the center of the subvolume. As with the above hyperparameters, this is explored further in Chapter 6.

After determining the subvolume's position, orientation, spatial extents, and sampling rate, the resulting voxel positions must be individually sampled from the CT volume. Since the subvolume center $\vec{v}$ and orientation $\vec{n}$ are not constrained to have integer values, the resulting sample points typically do not fall on integer volume voxel positions but instead must be interpolated. Trilinear interpolation and nearest neighbor were both implemented, and as they had similar results, nearest neighbor was chosen as the default for performance purposes.

Figure 2.22 shows some different visualizations of a subvolume sampled using the above methods. This is the input to the neural network, described later, from which the presence of ink will be determined. Ideally, the subvolume always captures the writing surface itself, as well as some depth into the writing surface and some distance into the air above the surface. When segmentation is precise, this results in a volumetric input that contains roughly half papyrus and half air, with the surface that would contain the ink running through the middle of the subvolume.



(a) Idealized 3D view. (b) 3D volume render. (c) Orthogonal slice planes.

Figure 2.22: Subvolume visualizations, showing idealized rendering alongside two views of the same subvolume from the surface of P.Herc.Paris. 2 fr. 47.

Inputs: implementation details

ink-ID samples subvolumes dynamically from the CT volume during training or inference, rather than precomputing subvolumes and storing them on disk. This is atypical for computer vision and machine learning, in which preprocessed image inputs are often read directly from disk during training and inference. Image operations performed during training or inference are typically limited to simple data augmentation operations such as crops, rotations, flips, and intensity changes, all of which have optimized implementations that frequently utilize the graphics processing unit (GPU).

The dynamic subvolume sampling in ink-ID is an expensive operation. The subvolumes are also sampled from a large CT volume, which must be held in random access memory (RAM) during program execution. As is discussed later, the CT volumes are sometimes prohibitively large, and become the limiting factor against moving forward with experiments using larger datasets. It would therefore be desirable to avoid not only the expensive dynamic subvolume sampling, but also to avoid loading the entire volumes into memory. This section will mention a considered alternative, before providing an overview of the ink-ID implementation.

One alternative would be to precompute and sample the subvolumes and labels,

storing them on disk as individual files. This way, during training or inference, the model could simply read the files as if they were standard image files, read the corresponding labels, and proceed. This more resembles the standard I/O pattern seen in computer vision and machine learning. No complex geometric computations are necessary for each subvolume, which needs only to be read from disk and then optionally augmented using standard methods. Further, the program running training or inference only needs to load the current batch of subvolumes into memory, and can discard them once complete. Based on this arrangement it would be possible to precompute subvolumes from many volumes at once, store them together on disk, and train across this entire dataset.

This method was not implemented for a few reasons. Primarily, the subvolume shapes and sizes are frequently adjusted, as they are an important part of the hyperparameter search space for ink detection models. Precomputing and storing a large set of subvolumes, only to have to recompute them every time one wishes to try another size, shape, or sampling method, is unwieldy enough to be prohibitive.

In addition, due to the high degree of overlap between the subvolumes from adjacent pixels in the PPM, the generated data would be highly redundant, increasing the data size on disk. In fairness, the data size would also *decrease* due to only storing the regions in the volume which lie on the segmented surface of interest. Rough calculations from a handful of real datasets revealed that the two effects mostly cancel each other out, so there is no real benefit with respect to disk space.

The inefficiency of sequentially sampling individual voxels is overcome by using compiled routines for the performance critical functions. ink-ID is primarily implemented in Python for quick development and interoperability with machine learning libraries such as PyTorch [55], but uses Cython [56] for improved performance when sampling subvolumes. The Cython language is a superset of Python, allowing type declarations using C types, and Cython functions can therefore be compiled to C. The

compiled functions are easily called from Python code, allowing one to write Python-like syntax that results in dramatically improved performance for certain operations.

The improved performance from Cython is sufficient to reduce the sampling compute time to less than that of training or inference, removing this particular performance bottleneck. The relative performance varies based on the subvolume sizes and other details of the implementation or experiment, but in one case selected as an example, the Cython implementation reduced the time necessary to sample a single $24 \times 80 \times 80$ subvolume (153,600 voxels) from 10.9s to 2.5ms, a 4,360x speedup.

The choice to dynamically sample the subvolumes during training and inference has the above benefits, but does result in the challenge of loading the entire CT volume into memory before starting the job. Effectively, this means that the system RAM on the machine running ink-ID must exceed the total size of all volumes to be loaded during a job. For small volumes this is not a problem, but at high resolution the individual volumes frequently exceed 100GB and in some cases exceed 1TB. Those numbers multiply when running jobs that run across more than a single volume.

The machine running ink-ID must meet the above RAM requirements in addition to having GPU(s) for the neural network operations. While single machines meeting both of these requirements do exist, they are somewhat unusual and are expensive. Many of the experiments in this work either operate on a single volume at a time, or rely on the following changes to the geometry to reduce the memory requirements.

Simplifying the geometric framework

As it has so far been described, the geometric framework is powerful but has some drawbacks. First, as implemented in ink-ID, the entire CT volume for each dataset had to be stored in memory during training or inference, preventing experiments across multiple full fragments.

Second, while the PPM defines a mapping between the 2D and 3D image spaces, the CT inputs and surface labels are not themselves registered to one another. Processing

them jointly relies on the pointwise mapping defined by the PPM, and there is not a natural way to use other sampling methods common to image-to-image domains such as aligned image patches.

Finally, the PPM itself is often multiple GB, a nontrivial increase to the already tight memory requirements.

Surface volumes

To alleviate these pain points, "surface volumes" were devised to sample thinner volumes from the original CT volumes, based on the segmented surface meshes. A surface volume contains the CT data intersected by a segmented mesh, as well as the neighboring CT data within a configurable thickness. This surface volume is sampled ahead of time and stored on disk as a set of image slices, so that during training or inference only the surface volume is required to be loaded into memory. Figure 2.23 shows a conceptual rendering of a surface volume being sampled, and Figure 2.24 shows a volume rendering of the actual surface volume sampled from P.Herc.Paris. 2 fr. 47.

The surface volume is sampled such that its X-Y slices are inherently aligned with the PPM image. The aligned image labels will therefore also be registered to the CT images directly. The PPM can then be discarded, having served its purpose. Figure 2.25 illustrates change to the geometric framework overview with surface volumes, where the volumetric data is directly aligned to the label images in X-Y.

Algorithm 3 shows the surface volume sampling method. The algorithm iterates over the PPM P, and for each point reads the PPM at that (u, v). This provides the 3D coordinate $\vec{v}$ and normal vector $\vec{n}$ of the same point in 3D, which are used to sample a linear neighborhood about that point in 3D using trilinear interpolation. This line fills the depth of the surface volume S for one (u, v). When repeated for each pixel, the result is a surface volume containing the CT data surrounding the segmented mesh. For the surface volumes in this work, the sampling radius was set

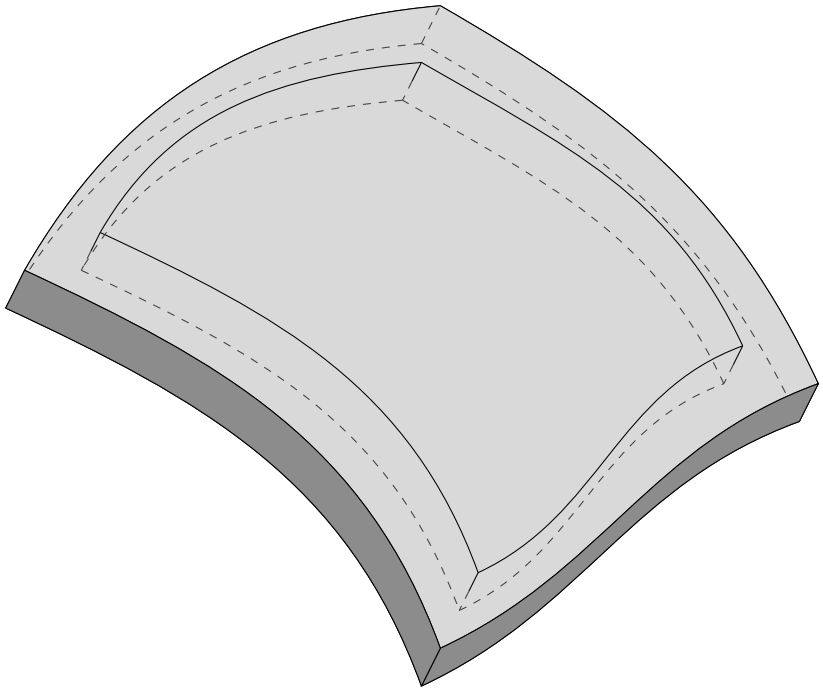


Figure 2.23: Sampling a surface volume from the CT volume. Idealized papyrus surface shown in gray. Compare with subvolume sampling in Figure 2.20.

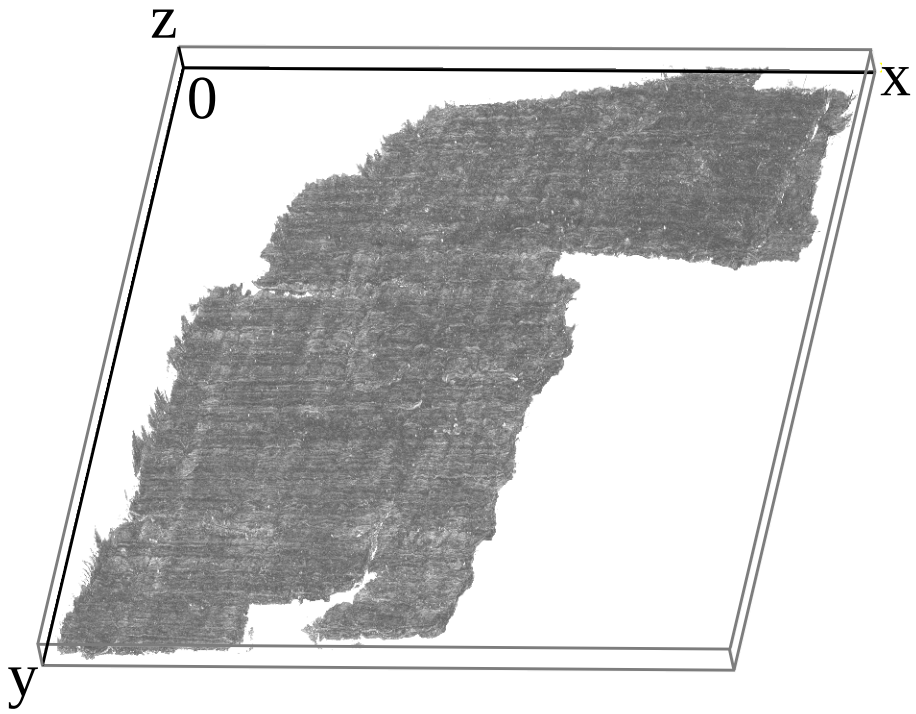


Figure 2.24: Surface volume of the segmented surface of P.Herc.Paris. 2 fr. 47. Volume rendering in Fiji/ImageJ.

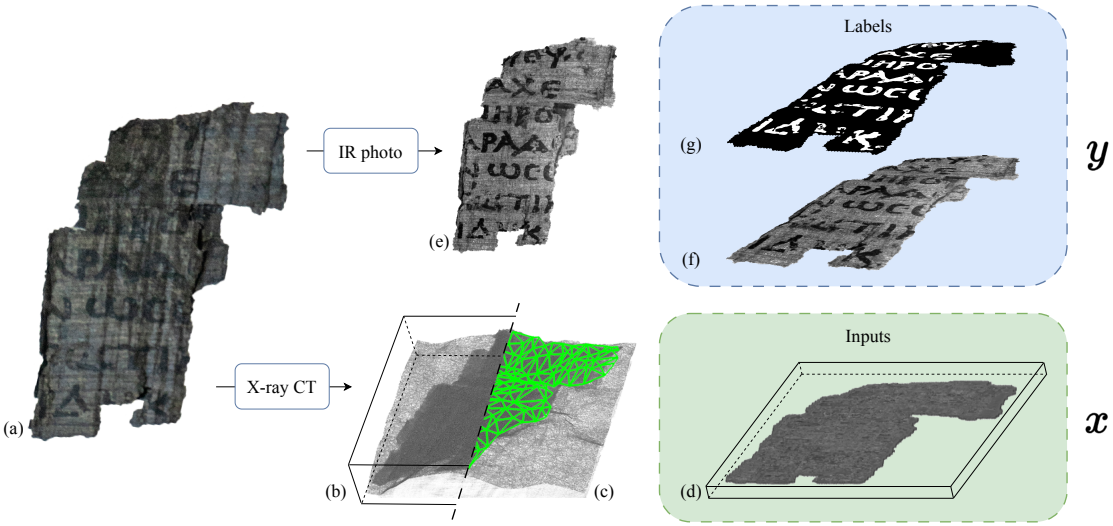


Figure 2.25: Geometric framework with addition of surface volumes, visualized with P.Herc.Paris. 2 fr. 47. (a) Scroll fragment, RGB photograph. (b) Volumetric X-ray CT image. (c) 3D surface segmentation. (d) Flattened "surface volume" sampled about the segmented surface mesh. (e) Infrared photograph. (f) Infrared photograph aligned to surface volume. (g) Aligned binary ink labels.

to $r = 32$ (creating a surface volume of 65 channels) and the interval $\delta = 1$ was chosen to preserve the original image resolution.

Generating surface volumes is a known compromise, as it introduces slight aliasing effects where there is curvature on the mesh surface. In our experiments, this does not impact the model's ability to learn to identify the presence of ink. There are other approaches that would alleviate some of the above pain points without introducing aliasing, such as storing volumes as 3D chunks on disk and using caching to limit the memory usage. Initial implementations of these methods have been prohibitively slow, but this is a matter of engineering and will likely be addressed in the near future.

Whether loading the entire CT volume into memory or using surface volumes, the purpose of data generation is the same. Subvolumes are sampled as the inputs to the machine learning steps to follow. The model learns to identify the presence of ink in these subvolumes during training.

Algorithm 3 Surface volume generation. Given a PPM P, CT volume V, sampling radius $r \in \mathbb{N}$, and interval $\delta \in \mathbb{R}_{>0}$, returns the sampled surface volume S.

procedure SURFACE-VOLUME(P, V, r, δ)
  $d = 2r + 1$ ▷ surface volume depth
  $S = \text{ZEROS}((d, P.\text{rows}, P.\text{cols}))$ ▷ init. empty surface volume
  for $v_i \in [0, P.\text{rows} - 1]$ **do** ▷ iterate PPM X-Y
    for $u_j \in [0, P.\text{cols} - 1]$ **do**
      $(x, y, z, n_x, n_y, n_z) = P[v_i, u_j]$ ▷ read PPM
      $\vec{v} = [x, y, z].\text{T}$ ▷ 3D coordinates
      $\vec{n} = [n_x, n_y, n_z].\text{T}$ ▷ normal vector
      for $z_k \in [-r, r]$ **do** ▷ iterate Z (depth)
        $\vec{s} = \vec{v} + z_k \delta \vec{n}$ ▷ compute 3D position
        $v = \text{INTERPOLATE}(V, \vec{s})$ ▷ sample
        $S[z_k + r, v_i, u_j] = v$ ▷ store
  return S

Labels

Computing the labels is comparatively straightforward. For a given $\vec{p} \in P$, the same pixel location $\vec{p}$ in the label image L (Figure 2.18) is sampled. Using binary classification as an example task, a binary label is generated and associated with a subvolume:

$$L(\vec{p}) \rightarrow \{0, 1\} \tag{2.4}$$

The (input, label) pair is now ready to be used in supervised learning. The goal is to train a model that can look at a 3D piece of CT data and predict the presence of ink there.

The label generated in this process corresponds to a single point. So that the model can use spatial context, a subvolume around that point is generated as the input. The label however only corresponds to the central voxel of the subvolume, and can only be interpreted as the presence of ink there, regardless of the presence or absence of ink elsewhere within the subvolume.

These pointwise or pixel-wise labels result in models that can only operate in a pointwise fashion. The task is not optical character recognition (OCR) or any similar

task that would rely on a more complete view of written characters. Inference is also performed on individual PPM pixels independently, so if a trained model generates a prediction image revealing the shapes of characters, this is due purely to the ink presence having those shapes, not because the model has learned to hallucinate written characters.

2.4.2 Machine learning

Having generated the subvolume inputs and associated labels, ink-ID proceeds to its second primary function, which is to perform machine learning on batches of these pairs, whether training or inference. Though training and inference routines ostensibly comprise the overarching purpose of ink-ID, their actual implementations are the simplest of ink-ID's three objectives outlined in Section 2.4. Data generation (Section 2.4.1) and dataset management (Section 2.4.3) are ultimately ink-ID's more significant contributions, compared to the relatively off the shelf components used in the neural network models at the core of the implementation.

Models in ink-ID are implemented using the PyTorch library [55] for deep learning, providing builtin, optimized neural network components, and enabling the easy use of hardware accelerators such as GPUs. This choice is specific only to the reference implementation provided, and is not necessary to implement the ideas of ink-ID. For example, early versions of ink-ID used TensorFlow [57] and performed identically; the transition to PyTorch was only for usability improvements, many of which are likely unnecessary at the time of this writing based on the rapid development of both libraries.

The standard classification model in ink-ID takes a subvolume as input and outputs a binary classification corresponding to the presence of ink at the central voxel of that subvolume. Figure 2.26 shows the convolutional [58, 59] architecture of this model for an single-channel input subvolume of dimensions $24 \times 80 \times 80$, a typical size (see

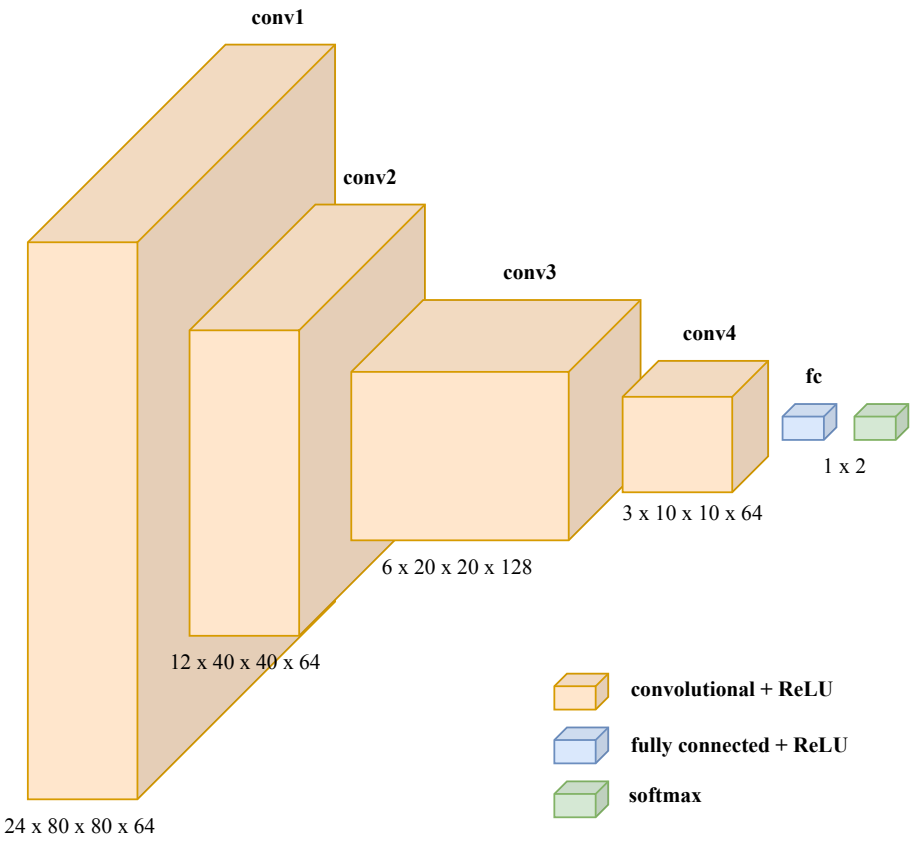


Figure 2.26: Network architecture for default binary ink classifier with input subvolume size $24 \times 80 \times 80 \times 1$.

Section 6.5). The model is trained using binary cross entropy loss:

$$l(x, y) = \frac{\sum_{n=1}^{N} l_n}{N}, \quad l_n(x_n, y_n) = -\sum_{c=1}^{2} \log \frac{\exp(x_{n,c})}{\sum_{i=1}^{2} \exp(x_{n,i})} y_{n,c} \tag{2.5}$$

where x is the input, y is the target, and N is the batch size.

ink-ID implements a number of model architectures for different purposes. The model and input size in Figure 2.26 are generally representative, and continue the P.Herc.Paris. 2 fr. 47 example from the preceding sections. Other models are described later in this work as the need for them arises to achieve different objectives.

2.4.3 Dataset management

ink-ID also performs higher level management of the images and labels in order to guide the training process and visualize the inference outputs.

PPM masks

Every (u, v) pixel $\vec{p} \in P$ is considered as a seed point, each of which will generate an (input, label) pair. The points are first filtered by the PPM mask, as the surface itself is typically irregularly shaped and also smaller than the PPM image P, which bounds the surface in a rectangular box. The PPM mask P_M (Figure 2.13c) is a separate binary image specifying which pixels are considered part of the surface:

$$P_M(\vec{p}) \rightarrow \{0, 1\} \tag{2.6}$$

The filtered set of seed points S is then defined by $S = \vec{p} \in P \text{ s.t.} P_M(\vec{p}) = 1$.

Cross validation

Ultimately, the goal of ink-ID is to train on all available training data for a given collection. In the case of the Herculaneum scrolls, this means some fragments with exposed text would have their surfaces imaged with photography and would also be scanned in CT. The model would be trained on all of these, and the purpose of doing so would be to then perform inference on the hidden layers inside the intact, rolled scrolls. There is no ground truth for these hidden layers, so it is a challenge to evaluate the model's performance on them. Instead, for the proof of concept and for the experimental development of the methods, a spatially-aware version of k-fold cross-validation is used.

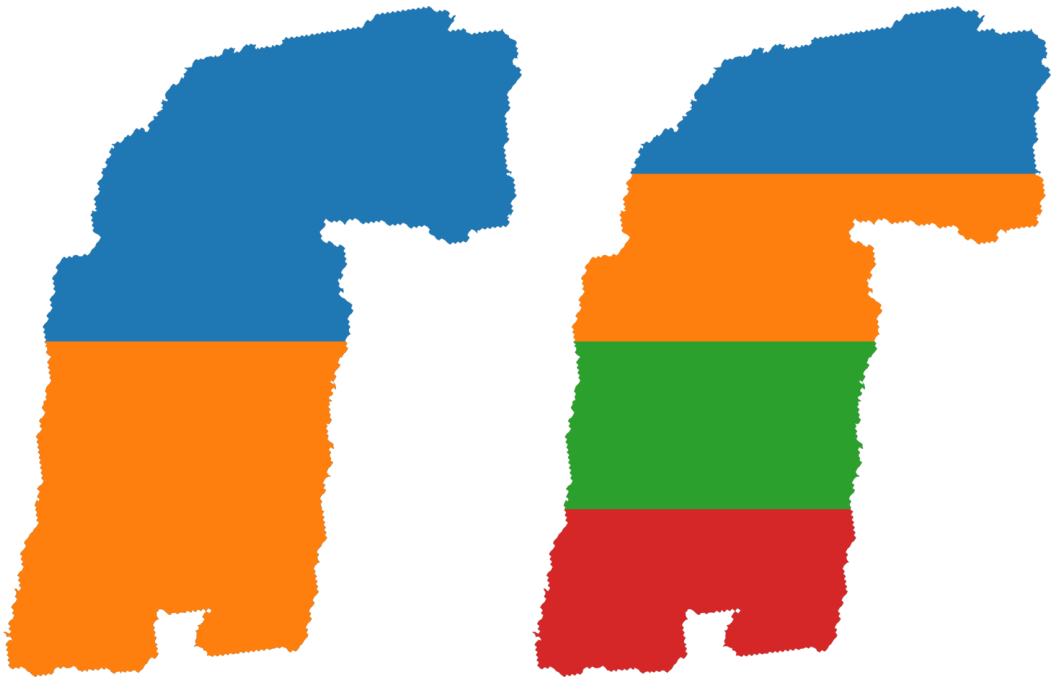
Traditional k-fold cross-validation uses uniformly sampled random splits of the training set S to create k subsets of equal size. This validation split is inappropriate for ink-ID, where adjacent pixels in the PPM yield significantly overlapping subvolumes. Even if a model has not seen a subvolume from the exact pixel being sampled,

subvolumes from neighboring pixels are similar enough that memorization is still a concern.

One solution to this problem would be to use a training dataset of k scroll fragments, using individual fragments as the cross-validation splits. The model could train on all but one, and then be evaluated on the held out fragment, repeated across the dataset. This approach was not used initially due to the practical memory constraints that prevented ink-ID experiments across multiple fragments simultaneously.

Instead, a spatially-aware version of k-fold cross-validation is used, where the PPM surface for a single fragment is divided into non-overlapping rectangular "regions". The regions are then assigned to training, validation, and test sets. In this arrangement, memorization is avoided within the region interiors, and model performance can be evaluated more effectively. This region-based approach continued to be helpful even after the introduction of experiments across multiple fragments simultaneously, as will be seen throughout this work.

Figure 2.27 illustrates two example cross-validation region splits for P.Herc.Paris. 2 fr. 47. The fragment surface is split into a top and bottom half in the first example, or four rows in the second example. The regions do not need to be equally sized image rows as they are here, but can be arbitrary rectangular bounding boxes. It is however often convenient to split the surface into roughly equal regions in a grid pattern, so a script is provided to split the surface into a desired number of rows and columns. Since the bounding boxes are also masked by the irregular PPM mask, the resulting cross-validation splits are not of identical size, but in practice this does not prevent the cross-validation method from providing valuable feedback on model performance. For each produced region, a region JSON file is generated which specifies the PPM file, the surface mask, the corresponding CT volume, the region bounding box in the PPM space, and other metadata such as any associated label images. A text file is also generated by the script, creating one dataset of the generated grid cell regions,



(a) 2-fold. (b) 4-fold.

Figure 2.27: Example region-based cross-validation splits for P.Herc.Paris. 2 fr. 47.

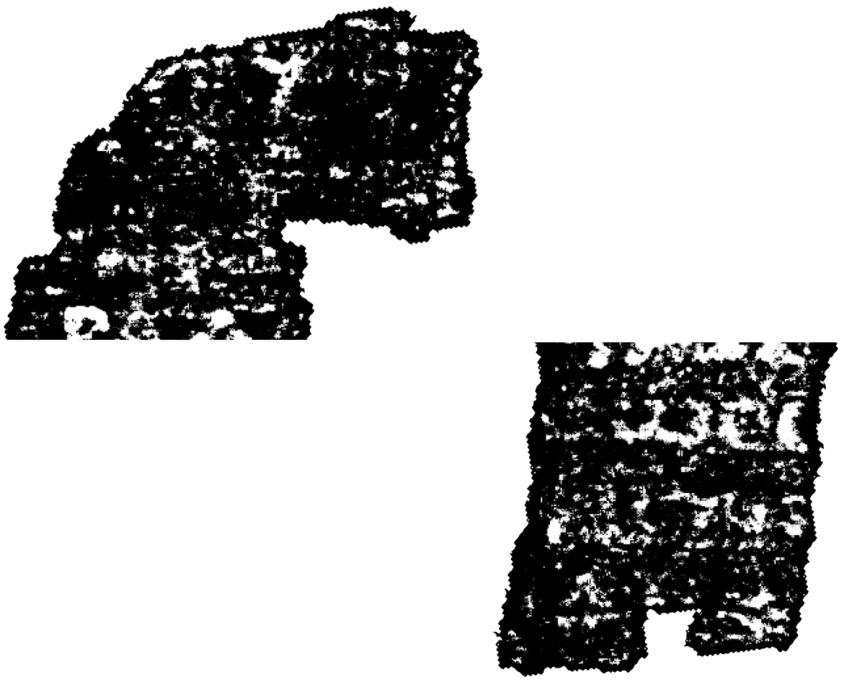
which can be iterated over to perform k-fold cross-validation.

Prediction images

ink-ID also manages the generation and organization of "prediction images," the images visualizing the output of the inference process. Figure 2.28 shows examples of such images generated from an ink-ID experiment with the 2-fold cross-validation split of P.Herc.Paris. 2 fr. 47 illustrated in Figure 2.27a. The model is trained on one of the regions, generates an ink prediction on the other, and vice versa.

Similarly to the training process, prediction images are generated during inference by sampling pixels from the PPM and then fetching their corresponding subvolumes. Unlike during training, shuffling the pixels is not necessary, so they are processed sequentially. For each pixel, the subvolume is sampled and then fed to the trained network. In the binary classification example shown here, the resulting [0,1] probability for the ink class is mapped to [0,255] and then plotted directly on the image.

Many quantitative evaluation metrics for binary classification require binarized or



(a) Region 1/2. (b) Region 2/2.

Figure 2.28: Example prediction images generated from an ink-ID experiment with the 2-fold cross-validation split of P.Herc.Paris. 2 fr. 47 illustrated in Figure 2.27a.

thresholded outputs. As discussed in Section 3.3, quantitative metrics are used in ink-ID to provide objectivity. The true metric of interest, however, is in reality the legibility of the text revealed by the detected ink. As this does not map perfectly to any particular metric, visual inspection of the generated prediction images is crucial. Prediction images are therefore not binarized, to retain as much visual detail as possible from the predicted ink probabilities. This is true during the development process and also of the ultimate desired output of ink-ID, which is an image generated for a scholar showing as much detail as possible about the ink presence. As thresholding only removes information, this is not necessary and is not performed.

As mentioned above, visual inspection is often preferred over quantitative metrics for evaluating a model's ability to reveal legible ink presence. It is therefore frequently helpful to visualize the model output not only after training, but also throughout the training process. This is visually equivalent to examining plots of various metrics

to show their changes over the course of training. To this end, ink-ID periodically pauses training and generates prediction images for the regions in the prediction set. These images are saved such that after training, one can see how the model converged throughout the training process. To ease this visualization, the intermediate prediction images are aggregated sequentially in an animated .gif image that is automatically generated for each ink-ID job. These animations not only show the final visual performance of a given model, but also provide useful information about the training process at a glance. Quick inspection of these animations often gives clear insight into model stability and convergence patterns that is not evident in the quantitative plots or in the separate examination of individual prediction images.

Generating prediction images can take a long time due to the high number of pixels in the PPM image, the resolution of which is based on the resolution of the original CT volume. Due to the high-resolution CT images, PPM images often have widths and heights on the order of tens of thousands of pixels, resulting easily in hundreds of millions of individual predictions necessary to create a prediction image if each pixel is processed. To address this, ink-ID does not typically generate a prediction for every single pixel in the PPM, but instead samples pixels from a sparse grid of configurable cell size. For example, by sampling only every 16^{th} pixel in x and in y, the number of samples is reduced by 256. To ease later processing steps, prediction images are generated at full resolution, but when using the grid sampling the generated predictions are simply plotted as larger rectangles (e.g. a 16×16 pixel square following the above example) instead of single pixels. The grid spacing is configurable in ink-ID and the optimal value depends on the dataset and visualization priorities, though 16 is the default and is commonly used.

In practice, this grid sampling during prediction image generation saves time and also has no discernable impact on the generated images. Imperfections in the segmentation, alignment, and labeling steps lead to labels that are not pixel perfect but

instead contain some noise. The model therefore already does not learn pixel perfect outputs, so even prediction images generated at full resolution are not perfectly sharp. Effectively blurring the prediction image by a small amount using the sparse sampling therefore has no impact on a prediction that is already slightly blurred. As discussed in Section 6.4.2, this is not a limitation of the model, but of the datasets as they have so far been constructed. For datasets with perfect label alignment, ink-ID is capable of very sharp prediction images. In this scenario, it is worthwhile to spend the extra compute time in order to achieve the sharpest images possible.

During the experimentation and development of this work, it was necessary to run many training jobs and examine their output. This can be a surprisingly laborious process, despite the expectation that one could have a single metric value for each experiment that would tell the story of its performance. Consistently, it was helpful to spend more time looking at the visual results throughout training in addition to the plots of various metrics. This is made somewhat challenging by the frequent use of k-fold cross-validation, leading to many trained models and prediction images for a single experiment.

To ease the review process as much as possible, ink-ID also has a script to compile the results of an experiment in a small set of visuals. Figure 2.29 shows an example of the output of this script. The generated "summary image" shows the label image alongside the prediction images. This allows one to intuitively establish the expected scale of the text in the output, and to visually validate the predictions for correctness. Other metadata is also added to the image, such as the training iteration (or "final" if training has completed).

The prediction images from the regions used for cross-validation can be visualized separately, as in Figure 2.29, which provides an accurate visualization of the dataset split used in the experiment. In this visualization, the models have also generated predictions of their respective training regions after the job has completed, and these

Figure 2.29: Example output of summary images script run after ink-ID experiment, in this case a 2-fold cross-validation experiment across P.Herc.Paris. 2 fr. 47. Memorization is evident where models have been asked to predict on their respective training regions (outlined in red).

are outlined in red. Memorization is evident here, showing that the models are capable of fitting the training data very well and highlighting the need for cross validation for evaluation. Another visualization is also generated, combining all regions from the same PPM into the same image space (Figure 2.30). This summary image is helpful to quickly obtain a single view of the effective model performance across the fragment surface.

As mentioned above, it is best for visualization to preserve the model output probabilities rather than reducing the information by binarizing the output using some threshold. In some cases it is also helpful to increase the visual contrast of the plotted probabilities, instead of relying always on a linear grayscale color map. The summary image script in ink-ID therefore generates output images using a variety of color maps, so one can review them quickly to see if one more clearly reveals the desired information. Figure 2.31 shows some examples of the generated images using different color maps. In all cases, including the linear grayscale color map, a color map legend is included indicating which colors indicate predicted ink and predicted ink absence.

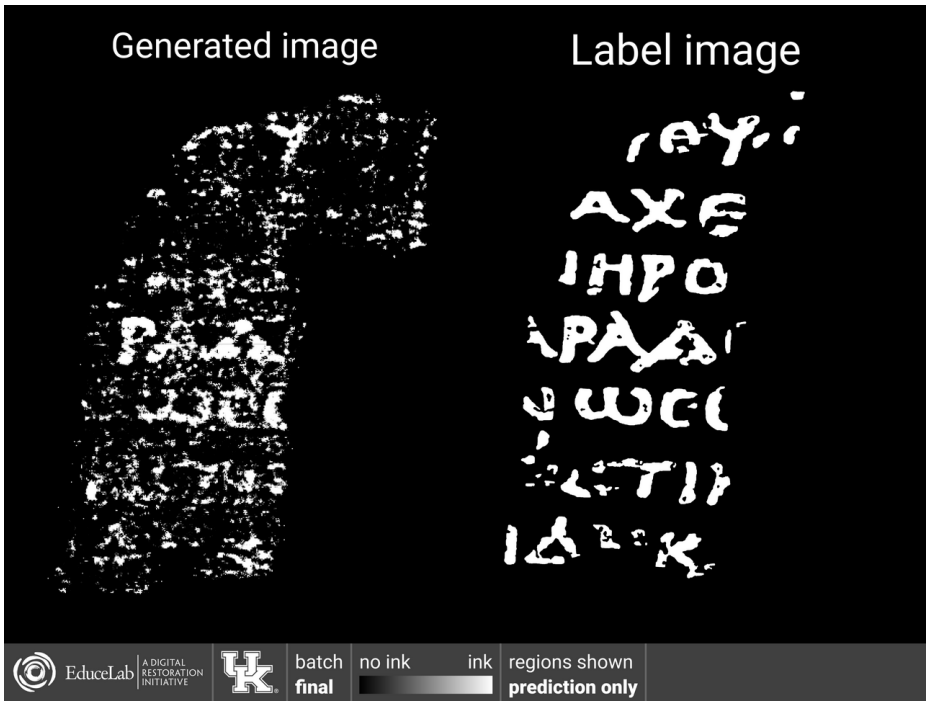
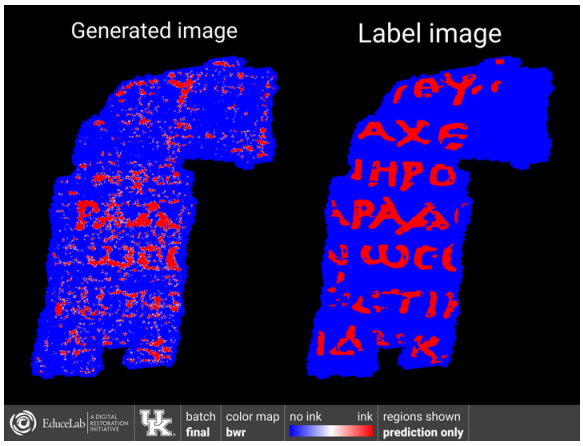


Figure 2.30: Summary image from example ink-ID experiment (Figure 2.29), now with regions from the same PPM plotted in the same space to give a composite view of the network's predictive ability across the dataset.

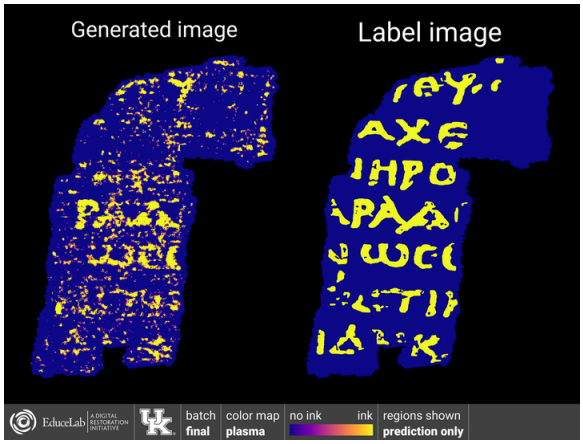
As mentioned above, these summary images are also animated, with individual frames created using the prediction images generated throughout the training process. The example summary images here have so far shown only regions from a single PPM, but they are also generated when multiple surfaces are involved, as will be shown later.

2.5 Summary

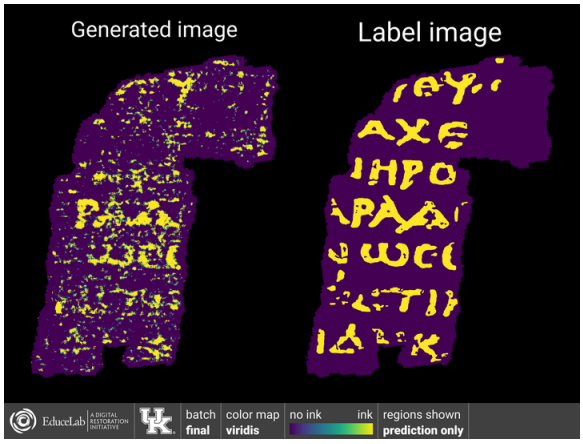
This chapter has presented a novel geometric framework linking volumetric images with 2D surface imagery of the same objects. This framework enables the creation of labeled image datasets for supervised learning, allowing the detection of patterns in volumetric images that otherwise seem invisible. The reference implementation of the machine learning component of this method was also presented. Called ink-ID, this library implements dynamic data sampling, training and inference routines, and dataset management and visualization. The next chapter will show the experimental results enabled by this method, including for the first time the successful recovery of



(a) “bwr”.



(b) “plasma”.



(c) “viridis”.

Figure 2.31: Summary images generated using different color maps to enhance visual contrast.

text from Herculaneum scroll surfaces using X-ray CT images.

CHAPTER 3. INK DETECTION: RESULTS

3.1 Introduction

The chapter outlines the experimental path taken to develop this method, which follows a sequence of proofs of concept, ascending towards the ultimate objective of revealing the hidden texts of the intact scrolls. This path leads to the following contributions:

- **Experimental results**: the successful recovery, for the first time, of Herculaneum ink from X-ray CT.
- **Evaluation**: techniques for the verification and evaluation of ink detection, including visual, quantitative, and papyrological methods.
- **Exploratory data analysis**: inspired by the experimental results, and using the tools developed in Chapter 2, a nuanced view of carbon ink in X-ray CT is presented.

3.2 Proofs of concept

The geometric framework in Section 2.3 and implementation in Section 2.4 were developed using a series of proofs of concept. The experiments have been designed to probe the general hypothesis that carbon ink presence can be captured in X-ray CT, and to explore its boundaries. This process begins with a lab-made proxy scroll, designed to remove some other challenges of the real materials and focus purely on the ink detection problem using the easiest case possible. Experiments then proceed to real Herculaneum scroll fragments of increasing sizes.

3.2.1 The Carbon Phantom

The Carbon Phantom is a lab-made proxy scroll, designed specifically to test whether a model may be able to detect the presence of carbon ink where it is not otherwise easily observed in X-ray CT images. Shown in Figure 3.1, the Carbon

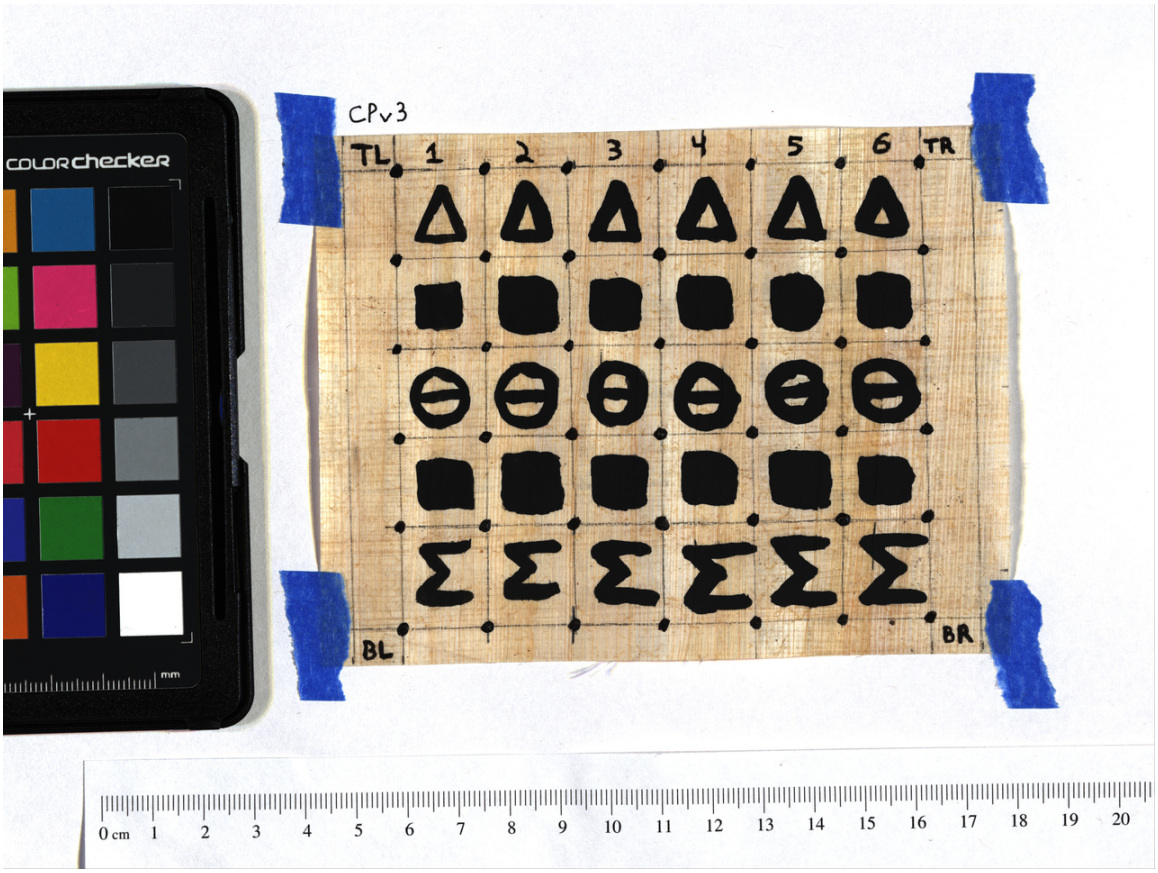


Figure 3.1: The Carbon Phantom proxy scroll, photograph.

Phantom has six columns of large Greek characters and squares, written with carbon ink. The number at the top of each column indicates the number of layers of carbon ink that were applied to the characters in that column, so the leftmost column has a single ink application and the rightmost column has six applications. Column six therefore has the thickest ink layer, which should have the strongest signal for detectability in CT. If a model were unable to reveal the presence of ink in this column, there is no reason to believe it would be able to detect the more subtle ink signal of the Herculaneum scrolls.

The smaller characters and grid dots are written with iron gall ink, so that they appear clearly in CT. These fiducial markers are used as alignment reference points for the labeling process. Figure 3.3 shows the result of running the Carbon Phantom through the Virtual Unwrapping pipeline after being imaged in X-ray CT at 12µm

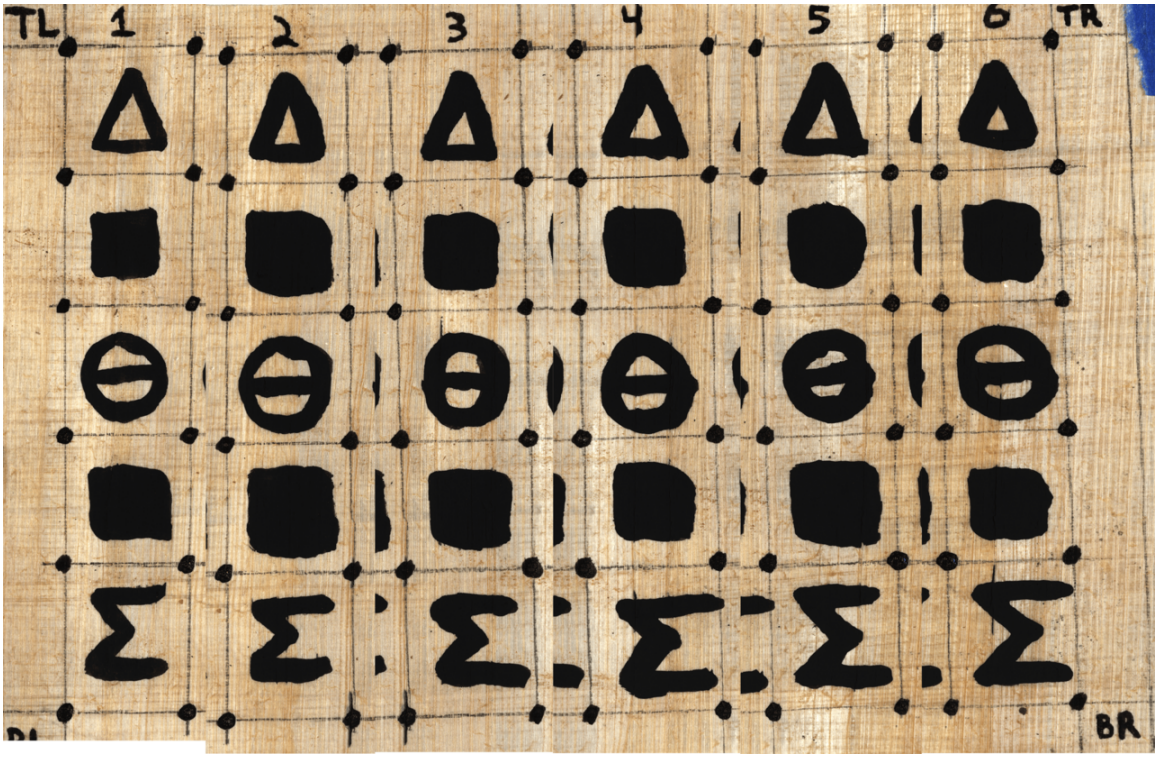


Figure 3.2: Carbon Phantom columns in color photography, processed independently and concatenated horizontally.

voxel size. This pipeline is as described in Section 2.3, except a midpoint based texturing method [21] was used, as shown in Figure 2.4b. The small characters and dots written in iron gall show up clearly, while the larger carbon ink characters are less apparent.

In the rightmost columns, where the carbon ink has been applied in up to six layers and is very thick, some of the carbon ink characters begin to faintly appear in the texture image. This effect does not continue, however, for the more leftward columns with fewer layers of carbon ink. By experimenting with different filters in the texturing phase, one can generate images that highlight slightly different features, but none clearly reveal the carbon ink.

ink-ID was used to see whether the trained model could highlight the ink presence. Many experiments were tried in order to understand the boundaries of ink-ID's predictive ability on this dataset. Figure 3.4 shows a representative experiment, which is

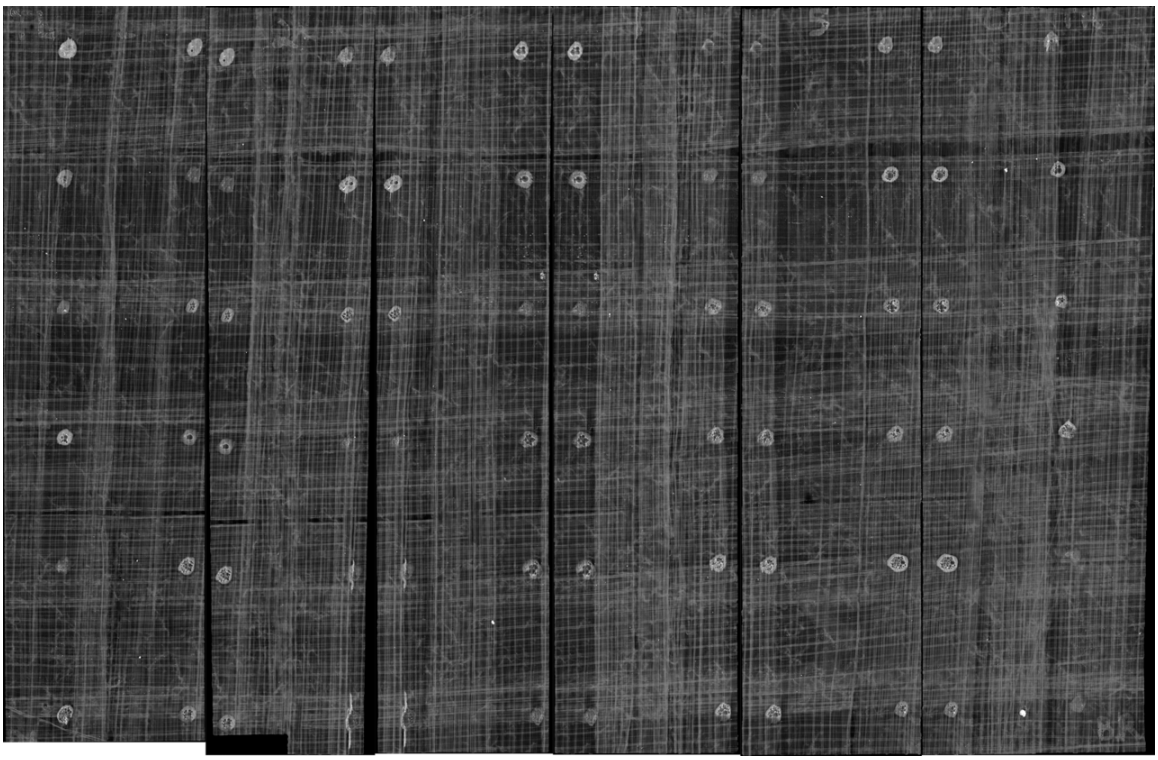


Figure 3.3: Carbon Phantom columns after traditional virtual unwrapping from X-ray CT, showing iron gall ink clearly but not the larger carbon ink characters.

the combined output of 30 independently trained models. Each column was treated as a separate job, and regions were defined around the five carbon ink characters in the column. A 5-fold cross-validation job was then run, training five separate models based on the single column. The figure shows this process repeated for each column, with the results concatenated horizontally to resemble the spatial layout of the Carbon Phantom.

Clearly, ink-ID is capable of accurately recovering the majority of the carbon ink on the Carbon Phantom. ink-ID performs well in all but the first column. This aligns with the morphology hypothesis: if a single layer is approximately 5µm thick, the CT scan voxel size would need to approach that order of magnitude to have a chance at capturing the ink signal. The Carbon Phantom was imaged with a 12µm voxel size for these experiments, close to the expected thickness of a doubled ink layer.

The Carbon Phantom is a promising initial proof of concept, showing that much

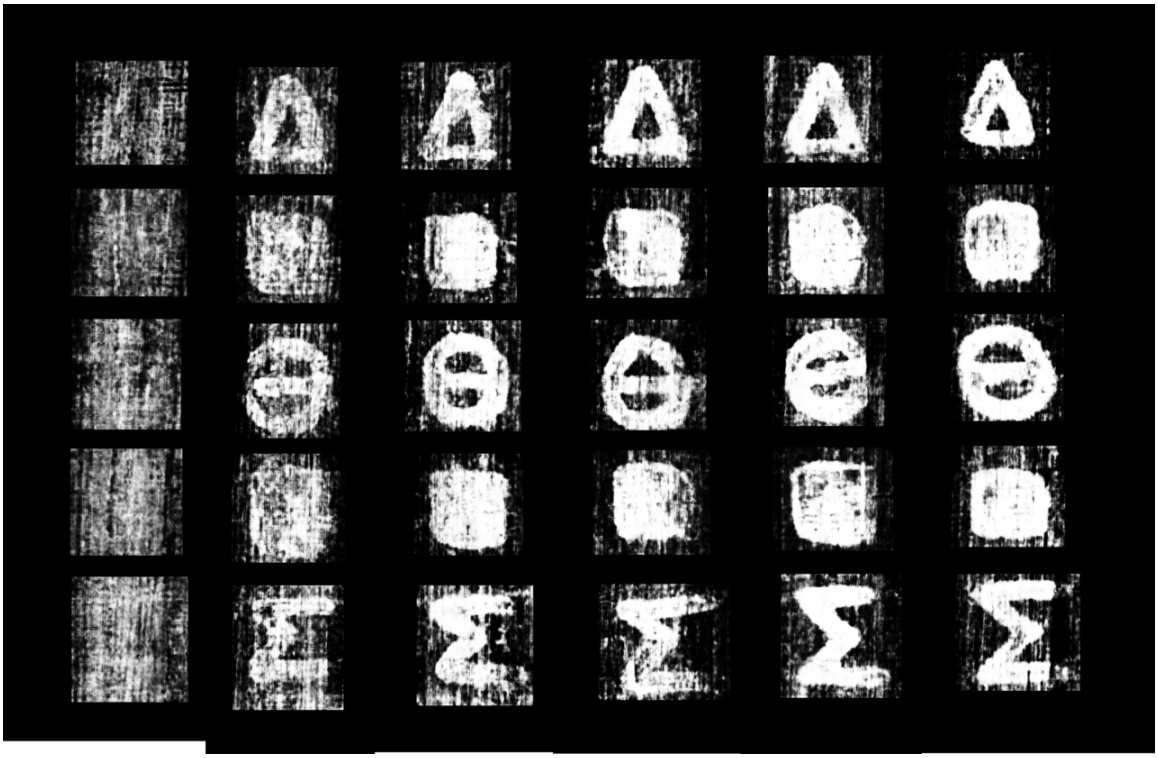


Figure 3.4: ink-ID result on the Carbon Phantom. Each column was a separate 5-fold cross validation experiment across the five characters of that column. Composite image therefore represents the output of 30 independently trained models.

(a) Color photograph. (b) Infrared photograph. (c) Binary ink labels.

Figure 3.5: P.Herc.Paris. Objet 59 reference images.

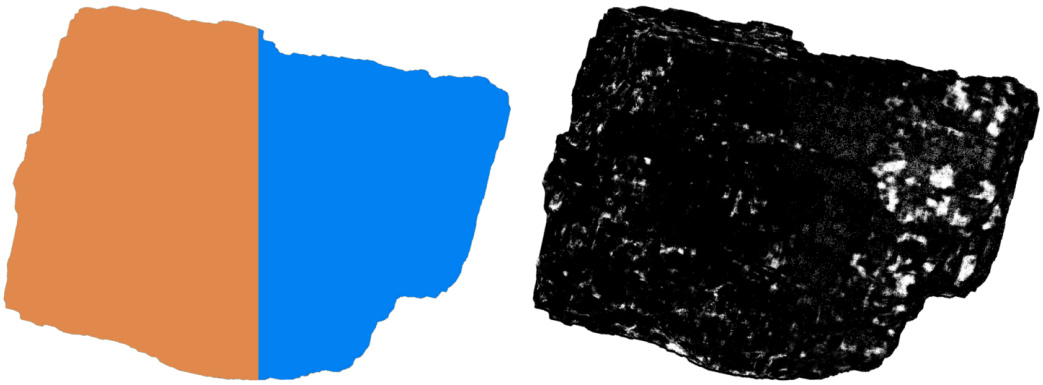
more carbon ink is recoverable in CT than is immediately visible in the raw data using virtual unwrapping. Though ink-ID is yet unable to recover the most realistic first column from this particular 12µm scan, the performance gradient across the column paints a clear picture of improved performance as the ink layer thickness meets and then exceeds the imaging voxel size. This would suggest both that the first column could be recovered if the Carbon Phantom were imaged again at a higher resolution approaching 5µm, and that the text of the real Herculaneum scrolls can also be recovered if imaged at sufficient resolution. Further, ink-ID and the associated software pipeline have room for improvement at each step, which may improve the ink recovery in the first column even using this existing 12µm data.

3.2.2 P.Herc.Paris. Objet 59

The same method was then applied to individual Herculaneum fragments to test whether the model is capable of learning the ink presence. The practical implementation of methods to handle more and larger CT volumes was developed in parallel with the theoretical framework. As a result, the earlier experiments operated on single fragments, beginning with the smallest.

First, a small fragment P.Herc.Paris. Objet 59 was used. Shown in Figure 3.5, the fragment has a few visible character pieces on the left and one clear ω (omega) on the right.

To test ink-ID performance, the fragment surface was split into a left and right



(a) 2-fold region split. (b) Resulting prediction image.

Figure 3.6: P.Herc.Paris. Objet 59 2-fold experiment and result.

half (Figure 3.6a). In the 2-fold cross-validation performed across these regions, the model trains on one half and predicts on the other, and vice versa. The result of the trained models, shown in Figure 3.6b, clearly reveals the omega on the right half, as well as some other ink spots. Though small and imperfect, this is the first ever successful recovery of a Herculaneum ink using X-ray CT. While the Carbon Phantom is a strong theoretical proof of concept, P.Herc.Paris. Objet 59 is the practical proof that Herculaneum ink specifically can be detected using only X-ray CT.

Some areas on the surface perform better than others, notably the ω character which is legible in the prediction image. Later experiments will go on to discover that this is common, and will show that the detailed structure of the papyrus surfaces and ink deposits vary considerably across the collection. By growing the training dataset, this situation will be improved, "lighting up" regions of the surface that performed poorly with smaller and less varied training data. For now, the primary conclusion from P.Herc.Paris. Objet 59 is that trained models are able to recover some of the Herculaneum ink from CT, a departure from the idea that the ink is altogether invisible in X-ray.

(a) P.Herc.Paris. 1 fr. 34. (b) P.Herc.Paris. 1 fr. 39. (c) P.Herc.Paris. 2 fr. 47. (d) P.Herc.Paris. 2 fr. 143.

Figure 3.7: Four fragments imaged in X-ray CT at Diamond Light Source in 2019, here shown in infrared photographs. Photos courtesy of BYU's Institute for the Preservation of Ancient Religious Texts.

3.2.3 Partial larger fragments

Building on this, four fragments were imaged and then processed in a manner similar to P.Herc.Paris. Objet 59. These four fragments, P.Herc.Paris. 1 fr. 34, P.Herc.Paris. 1 fr. 39, P.Herc.Paris. 2 fr. 47, and P.Herc.Paris. 2 fr. 143, are shown in infrared photograph in Figure 3.7. They were imaged in X-ray CT at Diamond Light Source in Oxfordshire, England in 2019.

The CT acquisition process was designed to image the fragments at the highest achievable resolution, in this case 3.24µm. Due to the small field of view required to capture at this resolution, the scan process required horizontal offset projections, merged before CT reconstruction, and also required the vertical stitching of multiple horizontal "slabs" after reconstruction. The vertical stitching for these datasets was not performed onsite, and so the dataset was initially a set of these horizontal slabs. This nature of the data, combined with the memory constraints of ink-ID at that time, led to initial experiments that operated on single slabs.

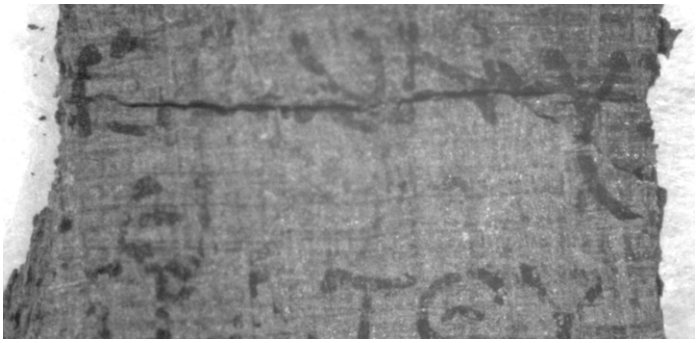
The first such experiment used individual slabs through P.Herc.Paris. 1 fr. 39. One

is shown in Figure 3.8. This and other early machine learning experiments on the "Diamond fragments" were not particularly promising. They are nonetheless shown here, as they illustrate one of the primary takeaways of this work: large gains in the ink detection performance are gained over time through the accumulation of small improvements to the various stages of the pipeline. These datasets looked early on like they had little to no detectable ink signal, something now known not to be true. This was achieved primarily with dataset development, not with sophisticated models. Improvements to segmentation, alignment, and labeling, along with increases to the training dataset size, all contribute significantly to performant ink detection.

These poorly performing early results also highlight two other lessons learned during this work. First, they are a reminder of the early emphasis that was placed on visual evaluation over purely quantitative evaluation. Most evaluation metrics for binary classification rely on binarized or thresholded outputs. As mentioned earlier, and shown here to be particularly true in the early stages, the recovered ink signal was faint and often occupied subtle ranges of the predicted ink probabilities. Binary metrics would fail to capture this, and it was therefore felt they may not provide feedback that was as helpful as visual evaluation. As the datasets and methods improve, quantitative metrics become more helpful in the upper ranges of performance.

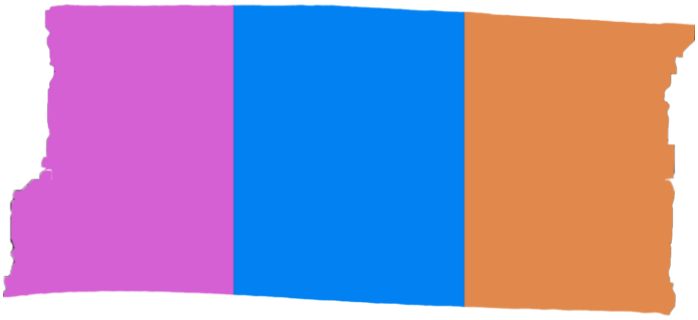
Second, these initial results compared against later, improved results suggest generally that optimism is warranted where ink does not yet appear in these images. As of this writing, refinements to the various pipeline stages and increases to the dataset size have lead to substantial improvements in regions such as that in Figure 3.8, where initially little ink was recovered. The pipeline stages still have room for improvement, and as new data is acquired, it will only grow the total training dataset.

Other slabs had similar initial results. Slab 2 initially performed quite poorly (Figure 3.9). As a result of these disappointing prediction images, the pipeline steps were reexamined to see if there was clear room for improvement. Upon close in-

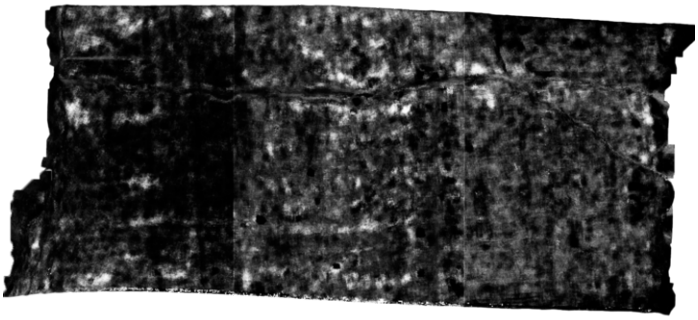


(a) Infrared.

(b) Ink labels.



(c) 3-fold cross-validation regions.



(d) Initial ink-ID result.

Figure 3.8: Initial ink-ID results on slab 4 of P.Herc.Paris. 1 fr. 39 using 3-fold cross-validation. Slight ink signal is visible in the correct areas, but no legible text is recovered.

spection, it became clear that the registration could be improved visibly when more alignment points were used. Another experiment was run using the same fragment and the realigned, more accurate labels. The results in Figure 3.10 show that the improved labels visibly improved ink-ID's ability to recover ink from the surface, as now multiple strokes are visible in the correct locations.

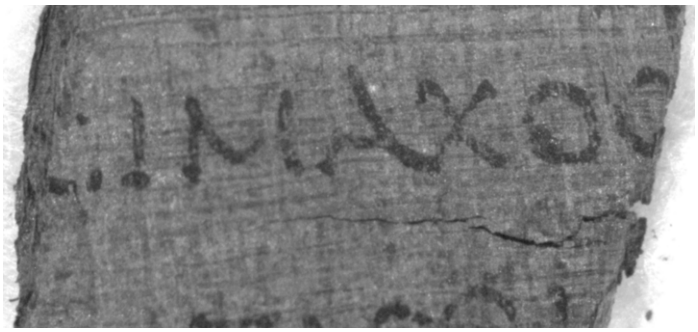
The improvement from redoing the label alignment for this slab led to another redo of the registration for all slabs of P.Herc.Paris. 1 fr. 39. Slowly, the results continued to improve. The experiments were still performed as 3-fold cross-validation experiments across individual slabs. This was done across all slabs, and the results were manually composited together in order to obtain a view of ink-ID's ability across multiple rows of text (Figure 3.11). In this arrangement, rows of text begin to appear in the correct locations, though still not many complete characters are present that could be transcribed accurately.

3.2.4 Full individual fragments

Though the total labeled training data was increasing, experiments still operated across individual slabs, so the training set for any given experiment was only as large as a single slab. The next step was to perform the volume merges required in order to treat the fragments as complete CT scans rather than collections of horizontal slabs.

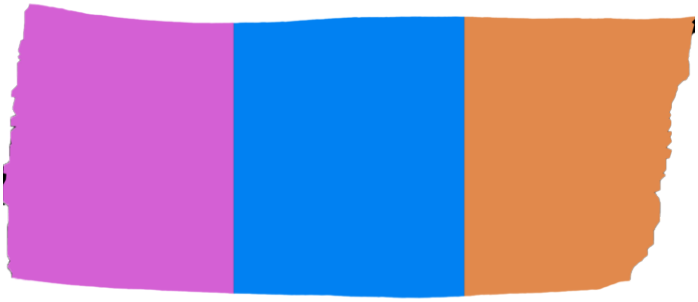
The horizontal slabs of a single slab contain vertical overlap, so it is necessary when merging them into a single volume to find the precise overlap slice between each pair of adjacent slabs. The imaging process used a physical vertical travel stage between horizontal scans, so there is no horizontal movement expected between slabs. If there were horizontal movement, an additional image registration step would be necessary to align the slices from adjacent slabs. Instead, the slices are assumed to align horizontally and require only the vertical merging.

To find the overlap point between adjacent slabs, the end slice from one was chosen as a fixed image, and then slices from the other slab were compared against this fixed



(a) Infrared.

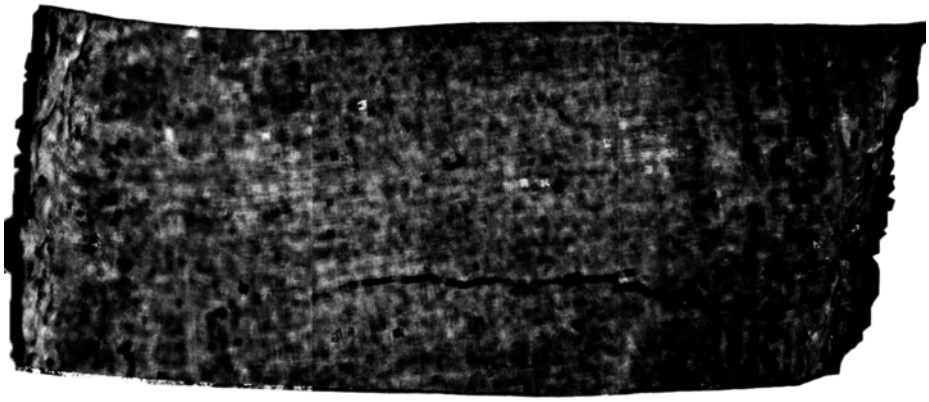
(b) Ink labels.



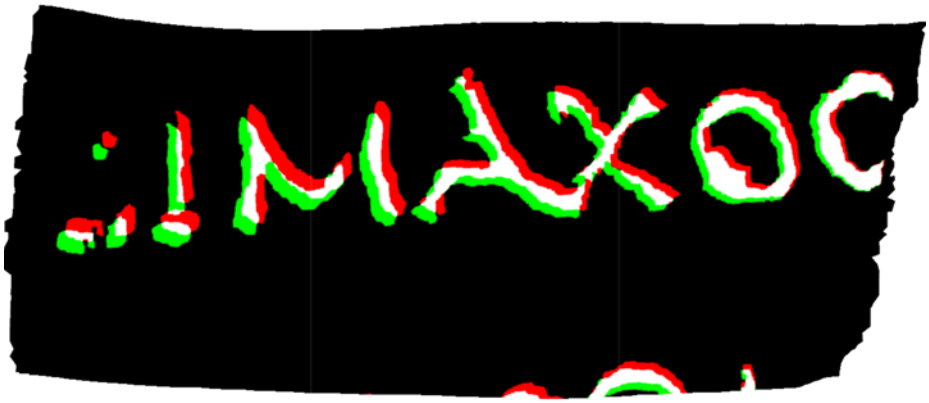
(c) 3-fold cross-validation regions.

(d) Initial ink-ID result.

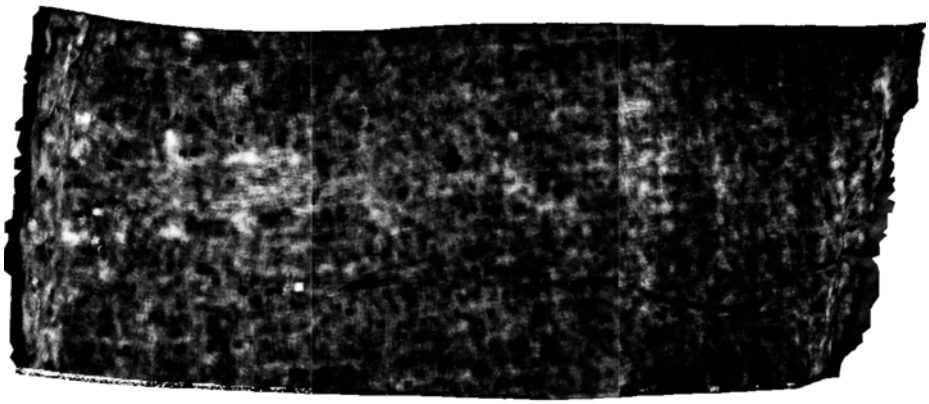
Figure 3.9: Initial ink-ID results on slab 2 of P.Herc.Paris. 1 fr. 39 using 3-fold cross-validation. This is even less promising than the slab 4 results in Figure 3.8.



(a) ink-ID results before improved registration (from Figure 3.9).

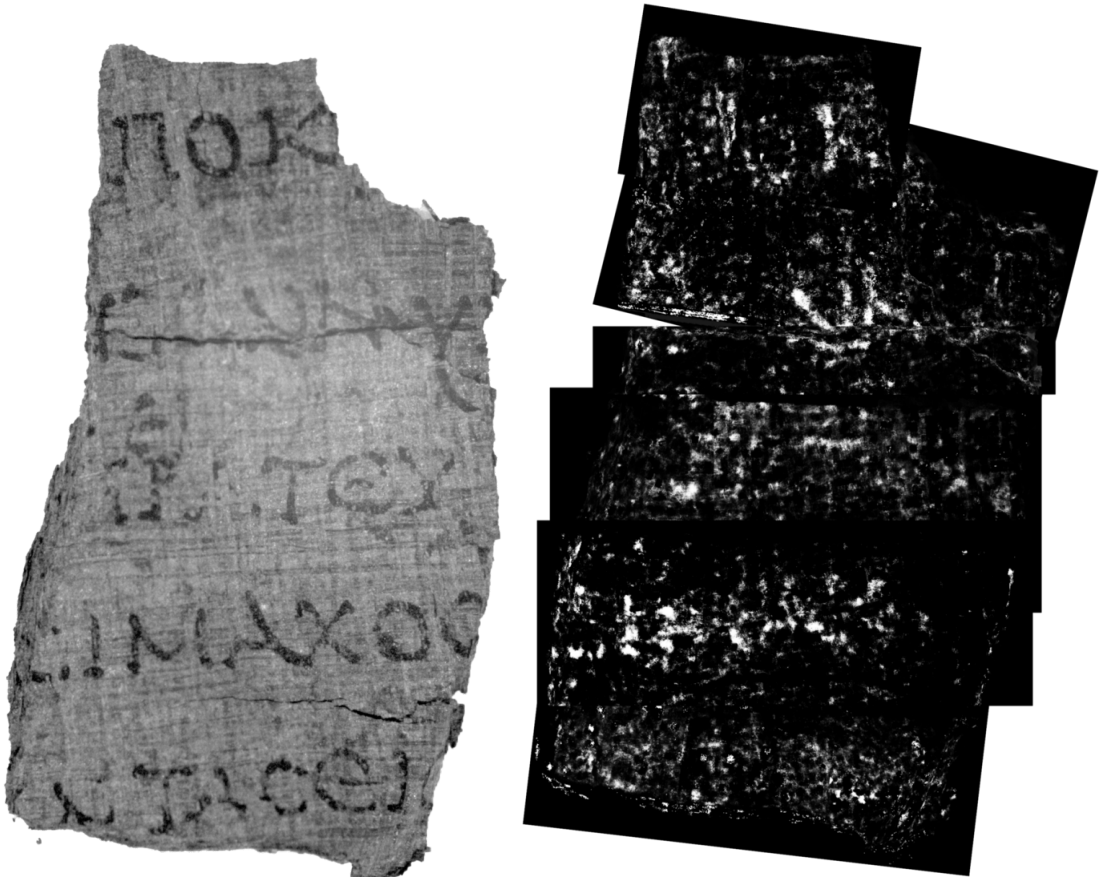


(b) Label diff after improved registration.



(c) ink-ID results after improved registration.

Figure 3.10: Improved ink-ID performance on a 3-fold experiment of P.Herc.Paris. 1 fr. 39 slab 4 after redoing label alignment. In label diff image, red shows old labels, green new labels, and white the overlap. Still, no legible text is recovered, though certain ink strokes become clearer with the improved labeling.



(a) Infrared. (b) Compiled slab experiments.

Figure 3.11: Compiled visualization of the slabs of P.Herc.Paris. 1 fr. 39, each treated as an individual 3-fold experiment. Though not yet particularly legible, rows of text begin to appear in the correct locations.

slice in order to find which of them most closely matched it. The Pearson correlation coefficient [60] was used to compare slice images, with the most correlated moving slice being selected as the overlap point. Sanity checks were also performed:

- Visual inspection was first performed to predict the overlap slice. This was typically accurate within 1-2 slices of the objective slice chosen using the Pearson correlation coefficient.

- The program to find the overlap slice displays not only the slice with highest correlation, but the top 20 sorted by correlation. It was verified that the correlation smoothly decreased when moving in either direction from the overlap slice. In other words, if slice n is identified as the overlap slice, the next most correlating slices should be $n + 1$ and $n - 1$, and then $n + 2$ and $n - 2$ and so on.

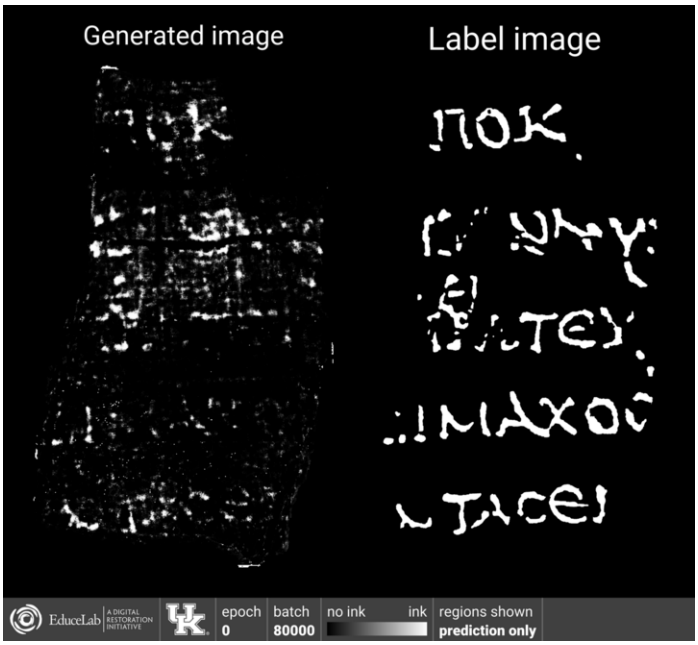
- Ultimately it was found that the overlap point always occurred between 1546 and 1550 slices into the slab volume, consistent with expectations that the vertical travel stage had precise and reproducible movement.

The result of this process was a merged volume for each fragment, allowing experiments across a complete fragment surface. Figure 3.12 shows the results of a 4-fold experiment across the surface of P.Herc.Paris. 1 fr. 39 after the volume was merged. Like most advancements in this work, improvements over the slab-based results in Figure 3.11 are marginal, but a step in the right direction.

The other fragments, treated similarly as individual experiments using 4-fold cross-validation on their surfaces, showed similar results. Figure 3.13 shows the result of such an experiment on P.Herc.Paris. 1 fr. 34, with the recovered text at least as clear as that from P.Herc.Paris. 1 fr. 39.

Some regions seem to perform better than others. For example, the MAX characters in the middle of the fourth row on P.Herc.Paris. 1 fr. 39 seem not to perform

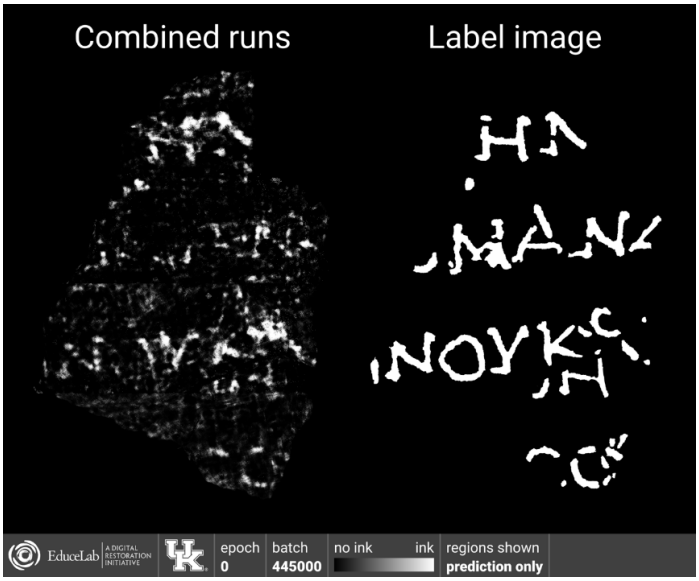
(a) 4-fold regions.



(b) ink-ID results.

Figure 3.12: Initial ink-ID results on P.Herc.Paris. 1 fr. 39 when treated as one merged volume or complete surface rather than processed as individual horizontal slabs. Results shown of a 4-fold cross-validation experiment.

(a) 4-fold regions.



(b) ink-ID results.

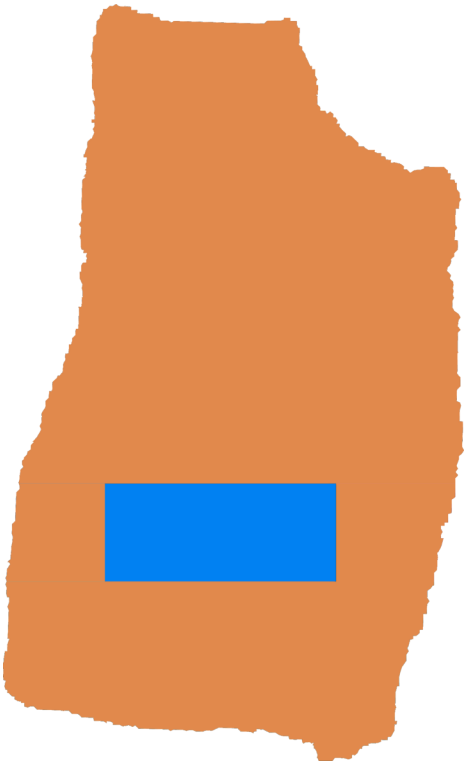
Figure 3.13: Initial ink-ID results on P.Herc.Paris. 1 fr. 34 when treated as one merged volume or complete surface rather than processed as individual horizontal slabs. Results shown of a 4-fold cross-validation experiment.

particularly well. It is not immediately clear if this is because there is not enough ink signal present in the CT there, or if the form of the ink signal simply differs from that on other parts of the surface such that the model has not been trained to detect it. In an attempt to disambiguate this, an experiment was run where those characters were isolated as the prediction region and the model was trained on the rest of the surface. The primary difference between this setup and the 4-fold row-based setup of Figure 3.12 is that the model predicting across the region of the MAX characters was able to train on the other characters from the same row of text. Perhaps, if something about the segmentation or ink signal in that row is unique, the model can learn this from the widened training distribution and improve its performance on that region.

Figure 3.14 confirms that a larger training set leads to improved predictions. By having exposure to a wider distribution of ink signal during training, the model improves its ability to recall the ink in this troublesome MAX region. This supports the hypothesis that the ink signal is varied across the collection, and that models exposed to as much variety as possible will generally have improved performance.

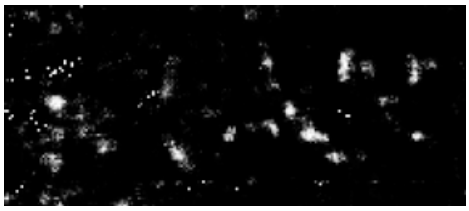
This effect is also observable more generally, by simply choosing a larger k for k-fold cross-validation, increasing the training dataset for each individual experiment. Figure 3.15 shows the result of an experiment on the surface of

At this point, it was clear that increases to the training dataset size and diversity led to significant performance improvements with ink detection. As experiments were now being run on entire scroll fragments each with multiple rows of text, the next step would be to train a model across multiple fragments. Memory constraints with the ink-ID implementation at that time prevented these experiments, as the entire CT volume of each fragment in the experiment had to be loaded into RAM during execution. It was not even possible to perform an individual experiment across the entire surface of P.Herc.Paris. 2 fr. 143, as the merged volume is 602 GB even after cropping it tightly to the fragment bounds, well exceeding the RAM on any of the

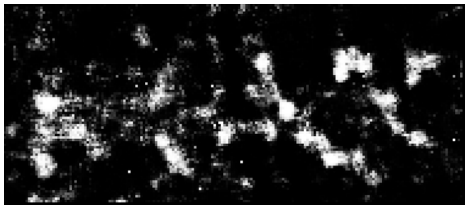


(a) Targeted 2-fold regions.

(b) Binary ink labels.

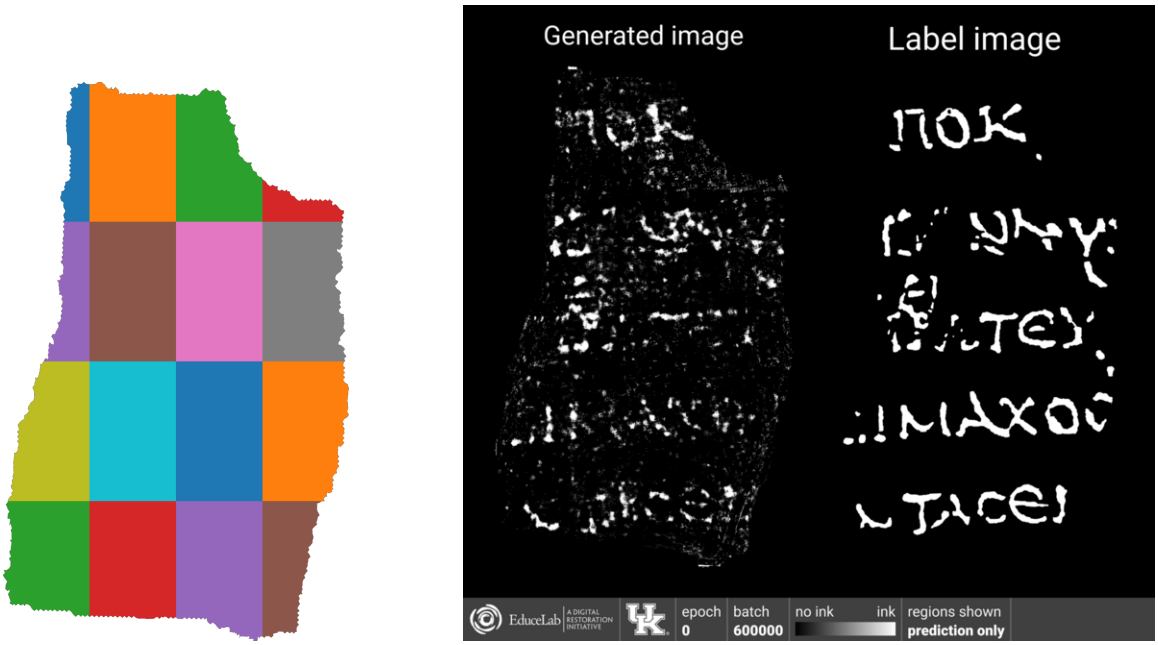


(c) Prior 4-fold (from Figure 3.12).



(d) Targeted 2-fold.

Figure 3.14: Impact on ink-ID predictions for MAX region when training on all but that particular region. Training region in orange, prediction region in blue. When training on more data, in particular other ink from the same text row, model has much better recall.



(a) 16-fold regions.

(b) ink-ID results.

Figure 3.15: ink-ID results on P.Herc.Paris. 1 fr. 39 when dividing surface into finer 4x4 grid. More training data and variety leads to better performance.

available compute resources. The surface volumes described in Section 2.4.1 were implemented as a solution to this problem, enabling the following experiments across larger datasets.

3.2.5 Multiple fragment experiments

The geometric change to use surface volumes enables the training of networks across multiple fragments at once, as their slimmed volumes can now all fit in memory during program execution. This is a step towards the vision originally described for this work, where a large body of scanned and labeled fragments would serve as a "reference library" of training data that would train a robust and diversified model. As has been thoroughly established elsewhere in machine learning, larger datasets often increase performance at least as much as improved models [61, 62].

The first experiment across multiple fragments used P.Herc.Paris. 1 fr. 34 and P.Herc.Paris. 1 fr. 39, with each surface split into top and bottom halves for a 4-fold experiment. The results are shown in Figure 3.16, demonstrating continued marginal

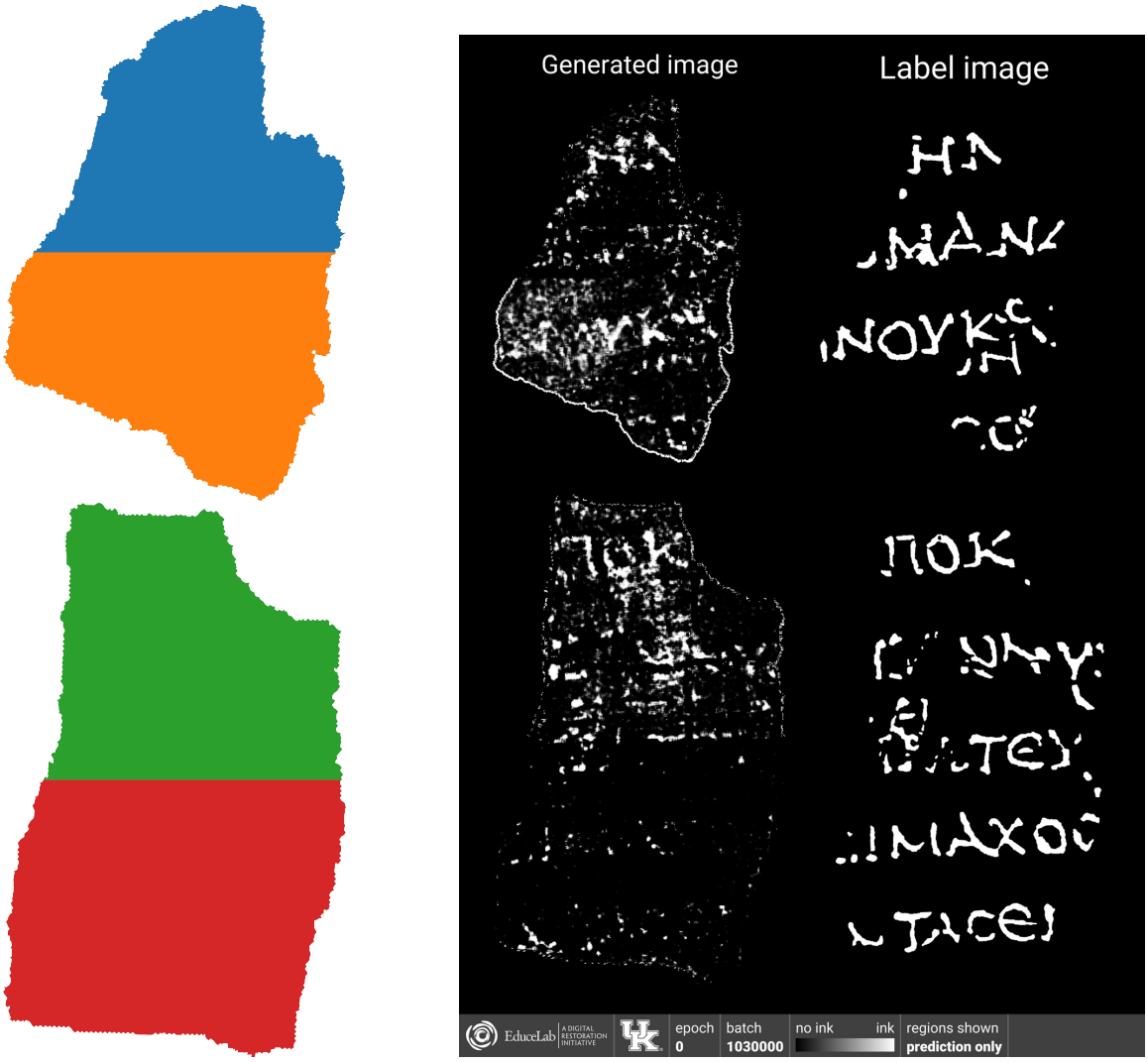
improvement.

Surface volumes also enabled the processing of P.Herc.Paris. 2 fr. 143, which due to its larger physical dimensions had a volume that did not fit in memory on the available compute resources, even for an experiment on it individually. The surface volume reduced the disk size (and subsequently RAM size during execution) of this dataset from 602 GB to 18 GB. After the data processing pipeline, a sanity check 2-fold experiment was run on this fragment individually, to make sure the data appeared coherent and some ink signal was present. Figure 3.17 shows the results, primarily for comparison against the improved results on the same surface once the model is trained on multiple fragments.

Finally, the four fragments were used jointly in an ink-ID experiment. Each surface was split into a top and bottom region, resulting in an 8-fold experiment across the four fragments. The 8-fold split was chosen to be able to compare the increased training dataset against the 2-fold experiments previously used on the individual fragment surfaces. Figure 3.18 shows the results of this experiment. These are the best visual results yet generated from this data. For a direct indicator of the impact of the larger training set, compare the results on P.Herc.Paris. 2 fr. 143 in this experiment against those in Figure 3.17, when it was the only fragment in the experiment. The false positive noise is significantly reduced, and the character boundaries are clearer.

3.2.6 Reprocessing data with more experience

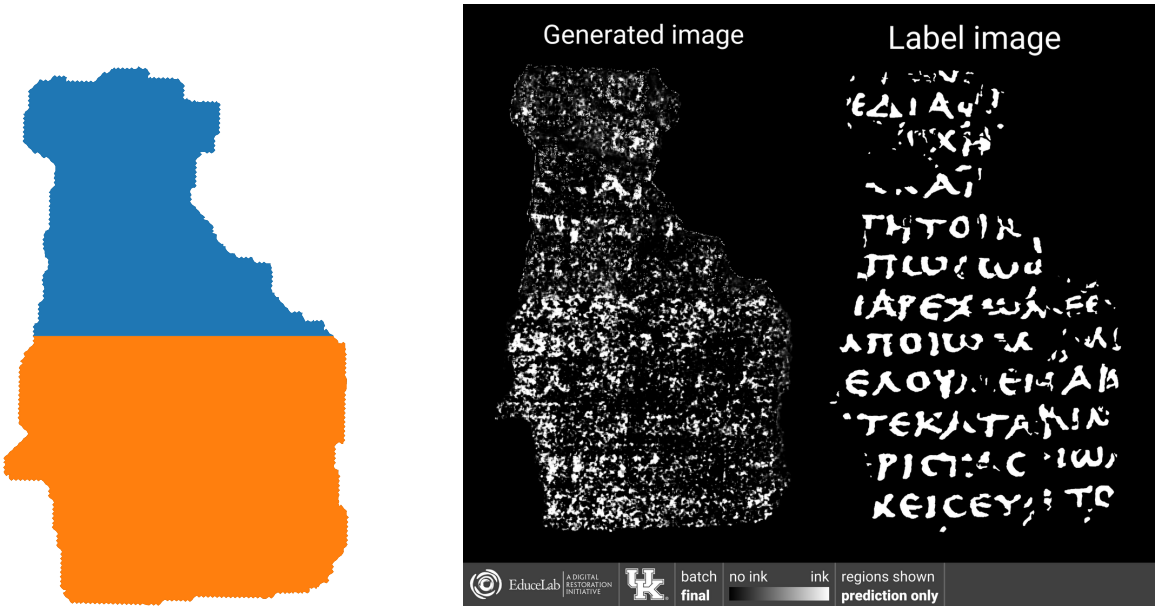
Though much of the data processing pipeline involves algorithmic components, there is still a fair amount of manual processing necessary to get from raw image data to the aligned images used in ink-ID. Throughout the work conducted to this point, I continued learning crucial details about various parameters and stages of the pipeline, and became more skilled and consistent at each of them. The fragments had been processed one by one and months apart, as there was engineering work required in between before it was even possible to process some of them, so they were done



(a) 4-fold regions.

(b) ink-ID results.

Figure 3.16: 4-fold ink-ID experiment across two fragments, the first instance of a training experiment across multiple CT scans or multiple scroll fragments.



(a) 2-fold regions. (b) ink-ID results.

Figure 3.17: Initial 2-fold ink-ID experiment on P.Herc.Paris. 2 fr. 143.

to varying standards. Reliably, when performing experiments, more and better data provided improved returns on ink detection performance than changes to the model architecture or similar adjustments. Based on this, it was decided to reprocess the Diamond fragments from scratch, beginning with the reconstructed CT slabs for each fragment and going through each step of the pipeline anew.

Using the lessons learned from the work to this point, particular care was taken in the following areas:

- Image intensity windowing is necessary to map the raw CT reconstructions in arbitrary float ranges to the 16-bit unsigned integer representation used by ink-ID. This had previously been done per-slab, using visual indicators or raw min/max values from that slab to set the intensity window. This time, windowing was done using a percentile method to discard outliers, and the same percentile window was applied to all slabs across all fragments. This removed some intensity streaking or banding that had originally been present in the merged volumes and texture images.

(a) 8-fold. (b) ink-ID results.

Figure 3.18: Initial ink-ID results across the four "Diamond fragments". 8-fold cross-validation used.

	Blur	Min	Max	Aperture	Contour	Closing size	Bilateral	Midpoint
P.Herc.Paris. 1 fr. 34	1	39	114	0	✓	10	✓	
P.Herc.Paris. 1 fr. 39	1	39	114	0	✓	10	✓	
P.Herc.Paris. 2 fr. 47	1	39	114	0	✓	10	✓	
P.Herc.Paris. 2 fr. 143	1	39	114	0	✓	16	✓	

Table 3.1: Canny edge detection parameters chosen for segmentations of the Diamond fragments.

- All slab overlaps were inspected visually, confirmed with the Pearson correlation coefficient, and confirmed to be in the expected range.
- Consistent parameters were used for the Canny edge detection used in segmentation as much as possible, shown in Table 3.1.
- Cleaning and meshing of the point clouds, a rather manual process, was done as consistently as possible.
- More time was spent on registration, roughly doubling the number of reference points used in the manual alignment.
- Binary ink labeling was performed with the supervision of a papyrologist, to help disambiguate tricky ink spots on the surface.

Using the reprocessed dataset, the 8-fold experiment across the four fragments was performed again. Figure 3.19 shows the results of this experiment, which represents the state of the art ink detection produced by this work at the time of writing (ink detection has been refined further as a part of the Vesuvius Challenge, see Chapter 7). The recall and sharpness of the characters that appear in this experiment exceed those of the same experiment with slightly inferior data. Once again, the accumulation of very small improvements across the dataset led to material gains in ink detection. As each step in the pipeline still has room for further improvement, there is good reason for optimism when looking forward on the ink detection problem.

(a) 8-fold. (b) ink-ID results.

Figure 3.19: ink-ID results across the four "Diamond fragments" after redoing data processing pipeline using lessons learned. 8-fold cross-validation used. This image represents state of the art ink detection produced directly by this work.

3.3 Evaluation

One of the cornerstones of the methods presented in this work is that they can be evaluated against fragment surfaces with known ground truth. This is crucial to build confidence in the ink detection, before it is used on internal, hidden layers for which there is no ground truth. For evaluation, recall and false positives are notable as they quantify subjective qualities of interest: how much of the true text can be recovered, and how much false text is generated in error? While it is acceptable not to detect challenging ink regions, the model should not substantially "hallucinate"; that is, it should not generate misleading images of characters that are not actually present.

3.3.1 Visual

Visual inspection allows even a non-expert to quickly verify the generated images against the ground truth, checking for rough accuracy, recall, and noise in the generated images. This also allows one to evaluate multiple subjective qualities that combine to create "readability," a measure that empirically did not always correlate precisely with quantitative metrics. A brief glance at the results and ground truth in Figure 3.19 suggests that ink-ID is correctly detecting the locations of text lines, nearing $\sim$ 50% recall of individual characters, and without notable false positives (characters that do not exist in the ground truth).

Visual validation was used extensively in the development cycle as an early indicator of model performance. As shown in early experiments (Figures 3.8, 3.9, and following), ink detection initially did not perform well enough to generate clear prediction images that would much resemble the labels after binarization. Initial quantitative evaluations did not always correlate with improved readability in the prediction images, so it was decided to prioritize visual evaluation in this phase. For the volume of experimentation done in this work, this was sufficient to guide the development.

(a) ink-ID (binarized) (b) Ground truth

Figure 3.20: Example ink-ID output and ground truth for binary ink classification task on P.Herc.Paris. 2 fr. 47.

3.3.2 Quantitative pixel-based metrics

Now that ink detection has converged to more accurate predictions, improved metrics are more likely to correspond with improved readability, and the risk of degenerate cases has been reduced. Going forward, a shift towards objective metrics should help refine ink detection methods, and allow the objective evaluation of larger volumes of experiments.

Ink identification is so far performed as a binary classification task, with predictions generated for each output pixel individually. Figure 3.20 visualizes this task, showing the ground truth binary label mask (3.20a) and the prediction image generated by ink-ID (3.20b). Once pixel predictions are aggregated at the image level, they can be evaluated against the ground truth using common metrics from semantic segmentation, another common computer vision task.

Here, the quantitative results of the latest ink-ID predictions in Figure 3.19 are reported. As this is a new task and dataset, these are presented primarily as a

Metric	μ	σ
Binary cross entropy	0.44	5.6×10^{-3}
Precision	0.58	4.2×10^{-2}
Recall	0.41	3.0×10^{-2}
FPR	0.051	9.9×10^{-3}
Dice/F1	0.48	1.5×10^{-2}
F0.5	0.53	2.4×10^{-2}

Table 3.2: Aggregate validation results for the pixel-wise binary ink detection task, across the four fragments in Figure 3.19, following 8-fold cross-validation. Mean (μ) and standard deviation (σ) reported for each, sampled from the second half of 620,000 total training batches.

benchmark against which future improvements can be evaluated. The widely used binary cross entropy and Dice [63] (also known as F1) metrics are reported. Precision, recall and false positive rate (FPR) are also included for their aforementioned ties to subjective qualities of interest. Additionally reported is F0.5, a measure similar to Dice/F1 that weights precision higher than recall in order to reduce false positives and encourage sharper characters:

$$\frac{(1+\beta^2)pr}{\beta^2 p + r}, \quad \beta = 0.5 \tag{3.1}$$

where p is precision and r is recall.

Table 3.2 shows the metrics associated with the results visualized in Figure 3.19. The results are aggregated across the 8-fold cross validation, meaning the 8 individual predictions are compiled and then metrics are computed. As output predictions fluctuate throughout training, the mean μ and standard deviation σ are reported, taken from the second half of 620,000 training batches. The second half is chosen as the model converges before this point.

The quantitative metrics align with the visual results, with recall higher than the false positive rate by a factor of 8. Both visual and quantitative results suggest the model is capable of accurate ink detection without significant hallucination.

Quantitative metrics are already being used more frequently in this area. As the baseline already met a performance standard that suggested quantitative metrics would be fruitful, the Vesuvius Challenge Ink Detection Progress Prize was launched on Kaggle using the F0.5 score as the evaluation metric. This seems to have done well at preferring submissions that struck the desired balance of precision and recall (see Chapter 7). Ablation studies in Chapter 6 also commonly rely on objective measures.

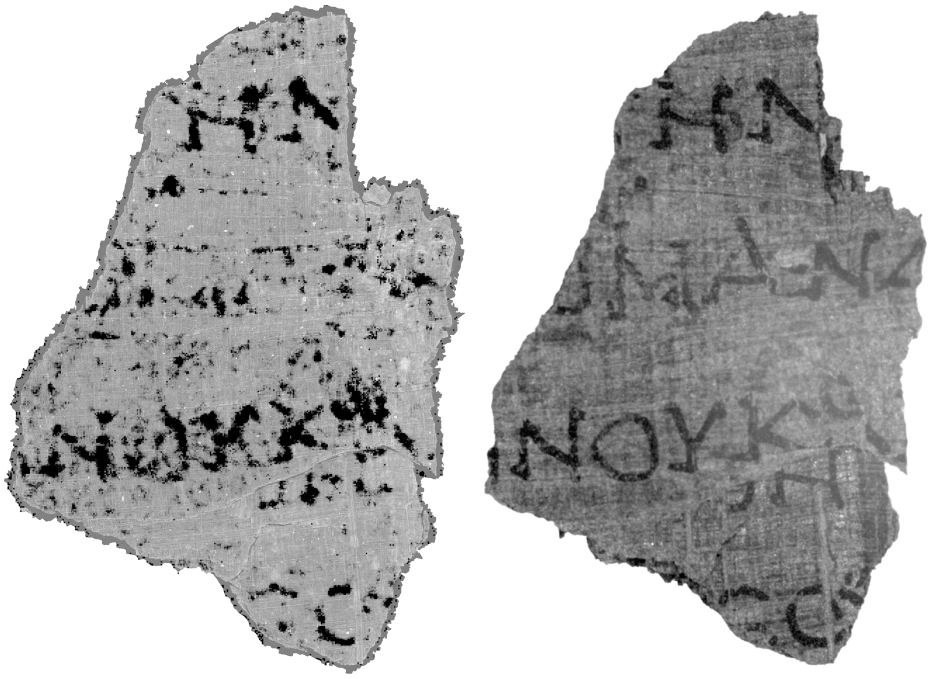
3.3.3 Papyrological character-based metrics

Quantitative image metrics are helpful for method development, as they can be automated and provide quick feedback. However, they do not exactly capture the ultimate objective, which is the readability of generated text to the trained papyrologists who study the Herculaneum papyri. In an effort to better capture readability, recall and false positives are again reported, this time with respect to individual characters as identified by papyrologists.

ink-ID prediction images from the experiment in Figure 3.19 were provided to a Herculaneum papyrologist[1], who was previously unfamiliar with these fragments. They transcribed each fragment based solely on the provided images, which are generated purely from X-ray CT. After this transcription, they were provided with ground truth infrared images, and another ground truth transcription was made. Using ink-ID images, the scholar was able in all cases to identify the correct number of lines of the fragment surface, even correctly identifying in P.Herc.Paris. 1 fr. 34 where the surface is broken and marks a transition between two separate writing layers.

For each fragment, the papyrologist was provided with the direct ink classification prediction image from ink-ID, as well as a composite image combining these predictions with the texture image from virtual unwrapping. An example of this image is shown in Figure 3.21 alongside the ground truth infrared photograph provided for transcription after the papyrologist had transcribed the ink-ID images. The compos-

[1]Special thank you to Marzia D'Angelo and Michael McOsker.



(a) ink-ID + texture image. (b) Ground truth (infrared).

Figure 3.21: Example images provided to papyrologist for transcription. P.Herc.Paris. 1 fr. 34 shown.

ite image is designed to show the predicted ink alongside useful spatial cues such as papyrus fibers, and is also meant to visually resemble the infrared images the scholars are used to examining.

Figure 3.22 shows the P.Herc.Paris. 1 fr. 34 transcriptions, made from ink-ID and ground truth images respectively. Of the 15 characters identified in the ground truth, 7 are recovered accurately using ink-ID. No characters are misidentified, or identified where none exist.

Table 3.3 summarizes these results across the four Diamond fragments. In all fragments but P.Herc.Paris. 2 fr. 47, recall exceeds the false positive rate by at least a factor of 2. The results are computed using a "strict" scoring, in which even those characters identified as uncertain by the papyrologist are scored. Excluding these characters would further improve the results. Further, characters that are mistakenly identified are typically the result of partial ink recovery, showing only some of the ink

	layer a		layer a
1	].HN[	1	].HN[
2	].....[	2	]Ọ̣MANΔ[
3	].NOΥK.[	3	].NOΥK.[
	layer b		layer b
1	]..[	1	]Ọ̣N[
2	].Ḍ.[	2	]Ẹ̣Ḍ.[

(a) ink-ID. (b) Ground truth.

Figure 3.22: Greek transcriptions for P.Herc.Paris. 1 fr. 34 from a trained papyrologist, of both ink-ID generated prediction image and ground truth infrared image.] and [indicate line beginning and end. A dot indicates indistinct ink traces and an underdot indicates an uncertain transcription.

Fragment	Characters	Recall	FPR
P.Herc.Paris. 1 fr. 34	15	0.47	0.00
P.Herc.Paris. 1 fr. 39	25	0.64	0.12
P.Herc.Paris. 2 fr. 47	26	0.23	0.27
P.Herc.Paris. 2 fr. 143	59	0.41	0.22
Cumulative	125	0.42	0.18

Table 3.3: Character recall and false positive rate (FPR) across the four Diamond fragments, comparing human transcriptions from ink-ID generated images against human transcriptions from ground truth. The number of characters in the ground truth transcriptions are also shown.

strokes making up a complete character. Importantly, the misidentified characters are not invented "whole cloth." They are at least identified in the correct location, and could be accurately identified with marginal improvements in ink detection.

These results are a tremendous step forward from the previously nonexistent baseline. This method already performs well enough to generate valuable text for Herculaneum papyrologists, who are used to working with fragmentary writing even in the best cases. That said, there is clear room for improved performance, which this work suggests is likely to follow.

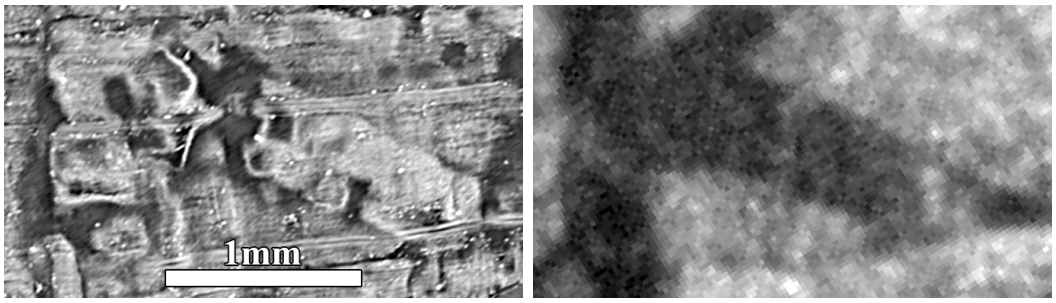
3.4 Revisiting direct inspection

Section 1.4 discussed the various imaging methods that have been attempted in order to achieve ink contrast from within a rolled scroll. The results then shown in this chapter are at odds with the initial conclusion that X-ray CT does not capture this ink contrast. Now that it has been shown the ink presence is indeed somehow captured in CT, it is worth revisiting the CT images to see whether anything can be observed that would suggest *what* is being detected by the trained models. This section presents the resulting findings, and aims to contribute a more thorough picture of what Herculaneum ink can look like in CT than has been published.

The aligned labels are helpful in this process, as they allow the X-ray CT images to be inspected alongside their respective ink labels. Perhaps by inspecting these images together, a human could learn to identify some visual pattern indicating ink that was missed while blindly inspecting X-ray CT without corresponding labels. If present, this visual pattern could be what is detected by the neural networks. The work of creating the aligned dataset required exactly this close side by side inspection of the X-ray CT and label images, and indeed resulted in occasional visual evidence of ink in the X-ray CT images.

Until an algorithmic registration process is discovered for these images, the manual alignment from Section 2.3.7 relies heavily on close manual inspection of the texture images and the infrared photographs in order to find reference points shared between the two images. These reference points are often features of the fragment boundary or papyrus fibers, as shown in Figure 2.17, but in less frequent cases it is clear that patterns corresponding to the written characters are themselves visible in the texture image.

Figure 3.23 shows an example of this from the surface of P.Herc.Paris. 1 fr. 34. In this case, the ink appears as a protrusion from the papyrus surface that has a smoother texture and even a slight intensity difference. It appears that the ink on



(a) X-ray CT. (b) Infrared photograph.

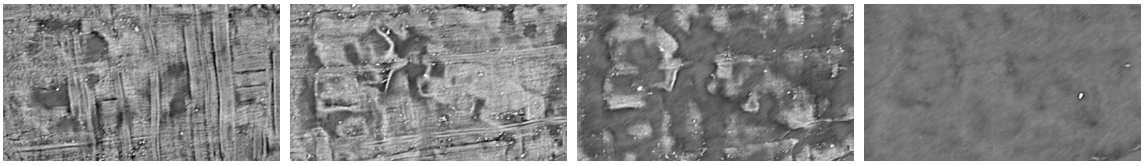
Figure 3.23: Example location from the K on the surface of P.Herc.Paris. 1 fr. 34 where the presence of ink is directly visible in the X-ray CT surface volume.

these strokes is thick enough that it is readily visible in CT. This is compatible with the morphological hypothesis, which suggests that when ink is thick or pronounced enough, the textural changes should be directly visible.

When looking for visual ink indicators, it is necessary to examine various depths with respect to the segmented surface. The segmentation is not perfect, so occasionally does not perfectly follow the papyrus surface. Different findings also have varying thicknesses, and seem to inhabit slightly different depths with respect to the surface. The surface volume simplifies this form of inspection, as the complete image stack can be loaded into an image viewer such as ImageJ/Fiji [64] or Photoshop and one can scrub through the slices while looking at a zoomed region of interest.

Figure 3.24 visualizes this process using some selected depths for the ink signal shown in Figure 3.23. When far down into the papyrus sheet, the vertical fibers from the verso (back face) are primarily visible. Moving towards the papyrus surface, the horizontal fibers of the recto become more prominent, and the ink begins to appear. The ink clearly has some physical thickness, as it stays in view even as the surrounding papyrus fades to air.

By closely examining aligned X-ray and infrared images during the alignment process, one develops an intuition for the occasional appearance of ink. To see how much of the ink on the surface is recoverable from direct visual observation, the en-



(a) (b) (c) (d)

Figure 3.24: Ink signal from Figure 3.23 at different depths from the segmented surface. (a) -11 layers from surface. Vertical papyrus fibers from back of sheet are prominent. (b) -3 layers from surface. Horizontal fibers from front of sheet are more prominent and ink begins to appear. (c) +2 layers from surface. Some protruding ink remains while papyrus begins to fade to air. (d) +10 layers from surface. Only air is visible.

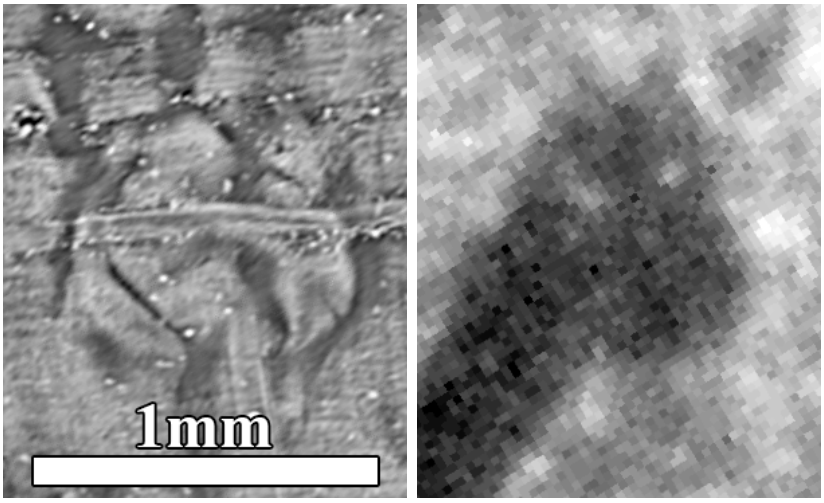
tire surface volume was later inspected again to look for patterns of interest, this time without the label image as a guide. Interesting surface features that seemed to resemble "inkiness" were flagged. Figure 3.25 shows those that were identified.

In addition to the ink spot already shown in Figure 3.23, only one other spot of interest on the surface is clearly ink, shown in Figure 3.26. Another spot, shown in Figure 3.27, is ambiguous. The character shape there would suggest there is no need for an ink stroke there, but there is a dark spot of some kind in the infrared, though it is less dark than the surrounding ink strokes. Ambiguous cases like these are one challenge of binary ink classification labeling, and are a vote in favor of multimodal methods like those in Chapter 4. The remainder of the flagged spots of interest appear not to be ink but other surface occurrences.

Figure 3.28 shows the results of the same process on the surface of P.Herc.Paris. 1 fr. 39, another fragment from the same scroll. Similar to P.Herc.Paris. 1 fr. 34, there is one clear instance of visible ink in X-ray CT, shown in Figure 3.29. The other spots of interest on P.Herc.Paris. 1 fr. 39, despite similar appearances to the ink spot, evidently do not correspond to ink on the surface but instead other unknown patterns. Figure 3.30 shows two examples of this, though there are more.

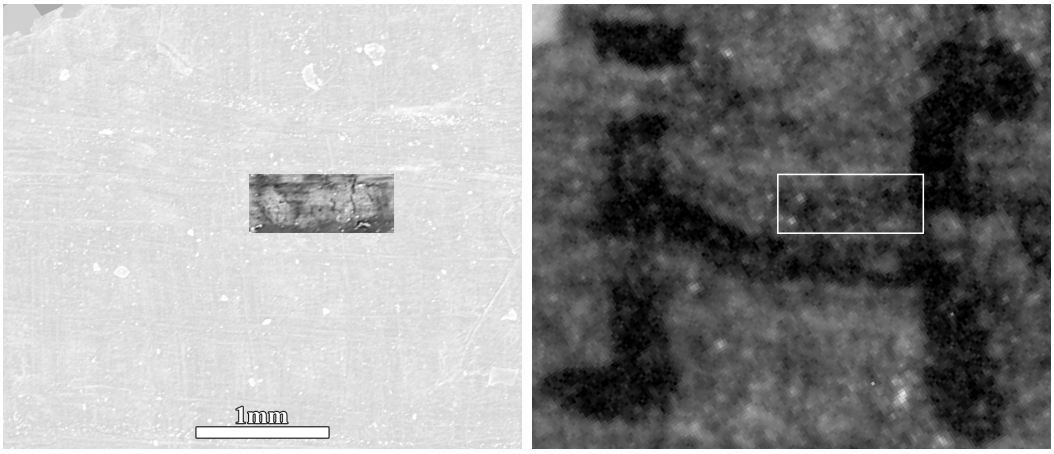
The same process was also performed on P.Herc.Paris. 2 fr. 47 and P.Herc.Paris. 2 fr. 143, which seem to have thicker ink layers (or a "heavier hand") than the fragments

Figure 3.25: Manually identified areas of interest on the surface of P.Herc.Paris. 1 fr. 34, overlaid on the texture image to show their position. Individual areas of interest may come from different depths within the surface volume. After comparison with infrared, only two of these regions actually correspond to ink. Inset: the same surface in infrared.



(a) X-ray CT. (b) Infrared photograph.

Figure 3.26: Example location from the ϒ on the surface of P.Herc.Paris. 1 fr. 34 where the presence of ink is directly visible in the X-ray CT surface volume.



(a) X-ray CT. (b) Infrared photograph.

Figure 3.27: Ambiguous instance where it is unclear if what is seen in CT is ink or something else. Region of interest highlighted.

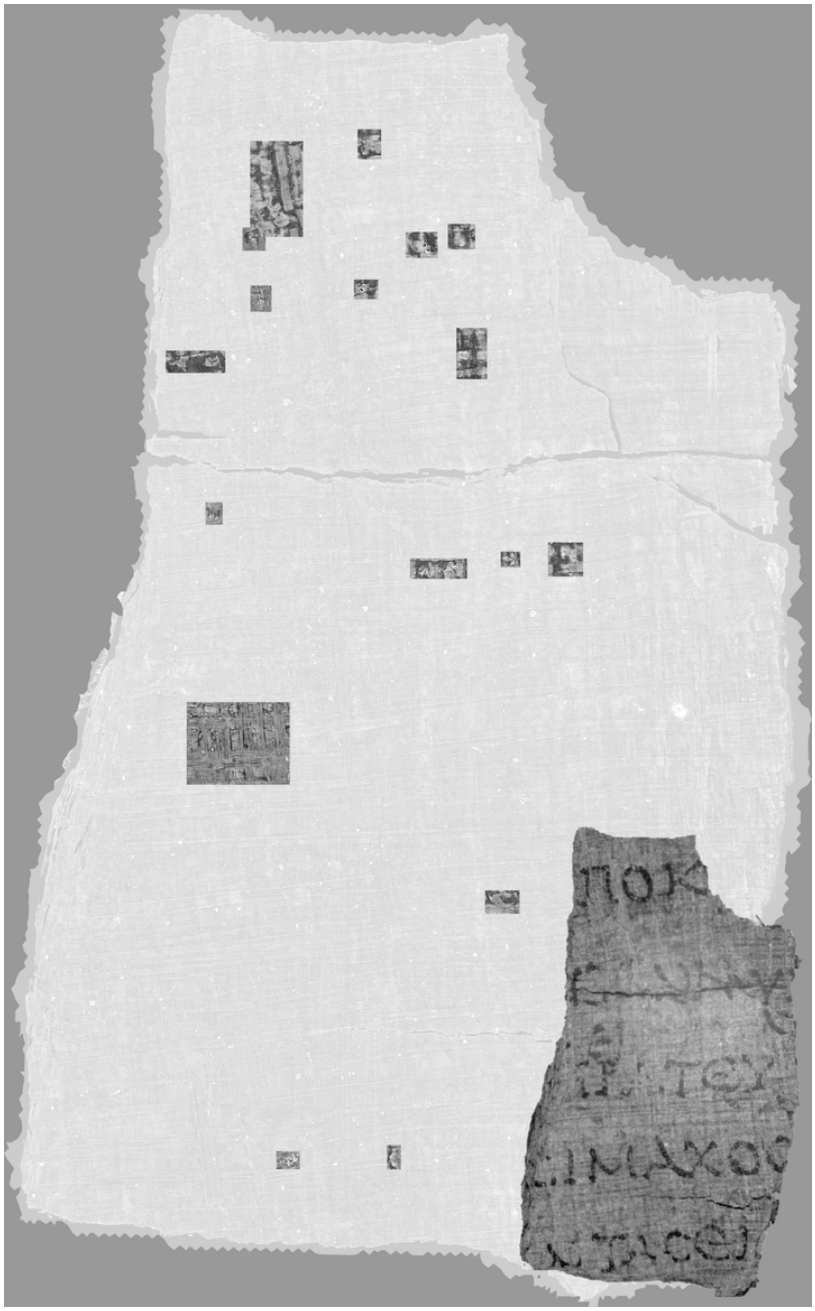
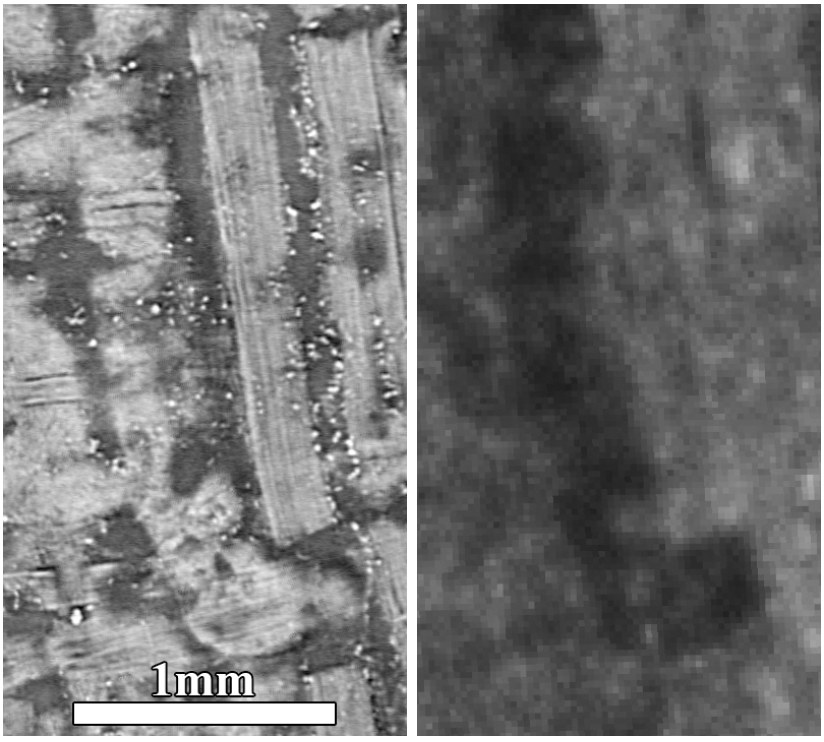
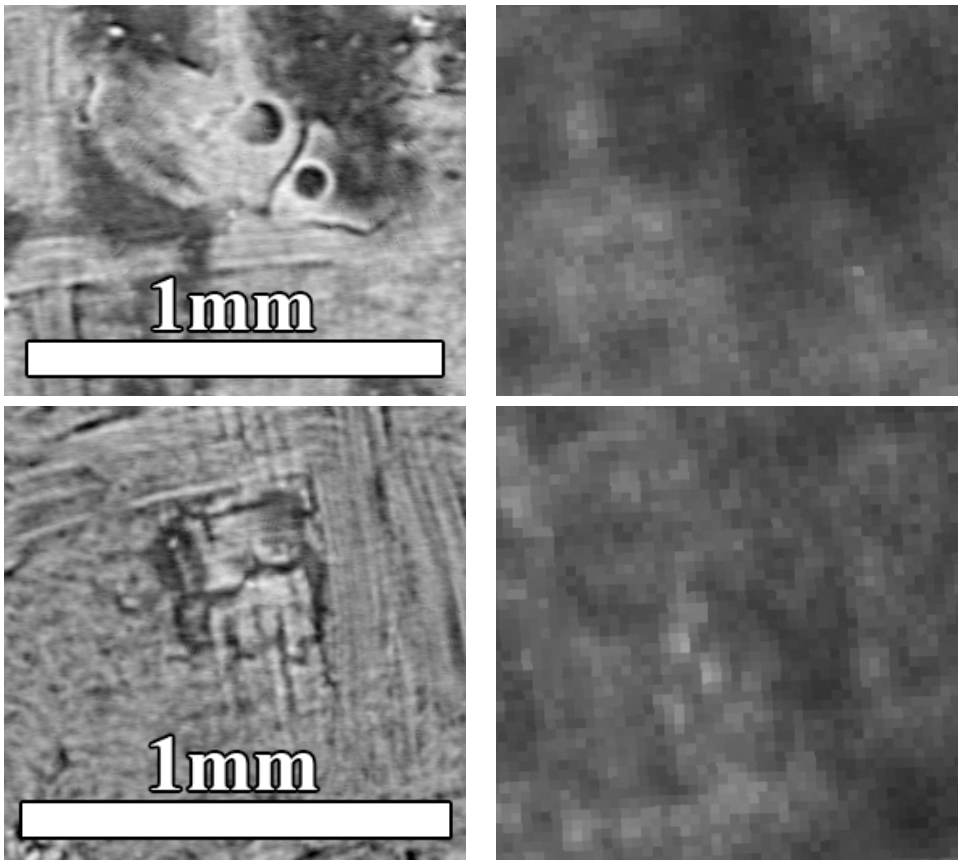


Figure 3.28: Manually identified areas of interest on the surface of P.Herc.Paris. 1 fr. 39, overlaid on the texture image to show their position. Individual areas of interest may come from different depths within the surface volume. Inset: the same surface in infrared.



(a) X-ray CT. (b) Infrared photograph.

Figure 3.29: Example location from the Π on the surface of P.Herc.Paris. 1 fr. 39 where the presence of ink is directly visible in the X-ray CT surface volume.



(a) X-ray CT. (b) Infrared.

Figure 3.30: Example locations from the surface of P.Herc.Paris. 1 fr. 39 where areas of interest in CT seem *not* to be ink.

from PHercParis1. For P.Herc.Paris. 2 fr. 47, the overview is shown in Figure 3.31, regions of interest that seem to be ink are shown in Figure 3.32, and regions of interest that seem *not* to be ink are shown in Figure 3.33.

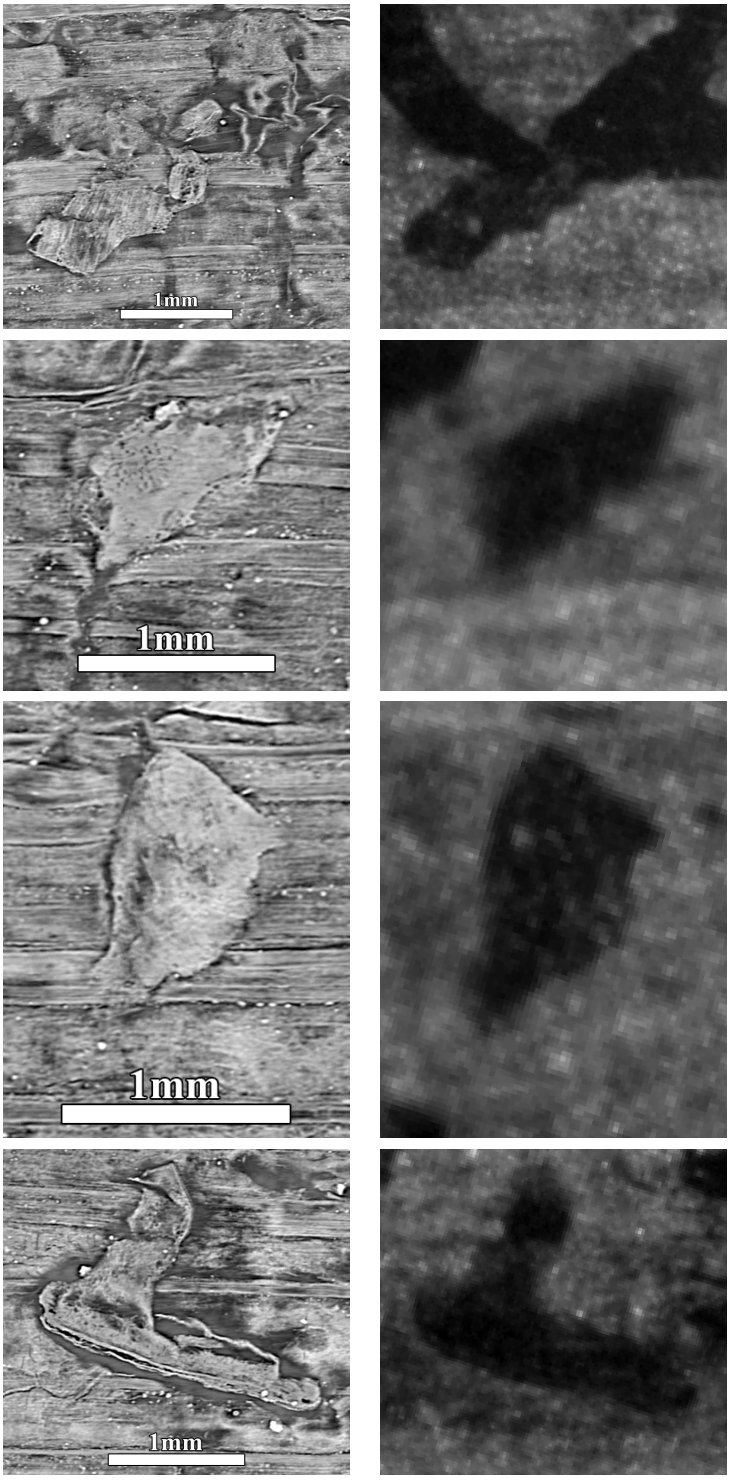
PHercParis2Fr143 was also studied, with the overview shown in Figure 3.34. Generally, similar results were obtained with this fragment to the others: there are some regions of interest in the CT that do end up being ink (Figure 3.35), and some that do not (Figure 3.36), with visual appearances resembling those from the other fragments.

In addition to these ink patterns which resemble those from the other fragments, P.Herc.Paris. 2 fr. 143 exhibits an interesting case where a different visual pattern corresponds to ink. Figure 3.37 highlights this region, showing the infrared image alongside a texture image and enhanced texture image of the same area. Clearly, grains of some high density material such as sand or pumice are associated with ink here. These characters are the only place across the Herculaneum scroll fragments in this work where this pattern is observed. Interestingly, this fragment has sand elsewhere on the surface also, but in regions where it is not associated with ink or writing (Figure 3.38).

It is unclear why the sandy material is associated with ink in the top region of the fragment and is not elsewhere. The correlation with ink at the top would suggest some contamination of the ink itself at the time of writing, but the presence of sand elsewhere on the surface would suggest later contamination after the ink had dried. In the middle of the fragment, the sandy material looks in photographs as if it is sitting loosely atop the papyrus surface, but this does not seem to be true. Even when the fragment was mounted vertically for CT scanning, the individual grains stayed in place. They are in the exact same positions as they were when imaged two decades ago in infrared. This would suggest even those scattered particles are somehow adhered to the fragment surface. The explanation for this, too, is unclear.

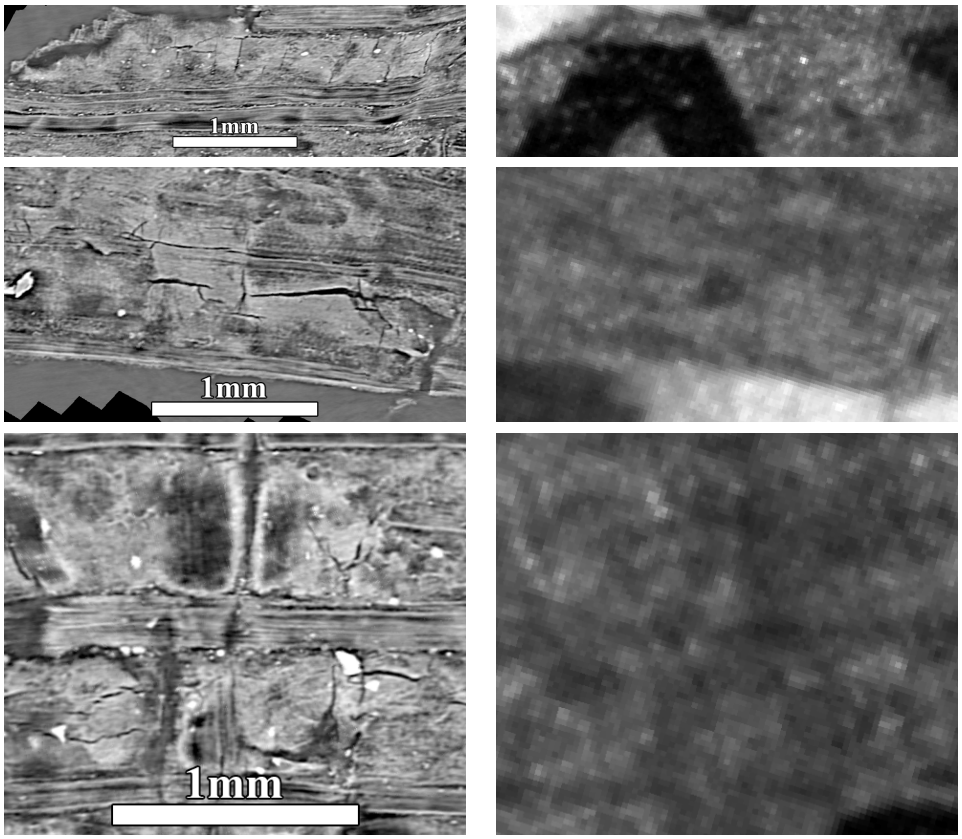
In any case, this pattern is the strongest case yet encountered of visible writing

Figure 3.31: Manually identified areas of interest on the surface of P.Herc.Paris. 2 fr. 47, overlaid on the texture image to show their position. Individual areas of interest may come from different depths within the surface volume. Inset: the same surface in infrared.



(a) X-ray CT. (b) Infrared.

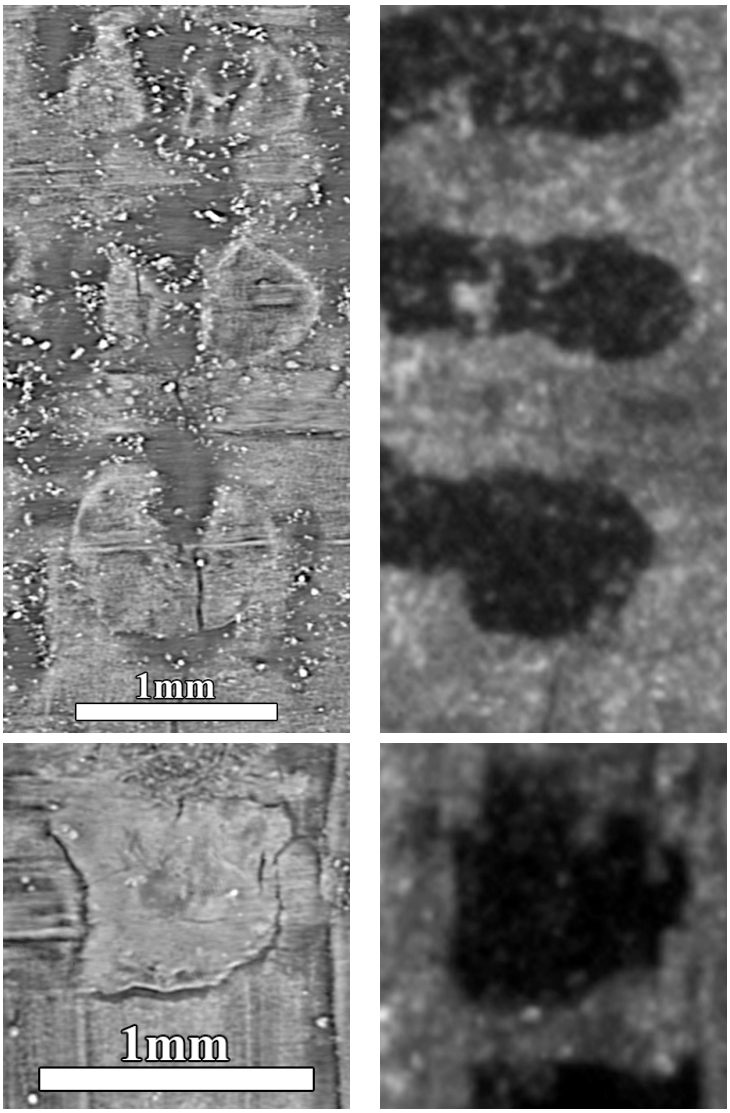
Figure 3.32: Example locations from the surface of P.Herc.Paris. 2 fr. 47 where areas of interest in CT seem to be ink.



(a) X-ray CT. (b) Infrared.

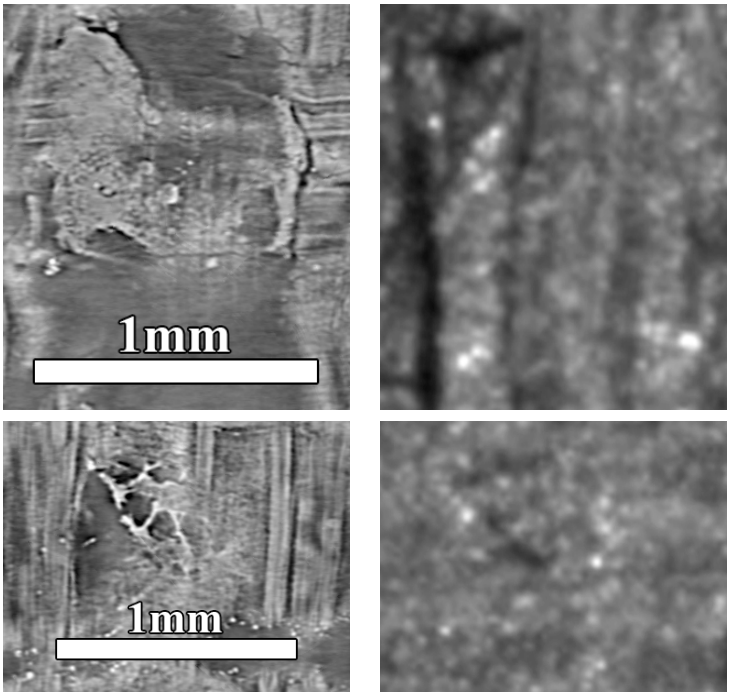
Figure 3.33: Example locations from the surface of P.Herc.Paris. 2 fr. 47 where areas of interest in CT seem *not* to be ink.

Figure 3.34: Manually identified areas of interest on the surface of P.Herc.Paris. 2 fr. 143, overlaid on the texture image to show their position. Individual areas of interest may come from different depths within the surface volume. Inset: the same surface in infrared.



(a) X-ray CT. (b) Infrared.

Figure 3.35: Example locations from the surface of P.Herc.Paris. 2 fr. 143 where areas of interest in CT seem to be ink.

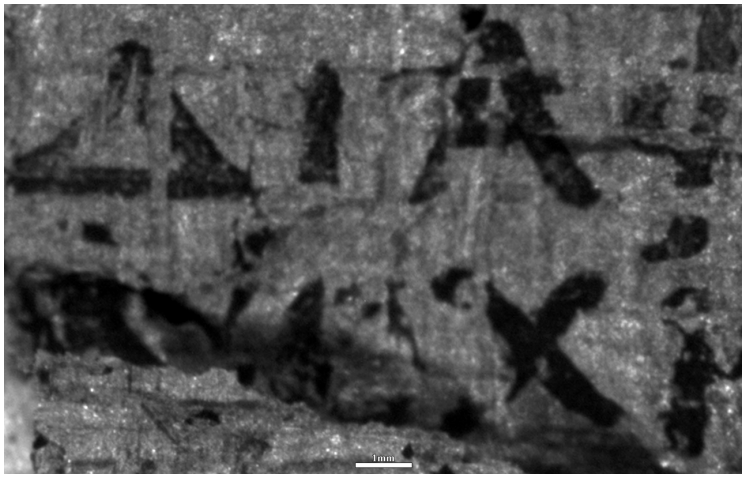


(a) X-ray CT. (b) Infrared.

Figure 3.36: Example locations from the surface of P.Herc.Paris. 2 fr. 143 where areas of interest in CT seem *not* to be ink.

in X-ray CT from Herculaneum material, and leads to two conclusions. First, the writing of the Herculaneum scrolls is occasionally visible directly in CT. Though it is a small fraction of the total text, this at least suggests there is occasional signal that could surely be recovered. Second, when compared against the other ink spots observed on this and other fragments, it becomes clear that the ink signal is itself quite varied. This is true across different scrolls, which is not surprising considering they came from different hands across a period of three centuries, but is also clearly true even within a single scroll. The variety of visual patterns corresponding to ink suggests that the signal is varied across the collection, which would imply that more training data across as wide a distribution as possible will be helpful in training models to identify ink in CT. This is consistent with the experimental path taken, where growing the training dataset always led to improved results.

The instances presented here are not meant to be complete, but rather to contribute

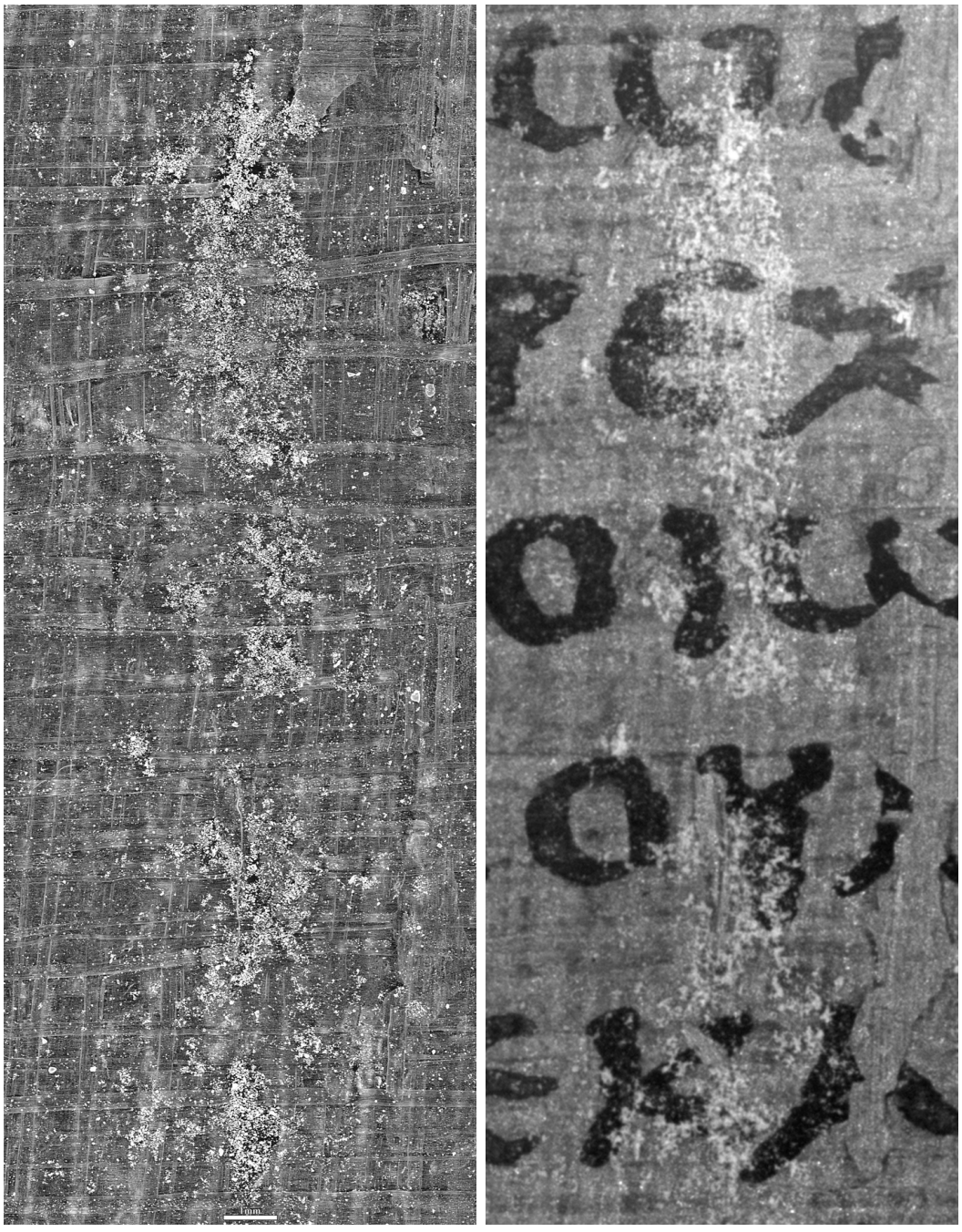


(a) Infrared.

(b) X-ray CT (texture image).

(c) X-ray CT (texture image), levels adjusted.

Figure 3.37: Location on the surface of P.Herc.Paris. 2 fr. 143 where multiple characters appear in CT with a sandy texture. High density grains were either present in the ink or later adhered to the ink here.



(a) X-ray CT. (b) Infrared.

Figure 3.38: Other sandy deposits on the surface of P.Herc.Paris. 2 fr. 143 that do not correspond to ink.

a survey of the visual patterns encountered on these surfaces in CT. These findings suggest the ink presence in CT is more complex than is sometimes suggested, either by works that claim carbon ink is invisible in CT or that text is easily read directly. Of the directly observable spots, they do follow some patterns that align with intuition. Notably, prominent ink spots in CT tend to occur at the beginning or end of individual pen strokes on the surface, where the ink has formed a thicker glob where the pen contacted or left the surface. This is consistent with basic expectations around hand written text, either with ancient reeds or modern ballpoint pens.

It is encouraging to observe these spots directly, as they suggest there is in fact a signal present that could be extracted from CT. Nonetheless, the small quantity and high false positive rate of these spots do not suggest that it is possible to recover full texts or even complete individual characters using direct inspection. Machine learning approaches, using the methods outlined in this work, are capable of recovering more text more accurately than this form of inspection. This is dissatisfying to our desire to see the ink for ourselves in the data, but I suggest it is ultimately not surprising. The high dimensionality of CT exceeds human visual capacity, and we are reduced always to 2D visualizations with less information than the raw volumetric data.

Even volume rendering, a visualization that is "more 3D" than viewing slice images, seems not to reliably capture the ink signal in a way that is visually discernable. Using the ink labels, sets of subvolumes were generated that were known to be ink and not ink respectively. They were visualized alongside their labels using volume rendering, similar to Figure 2.22b, in an unsuccessful effort to see if and visual patterns could be detected. Many transfer functions were tried for rendering, none of which seemed to enhance ink visibility. At least with the forms of visualization so far tried, learned models are able to detect something in the CT data that human eyes cannot find.

3.5 Reproducibility

The viability of ink-ID as a method for recovering the texts of the Herculaneum scrolls relies not only on validation using ground truth but also reproducibility. The method was validated using multiple techniques on fragments with known ground truth, confirming that the model is detecting ink in the correct locations. It is additionally important that all steps of the pipeline are reproducible by third parties, particularly when making claims about revealed hidden text (Chapter 5).

To this end, both the data and code from this work have been made available so that others can verify the results using the provided implementation or their own. The software pipeline components are available online for both Volume Cartographer[2] and ink-ID[3]. Where applicable, specific code and data versions are made available through archival open resources to accompany relevant publications. For example, the code and data for the original ink-ID paper [14] are available[4] through the Open Science Framework [65].

With the exception of one fragment held back for validation, the data from the Diamond fragments was released as part of the EduceLab-Scrolls dataset used in the Vesuvius Challenge. The initial release uses a slightly restricted license for the purposes of the research competition, but this will be changed to a Creative Commons license following the contest conclusion.

3.6 Summary

This section has shown the experimental results enabled by the geometric framework and machine learning pipeline introduced in Chapter 2. For the first time, the carbon ink of the Herculaneum scrolls is recovered noninvasively using X-ray micro-CT in the right conditions. Due to the associated labels, this ink detection crucially is *verifiable*, building confidence before it is applied to unseen hidden layers. When ink

[2] `https://github.com/educelab/volume-cartographer`
[3] `https://github.com/educelab/ink-id`
[4] `https://osf.io/zdkn4/`, DOI 10.17605/OSF.IO/ZDKN4

predictions are overlaid on texture images, a detailed composite image is produced purely from X-ray CT that recovers the salient surface features including text and papyrus fibers. Figures 3.39 - 3.42 display this result for the four fragments used primarily in this work.

Early efforts performed poorly, and were improved slowly through the gradual refinement of various stages of the pipeline. Improved performance resulted more often from dataset insights and refinements than from sophisticated machine learning models, and the steps of the data processing pipeline each still have room for improvement. Graphical tools were developed for various pipeline steps, enabling deeper visual understanding of the dataset and improved ink detection results. As each step is better understood and implemented, it is likely ink detection will continue to improve.

As the result of a series of incremental improvements, the experimental path taken followed essentially a depth-first search towards a viable ink detection method. The resulting geometric framework and software pipeline should therefore be considered one viable approach to ink detection. There is no indication this is the only such method, and it is possible the various pipeline components could be rethought, reordered, combined, or eliminated. Hopefully, having found one viable approach will make it easier to engage in these broader explorations. This is one of the objectives, for instance, of the ongoing Vesuvius Challenge (Chapter 7).

Using the lessons learned from developing the methods and dataset, an exploratory data analysis was conducted showing and characterizing the occasional appearances of Herculaneum ink in the X-ray CT data. This analysis shows that the appearance of ink in CT is more complex than is published elsewhere, with claims either suggesting it is invisible or can be read easily. This analysis also supports the assertion that, using existing visualization methods, trained models are able to observe some patterns in high-dimensional data that elude human eyes.

To validate these results and to encourage more work in this area, the data in this

chapter was released as a first of its kind open dataset, combining large volumetric images with aligned 2D image labels. I believe this is the largest dataset ever released in the heritage domain. Already, many people have independently reproduced the ink detection work, and have additionally been working on improvements for various parts of the pipeline.

Figure 3.39: Composite ink-ID + texture image of P.Herc.Paris. 1 fr. 34.

Figure 3.40: Composite ink-ID + texture image of P.Herc.Paris. 1 fr. 39.

Figure 3.41: Composite ink-ID + texture image of P.Herc.Paris. 2 fr. 47.

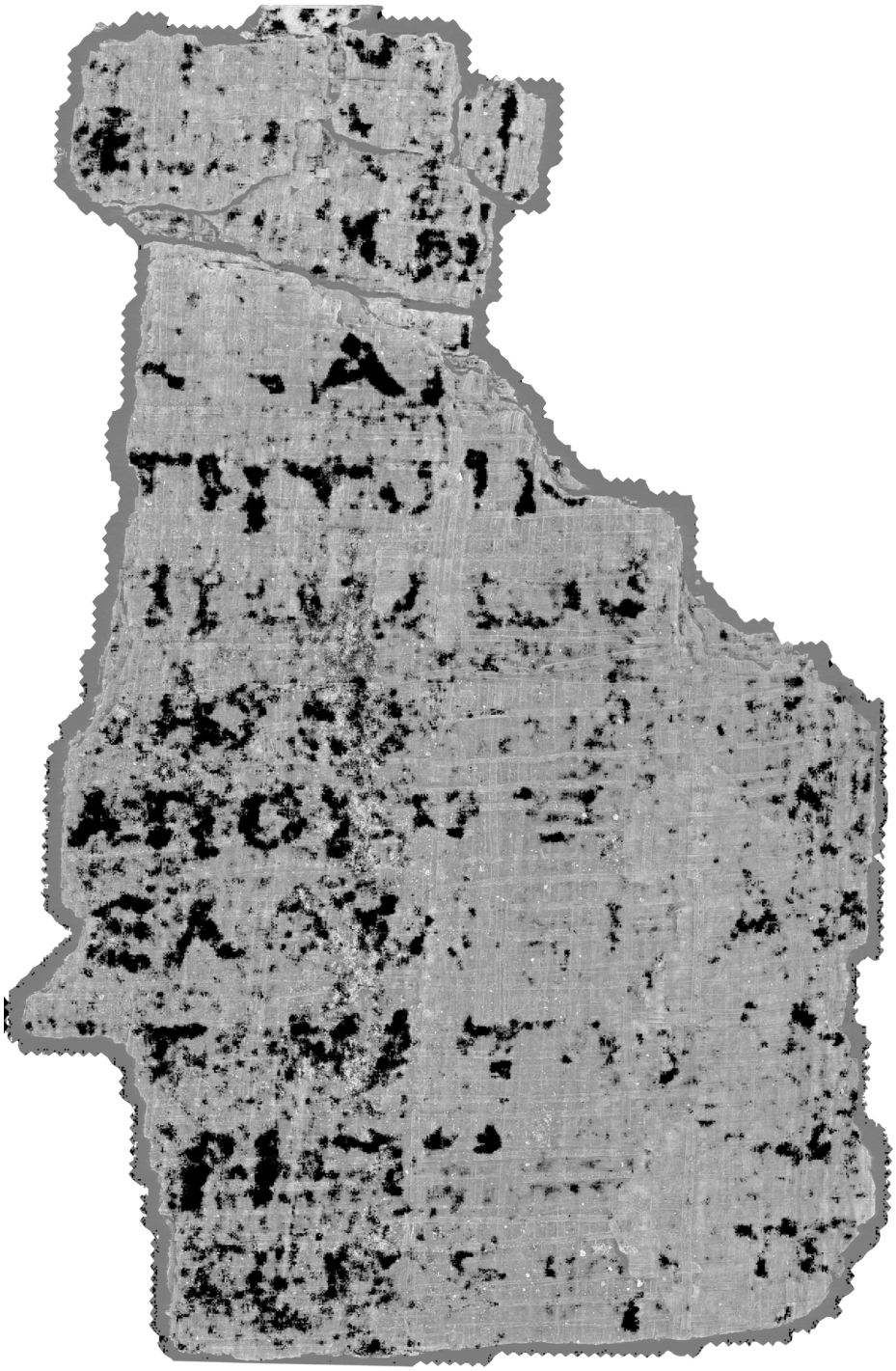


Figure 3.42: Composite ink-ID + texture image of P.Herc.Paris. 2 fr. 143.

CHAPTER 4. MULTIMODAL TRANSFORMATIONS

4.1 Introduction

Binary classification is useful in order to validate a model is capable of learning the ink presence. The implementation is straightforward, there are standard and consistent metrics such as cross entropy that make it easy to evaluate models and their progress, and the generated "prediction images" display maximum contrast between ink and not-ink predictions. Beyond proof of concept, however, binary classification has some limitations.

First, classifying voxels or neighborhoods of volumetric CT as ink or not-ink discards information. A neighborhood of such an image contains much more information than the ink presence alone: the local structure of the papyrus fibers, their orientation, the proximity to other nearby papyrus layers, and so on. A model trained with a binary ink classification task will learn only the parts of these features that are relevant to the detection of ink, and the rest will be discarded. For ink detection alone this may be acceptable, but it could be advantageous for other downstream tasks such as learned segmentation to have a model that has learned a more general and more powerful representation of the volumetric input data.

Second, another drawback of the binary classification approach is that it introduces an opportunity for human error in the labeling process. There are other such opportunities in a pipeline such as ink-ID already: the segmentation involves some manual steps, and crucially the label image registration is presently manual and not pixel perfect. Further manually labeling an infrared photograph into a binary ink/not-ink mask adds additional chance for human error. As seen in Section 3.4, there are sometimes places on the fragment surfaces where the correct ink label is not clear, even after examining both the infrared photograph and X-ray texture image, sometimes also alongside the corresponding CT slices. The resulting noisy labels can introduce

confusion to the model during training.

Finally, images generated from binary classification are useful to verify the model is detecting the ink signal, but they are not the ideal image representation for scholarly work, the ultimate goal of the generated images. First is the fact that they simply do not resemble what scholars are used to studying, which is often the physical material itself or an infrared photograph. Beyond that, though, they explicitly leave out information that is important to the process used by a scholar. The Herculaneum papyri in particular have broken and uneven surfaces, such that the exposed "surface" under study often contains letterforms from what were actually multiple, separate layers in the original rolled scroll. In order to navigate this 3D jigsaw puzzle, the scholars rely on close inspection of the surface itself, with its surface shape and the step edges between broken layers being highly important signals.

Compositing the binary ink prediction with the texture image as shown in Figures 3.39 - 3.42 achieves an approximation of an infrared photograph, but slightly different features are represented, particularly the papyrus fiber structure which has overlap but appears significantly differently between the two.

These limitations of binary classification are acceptable during the technical development of an ink detector, but could limit the ultimate utility of the resulting images to the scholars. It would be preferable to have the model learn to directly output images that more resemble what a scholar can use directly. Therefore, this chapter introduces the following contributions:

- **Multimodal transformations:** a generalized framing of ink-ID is presented, allowing models to learn image-to-image domain transformations such as CT $\rightarrow$ color or CT $\rightarrow$ infrared. These more powerful and expressive models inherently learn not only the presence of ink, but also other features such as papyrus fibers and surface cracks.

- **Domain shifts:** results show that ink-ID is capable of accurate predictions in

the presence of image domain shifts between training and inference, and that these predictions can be further improved as the domain shift is reduced.

- **Output sharpness:** results using perfectly aligned labels show that ink-ID models are capable of remarkably sharp prediction images under the right conditions, suggesting much further improvement to ink-ID results will be enabled by future work.

4.2 RGB

One approach is to train a model to predict, from X-ray CT alone, what the surface would look like under color photography. Mapping between modalities like this effectively allows one to extract images from an object interior that cannot be captured directly. In this case, the model can simulate a color photograph of hidden, interior layers, even though they are not visible to a traditional camera.

This approach requires no changes to the geometric framework, and simply exchanges the label image used during training. Instead of the binary ink mask, the label is sampled from the RGB reference photograph:

$$L(\vec{p}) \rightarrow \mathbb{Z}^3_{[0,255]} \tag{4.1}$$

As the pixel sample from RGB yields three values, the size of the model output vector is also changed to three. As the task has changed from classification to regression, the loss function is changed from binary cross entropy to a smooth L1 loss:

$$l(x,y) = \frac{\sum_{n=1}^{N} l_n}{N}, \quad l_n(x_n, y_n) = \begin{cases} 0.5(x_n - y_n)^2/\beta, & \text{if} |x_n - y_n| < \beta \\ |x_n - y_n| - 0.5\beta, & \text{otherwise} \end{cases} \tag{4.2}$$

where x is the input, y is the target, N is the batch size, and $\beta = 1.0$.

4.2.1 Carbon Phantom

Once again, the Carbon Phantom dataset is chosen for the initial proof of concept due to its controlled and deliberate construction. The Carbon Phantom shows clear ink contrast in the visible color wavelengths, so is a good candidate for the RGB prediction task. These experiments aim to recreate the surface as it would appear in a color photograph.

Using the same experimental framework as before for the Carbon Phantom, the surface is split into six columns, each processed independently as 5-fold cross-validation experiments. This time, the region bounding boxes capture not only the large carbon ink characters but also the papyrus and iron gall ink surrounding them, completely covering the surface with nonoverlapping regions. Once the models are trained, the final generated images from their respective prediction regions are composited (results shown in Figure 4.1).

As can be seen, the model implicitly learns the presence of the ink quite well in this setup, even though the ink was not explicitly labeled. Instead, it has learned this signal directly from the RGB color space. The ability of the model to detect the carbon ink in the various columns is on par with its performance in the binary ink classification setup, indicating that the ink signal was learned just as well without the explicit labeling.

Additionally, there is now much more information present in the generated images. Instead of ink presence alone, the model has also learned to represent the papyrus fiber structure in the resulting image. This makes the image itself more useful, but also suggests the model has learned a more general representation that may be more powerful for other downstream learning tasks.

4.3 Infrared

This concept is not limited to color photography. A model can be trained to predict the appearance of a surface in any modality, such as in the infrared wavelengths.

Figure 4.1: ink-ID results on RGB task for Carbon Phantom. Surface divided into six columns, each treated as separate five-fold cross-validation training experiment. Image shown is combined result of 30 trained models. From X-ray CT alone, ink-ID can recover not only carbon ink but iron gall ink and papyrus fiber structure.

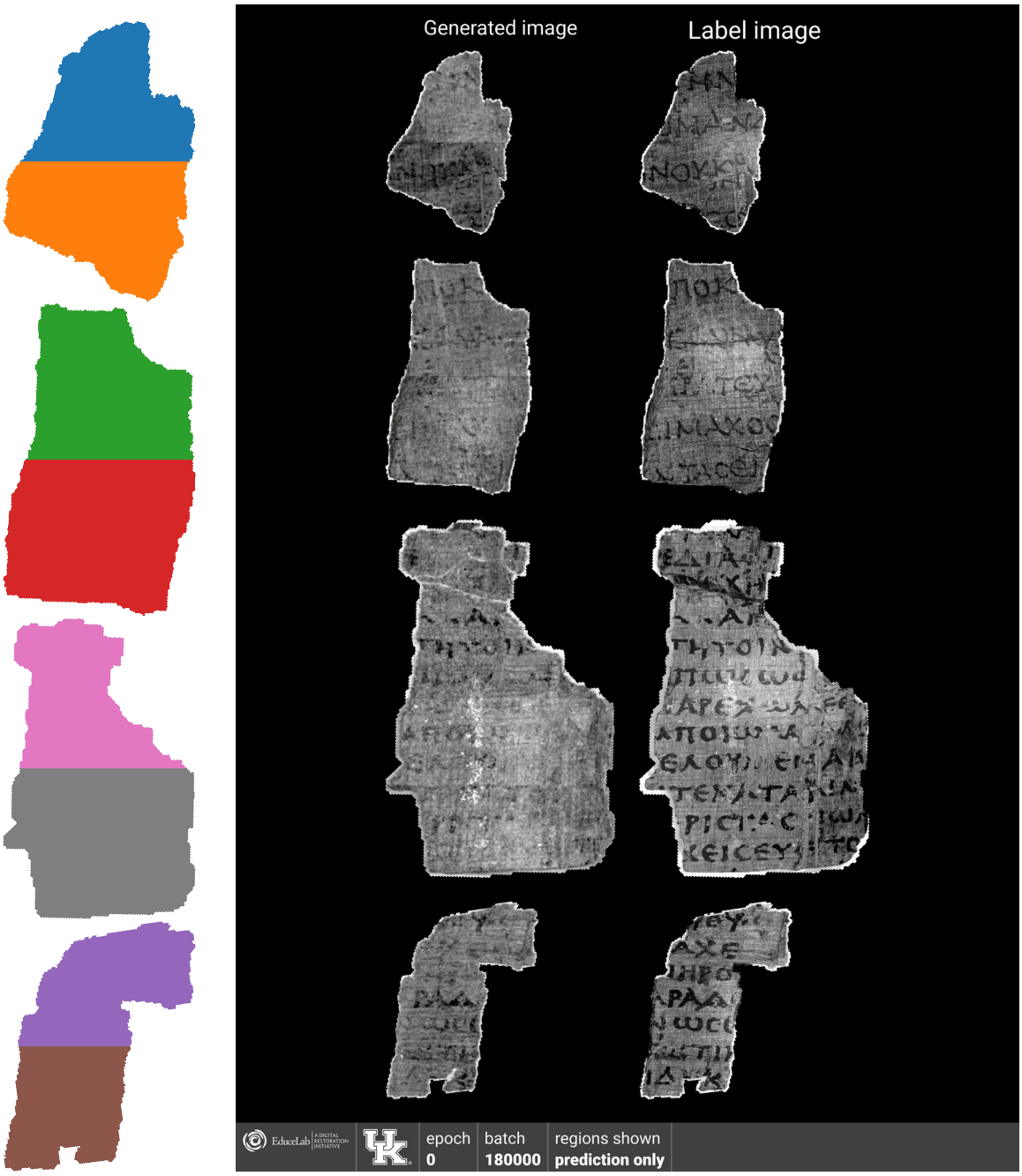
4.3.1 Herculaneum

Genuine RGB labels for the Herculaneum scroll fragments are available, but not necessarily desirable. The damage due to carbonization creates a low contrast black-on-black visual in which it is difficult in many places to observe the text (see Figure 1.3 or 3.5). It is for precisely this reason that infrared photography became the standard for digitization of this collection, as it reveals a higher contrast image showing the text much more clearly.

Figure 4.2 demonstrates the ability of ink-ID to reproduce the infrared appearance from X-ray CT. In addition to learning to detect ink, the model is capable of rendering papyrus fibers as well as other surface features such as the light-colored sandy material on the surface of P.Herc.Paris. 2 fr. 143.

The output of this infrared prediction is also useful to examine the downstream effects of possible labeling error introduced during the binary labeling step. Like the other steps in the pipeline, binary labeling is known to be imperfect, and introduces the opportunity for error, either where spots are mislabeled or where spots do not neatly conform to the ink/no-ink classification. Directly predicting the infrared appearance skips the binary labeling step, so there is no opportunity for this particular error to be introduced. If the labeling noise from binarization significantly impacts the ability of a model to recover ink, then the direct infrared predictions should show more legible text.

This does not seem to be the case, as the text recovery in the infrared predictions is on par with but does not exceed that of the binary classification predictions in Figure 3.19. This is based on a visual evaluation, as it is difficult to quantitatively compare these predictions on different tasks or modalities. It is also possible the text contrast in the infrared predictions could be further boosted by more dramatically adjusting the contrast in the infrared label images. That said, it fortunately seems clear that binary labeling does not introduce enough error to significantly impede the model's



(a) 8-fold. (b) ink-ID infrared predictions.

Figure 4.2: ink-ID results across the Diamond fragments, using the infrared images as training labels. 8-fold cross-validation used.

ability to detect ink.

4.4 Simulated RGB

Particularly where genuine RGB labels are not desirable, a simulated "RGB" approach creates a new opportunity: by creating the desired RGB labels manually using some other label image, one can control the generated images such that they take on any desired appearance.

4.4.1 Herculaneum

In the case of the Herculaneum scrolls, the model can learn to undo the damage from thousands of years ago, creating an image that represents what the scroll could have looked like to the eye when it was originally written.

Figure 4.3 shows an example "RGB" image manually created for P.Herc.Paris. 2 fr. 47. The original infrared photograph is pictured alongside the manually created RGB version. The RGB image is created in Photoshop through a combination of contrast stretching, colorization, and levels adjustment. The binary labels are also utilized to select and further darken the characters on the surface, boosting the color contrast between papyrus and ink. In the interests of creating a consistent dataset for training, the same exact settings are applied to the infrared photographs for each of the four Diamond fragments.

Figure 4.4 shows the results of using this training objective under the same training framework as used in the other 8-fold experiments across these fragments. As can be seen, the model is capable of generating realistic color images of the fragment surfaces, recovering the ink signal as well as it does when trained as a binary classifier.

4.5 MS M.910

Learning multimodal transformations can manifest in ways other than simply altering the target imaging method (e.g. visible color or infrared). Consider for example a two-sided writing surface imaged in CT, where one wishes to generate images that differentiate the writing on each side. The Herculaneum scrolls are not a good fit for

Figure 4.3: Simulated RGB image made for P.Herc.Paris. 2 fr. 47 from the infrared, imagining what the surface may have looked like when it was written. Binary label masks used to artificially darken ink regions.

this particular problem as the text is almost always on only one side of the papyrus sheet. Even if the writing were two-sided, the physical papyrus layers are so fragile that there are not single-layer fragments able to be imaged from both sides. That said, there are other datasets that are a better fit.

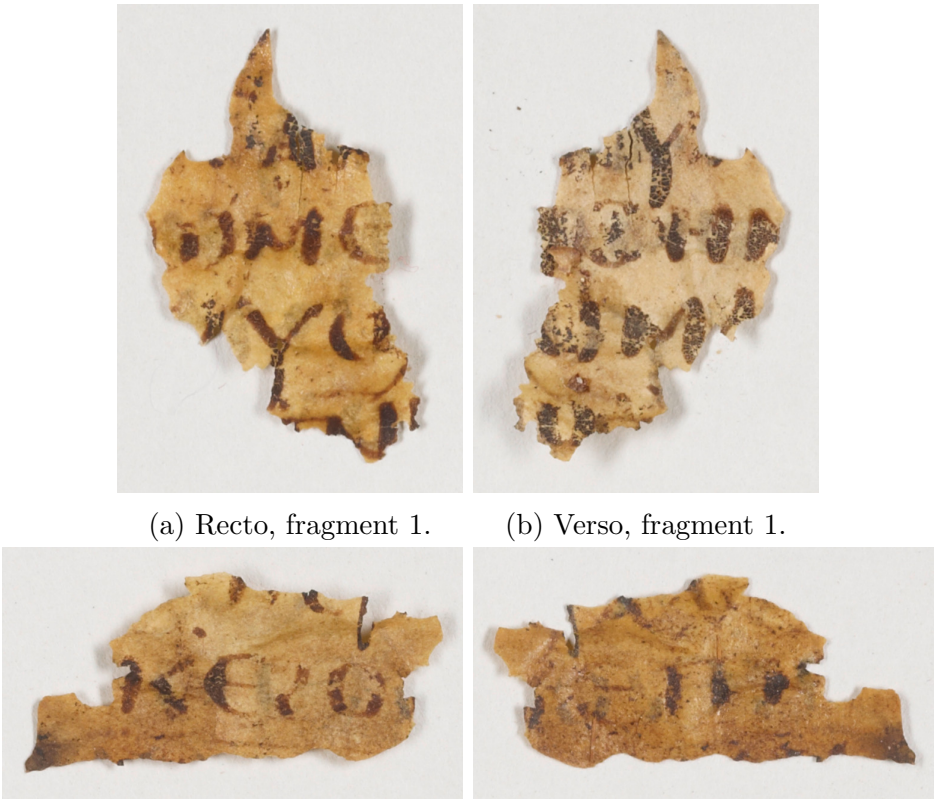
The M.910 manuscript (Figure 4.5) is a parchment codex stored at the Pierpont Morgan Library, containing a version of *Acts of the Apostles* in Coptic. The manuscript has been damaged by water and heat, so has warped pages that are too fragile to open physically. The pages also contain writing on both sides, so any non-invasive method to recover the text needs to be able to differentiate the recto and verso. The text is also written with an iron gall ink that reveals contrast in X-ray, easing the challenge of detecting the ink at all, and allowing the experiment to focus instead on the two-sided nature of the pages. MS M.910 also provides the opportunity to examine ink-ID's performance in the presence of a domain shift between training and inference CT images, a concept that becomes important with the Herculaneum scrolls.

(a) 8-fold.

(b) ink-ID "RGB" predictions.

Figure 4.4: ink-ID results across the Diamond fragments, using the simulated RGB images as training labels. 8-fold cross-validation used.

Figure 4.5: The M.910 manuscript. Photo courtesy of the Morgan Library and Museum.



(a) Recto, fragment 1. (b) Verso, fragment 1.

(c) Recto, fragment 2. (d) Verso, fragment 2.

Figure 4.6: MS M.910 fragments. Color photographs. Photos courtesy of the Morgan Library and Museum.

4.5.1 Detached fragments

In addition to the primary M.910 manuscript, there are also some detached parchment fragments. As they are each a single layer, have text on both sides, and are robust enough to withstand mounting and imaging, they are a good candidate to test whether ink-ID can differentiate between the writing on each side. The recto and verso of each fragment are shown in Figure 4.6.

An ink-ID experiment was conducted to see whether the model is capable of distinguishing recto from verso. The fragments were first both CT scanned at 8.00 µm voxel size and processed with the data pipeline. Segmentation was performed through the middle of the parchment cross section, as shown in Figure 2.4b. Each fragment was only segmented once, but two texture images were generated, using directional

sampling in opposing directions: one following the surface normal vectors to sample the recto, and one following inverted normal vectors sampling the verso. The color reference photographs in Figure 4.6 were aligned to the texture images to create RGB labels.

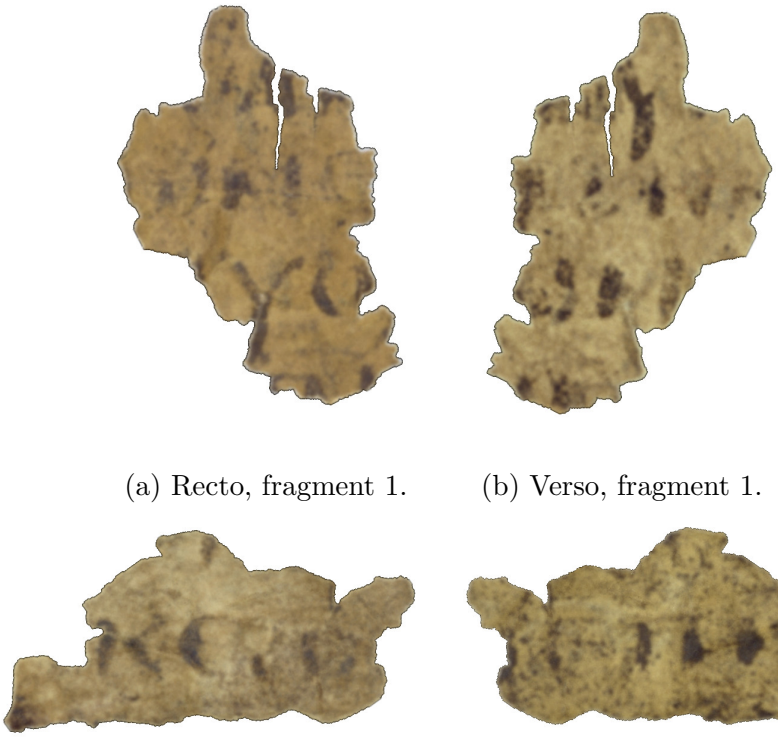
ink-ID was then trained in a 4-fold cross-validation across the four fragment faces. Unlike the subvolumes from Canny segmentation on the Herculaneum scroll fragments, which nominally capture only the recto face, the subvolumes sampled from these fragments are centered in the middle of the parchment layer and capture both the recto and verso faces. The ink-ID model has to learn to differentiate between ink on the top face in the subvolume, and that on the bottom face, so it cannot simply report whether there is a bright spot detected. Figure 4.7 shows the results of the 4-fold experiment, confirming that ink-ID has no trouble distinguishing recto from verso when the bidirectional ink signal is captured clearly in the CT data.

4.5.2 Manuscript

Section 4.5.1 showed that ink-ID is capable of recreating the two-sided appearance of the MS M.910 fragments in color. How does ink-ID perform when trained on the fragments and then applied to the internal surfaces of manuscript, which was imaged separately at a different resolution? For reference, Figure 4.8 shows the texture image for an internal page generated from virtual unwrapping. Writing is visible due to the absorption contrast in X-ray between ink and parchment, but the image still visually resembles X-ray instead of a more intuitive modality such as color. Bleed-through is also visible, as ink from both sides of the page is present in this texture image.

Initial ink-ID output

Figure 4.9 shows the results of training ink-ID on fragments 1 and 2, and generating prediction images for the same internal page from the manuscript. Subvolumes of size $48 \times 48 \times 48$ voxels were sampled in all cases. As the fragments were scanned at 8.00 µm and the manuscript was scanned at 35.69 µm, the spatial dimensions of



(a) Recto, fragment 1. (b) Verso, fragment 1.

(c) Recto, fragment 2. (d) Verso, fragment 2.

Figure 4.7: 4-fold ink-ID results on MS M.910 fragments. Color images rendered purely from X-ray CT inputs.

Figure 4.8: Virtual unwrapping texture image of internal page of MS M.910. Ink is visible due to absorption contrast in X-ray.

Figure 4.9: Initial ink-ID results when trained on fragments and applied to internal page of MS M.910 without spatial subvolume sampling or other domain adaptation measures.

the subvolumes from the fragments ($384 \times 384 \times 384$ µm) differed from that of those from the manuscript ($1.713 \times 1.713 \times 1.713$ mm). Surprisingly, ink-ID is capable of generating a reasonable prediction image, showing the correct location of some text on the surface.

Domain shift

In this experiment, there is a domain shift between the training images from the fragments and the inference images from the internal page of the manuscript. This difference is characterized primarily by resolution or voxel size: a subvolume of certain voxel dimensions corresponds to different spatial extents in the different domains. This affects the apparent thickness of the parchment sheet in the subvolume, as well as the nature of the ink signal on the parchment surface. As will be shown in later sections, this domain shift seems narrower than that evident between the Herculaneum scrolls and fragments. Nonetheless, it is helpful to see an example case where ink-ID is capable of reasonably bridging a domain shift even without domain adaptation measures, and further it can be shown that by taking some basic measures,

Figure 4.10: ink-ID results when trained on fragments and applied to internal page of MS M.910 with spatial sampling, ensuring training and inference subvolumes have the same spatial dimensions.

the performance can be improved significantly even if the domain shift is never fully eliminated.

Spatial subvolume sampling

One simple way to close this gap is to sample subvolumes of equal spatial dimensions from both domains. The subvolume size in voxels can be kept constant by oversampling or undersampling the CT volume. This principle is visualized in Figure 2.21a. Figure 4.10 shows the results of an ink-ID experiment where all subvolumes were set to spatial dimensions $384 \times 384 \times 384$ µm. This means the fragment subvolumes during training were sampled at their "native" resolution, and the subvolumes from the internal layer of the manuscript were oversampled. Oversampling changes the nature of the domain shift: now the physical features present in each subvolume are similar, but the training subvolumes capture higher frequency details than the blurrier inference subvolumes. Empirically it can be seen that this change was helpful: the ink-ID prediction image now shows much sharper details.

Two-sided pages

Though ink-ID could cleanly distinguish the recto and verso on the higher resolution fragment data (Figure 4.7), it is unclear whether this also transfers well to the lower resolution manuscript data. To check this, another prediction was generated for the same internal surface using inverted surface normal vectors to sample the verso. The result is shown in Figure 4.11. The correct orientation of this surface would be horizontally mirrored from that of the recto, but it is shown in the same orientation for easy visual comparison. Though some salient visual features do change between the two predictions, indicating some difference extracted by the two-sided approach, there is clear bleed-through visible also. This suggests the model is unable to completely distinguish the two sides. This makes sense considering the two-sided cross section of the parchment sheet is much less defined in the lower resolution manuscript scan, and the model has only been trained on sharply differentiated sides in the high resolution fragment scans. Perhaps this particular domain shift could be bridged further with downsampling during training, so the model learns to distinguish the two sides even when their ink signals are muddled together.

4.5.3 Discussion

This section has presented ink-ID results on the M.910 manuscript. MS M.910 is an instructive case study in a few ways. First, the pages have text on the recto and verso, and the detached fragments usable for training share this feature. This enables the training of models that are capable of distinguishing recto and verso, suggesting they have spatial reasoning through the depth of the subvolume.

MS M.910 is also helpful for studying a domain shift between training and inference data. The domain shift is largely characterized by resolution. It is shown that, where the signal is strong enough, ink-ID is capable of reasonably bridging some domain shifts even without special measures taken. The performance is only improved when measures such as spatial sampling are taken to minimize the domain shift. Even with

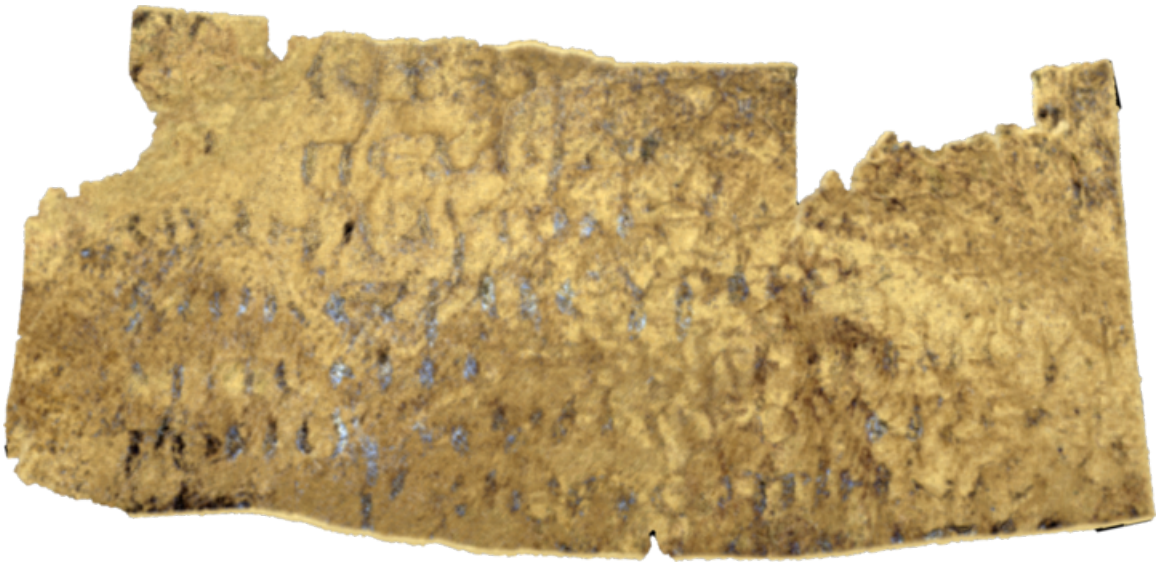


Figure 4.11: ink-ID results when trained on fragments and applied to internal page of MS M.910, applied twice with opposing normal vectors to sample recto and verso. Horizontally mirrored for visual comparison with Figure 4.10, demonstrating some bleed-through.

few or no special measures taken, ink-ID can perform well across separate CT scans of different resolutions.

This brief exploration with MS M.910 is featured primarily to suggest the larger domain shift of the Herculaneum scrolls and fragments can be bridged, or reduced enough for useful results. As the domain shift of the Herculaneum scrolls is larger and the ink signal is more subtle to begin with, this is helpful to keep in mind during initial experiments.

4.6 Dead Sea Scrolls fragment

So far, the ink-ID experiments have featured labels that are not perfectly aligned. The registration of the label image to the texture image is not pixel perfect, and like the other steps of the data processing pipeline introduces small errors. Empirical ink-ID results show that ink detection is possible even in the presence of these accumulated errors, but it is helpful to know what may be possible if perfect label alignment can be achieved or models can be trained with improved tolerance to misalignment. Assuming perfect labels, what is the ceiling?

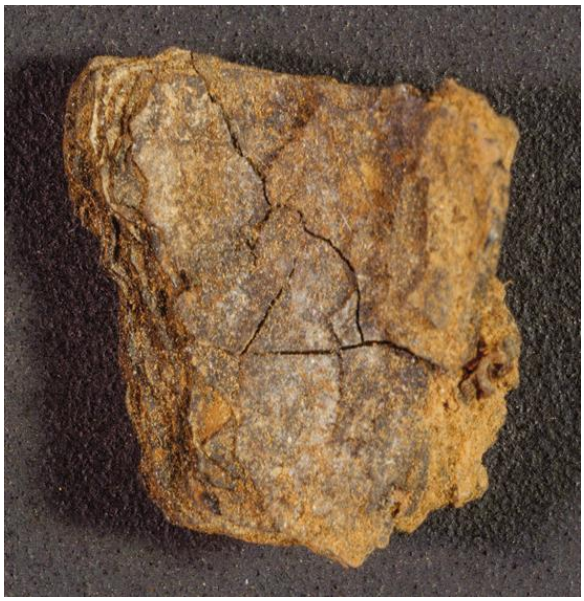


Figure 4.12: Dead Sea Scroll fragment 1032a, color photograph. Photo courtesy of the Leon Levy Dead Sea Scrolls Digital Library.

Another manuscript gives an opportunity to explore this question. This experiment uses CT data from Dead Sea Scroll fragment 1032a, pictured in Figure 4.12. This experiment will create a set of labels with known perfect alignment, and then examine the ink-ID output to see what the model is capable of.

The manuscript has parchment pages and an ink that appears with strong contrast in X-ray. Figure 4.13 shows the results of segmenting five internal layers from the fragment and running them through the virtual unwrapping pipeline. The texture images clearly show Hebrew characters in great detail.

Next, a "color" label was generated for only one of these recovered layers. This was done by taking the texture image, which as mentioned already reveals the ink presence, and modifying it manually in Photoshop. As with the "color" labels for the Herculaneum papyri, the modifications are done to generate an image representing what this surface could look like if photographed in its original, undamaged state. The result of this process is shown in Figure 4.14.

Perfect alignment is achieved, because the label images are synthesized directly

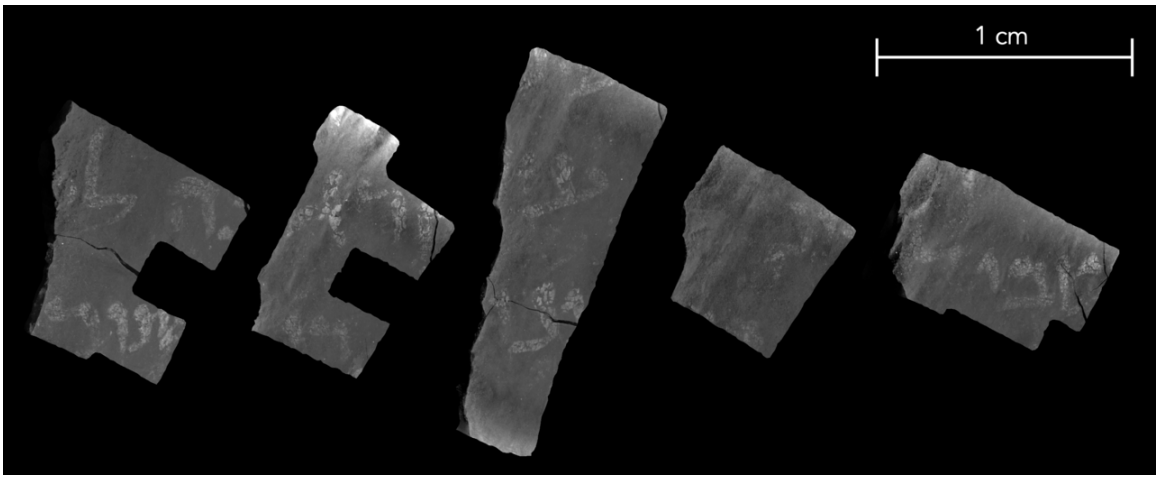


Figure 4.13: Texture images of five layers segmented from inside fragment 1032a.

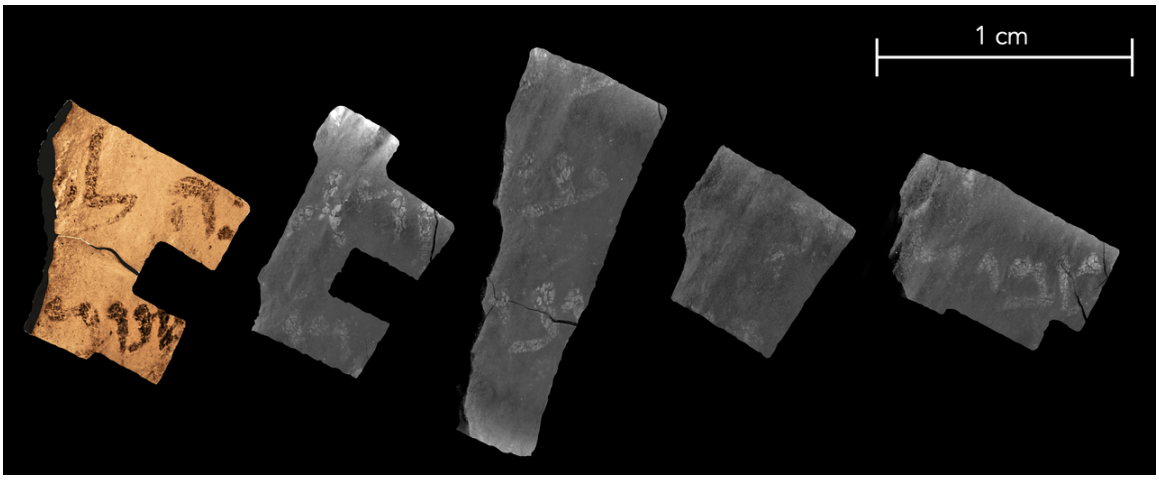


Figure 4.14: Leftmost surface manually colorized to simulate its appearance before damage.

Figure 4.15: Using leftmost surface as training data, ink-ID predicts on the other surfaces. Very sharp images are produced, suggesting model architecture is capable of high spatial precision when trained with perfectly aligned label inputs.

from the texture images, ensuring each pixel of the two images is aligned perfectly. As the texture image itself is generated from the PPM image space, its pixels are in alignment with the sampled subvolumes by definition.

Finally, an RGB ink-ID model is trained using only this fragment as training data. The model is then applied to the other layers during inference, generating RGB prediction images. These generated images are shown in Figure 4.15.

As can be seen, these generated images are much sharper than any others so far generated using ink-ID. This finding is important for two reasons. First, it shows that the model architecture itself is not limited in sharpness, but is capable of generating accurate per-pixel predictions. Though other model architectures may be sought for increased general performance, the existing implementation is already capable of sharp predictions given sufficiently aligned training data. Second, it hints at what will become possible for other datasets such as the Herculaneum papyri, once either label alignment is further improved or models are created with more noise tolerance. Eventually, it will become possible to generate sharp images like those shown here for the interiors of even the Herculaneum scrolls.

4.7 Summary

This chapter has presented an extension to the ink identification framework of Chapter 2, generalizing it to learn image mapping between modalities. In the example of the Herculaneum scrolls, this means hidden surfaces can be imaged in X-ray CT and then viewed as if they had been imaged in infrared. By coloring the infrared training images, it is possible to simulate the possible appearance of these surfaces before they were damaged by the eruption.

This can be applied with other manuscripts as well. For manuscripts with two-sided pages, ink-ID is capable of distinguishing the recto and verso. Further, in the right circumstances, ink-ID can generalize well across CT scans, even of differing resolutions. When basic steps are taken to reduce the domain shift between different CT scans, this performance further improves.

In other works, false coloring of X-ray or CT images is sometimes used to simulate certain visual appearances. This is a legitimate visualization method, but is only capable of extracting what is already visible in the X-ray image. For instance, imagine trying to false color the texture image of a scroll fragment to resemble its appearance in infrared or color photographs. There would be no ink or text evident in the resulting photograph, as there is none visible in the source texture image. The presented method is capable of more: any pattern that is recoverable by the machine learning model can be mapped to the target domain. In the case of the Herculaneum papyri, this means not only the ink, but cracks, papyrus fibers, holes, sand, and other features are also visualized. This is accomplished without explicit labeling of the desired features, instead automatically capturing anything appearing in the target domain that is detectable in the source.

Using the IAA-1032a manuscript fragment, this chapter also examined the possibilities enabled by perfect label alignment. With improved label alignment, ink-ID models are capable of remarkably sharp prediction images. As methods improve, I

suggest sharp images like these will be produced of the Herculaneum scroll interiors.

This work has introduced the concept with X-ray → infrared or X-ray → RGB, but neither the source or target domains are limited to these modalities. With respect to the Herculaneum scrolls, at least the following configurations could be promising. Multiple registered CT scans of varying incident energies could be combined to a single multi-energy input, which may increase the detectable signals. The output could similarly be extended from infrared or RGB to multi- or hyper-spectral image stacks, extracting many wavelengths simultaneously. This could even be reversed, possibly enabling spectral → CT models that could predict the partial volumetric appearance of an object from surface photography.

Damaged manuscripts have been used for the introduction of these ideas, but this paradigm could be applied to other domains. As an example, consider noninvasive medical imaging. Most machine learning-based detection work using medical images relies on human expert labeling. This work suggests there may be more information captured in the images than could possibly be annotated by a human. As with the scroll fragments, specimens could instead be imaged in two modalities, one the noninvasive method of interest (e.g. CT) and one that highlights the features of interest more clearly (e.g. biopsy). By using these aligned images as a labeled dataset, perhaps more could be extracted from the noninvasive modalities than is presently assumed possible.

CHAPTER 5. REVEALING HIDDEN TEXTS

5.1 Introduction

The cross-validation experiments on fragment surfaces are exciting proofs of concept but are ultimately just that: the end goal of learning an ink detector is to apply it to unseen, hidden layers, revealing new texts. Regarding the Herculaneum collection in particular, there are multiple options for where to pursue these hidden layers. In addition to the hidden layers of the intact, rolled scrolls, the fragments themselves have subsurface layers with unknown text.

As with other parts of this work, an experimental pathway is chosen to start with the lowest hanging fruit, and iterate from there. This path starts with some particular layers from within the scroll fragments, as they are more easily segmented and come from the same CT scans as the training data on the fragment surfaces, minimizing the domain gap between training and inference. This chapter will show that ink-ID is able to reveal some Greek characters from these layers, noninvasively recovering hidden Herculaneum texts for the first time in history.

The hidden layer segmentation method developed in this chapter is also applied to the intact scrolls, yielding the largest segmentations ever generated of papyrus surfaces from within Herculaneum scrolls. There are multiple obstacles to ink detection from within the intact scrolls that were not present for the fragments, primarily concerning the difference in image distributions between the CT scans corresponding to the training and inference datasets. Though ink-ID has not yet revealed text from these surfaces, this chapter discusses some efforts to get around these obstacles and presents a case for optimism in the next steps. The contributions of this chapter can be summarized as follows:

- **Coarse-to-fine segmentation method:** a novel segmentation method is introduced, enabling state of the art segmentation results on the hidden layers of

the Herculaneum scrolls.

- **Hidden text:** using the above segmentation method, written characters are revealed from the hidden layers of the Herculaneum scroll fragments, marking the first noninvasive recovery of hidden text from the Herculaneum papyri.

- **Domain shift:** the domain shift between fragment and intact scroll CT scans is characterized, and next steps are suggested for ink detection on the intact Herculaneum scrolls.

5.2 Segmentation

For the Herculaneum papyri, a new segmentation method was required for the hidden layers of either the fragments or the intact scrolls. As discussed in Chapter 6, ink-ID is at least for the time being quite sensitive to segmentation, relying on a mesh that closely follows the papyrus surface. The Canny edge detector-based method developed in Section 2.3.3 accomplishes this level of precision, but relies on an exposed surface without occlusion by other papyrus layers, as the rays projected from the image boundary stop at the first edge encountered. Segmentation methods previously developed for virtual unwrapping [21] are designed for internal layers, but have other problems: they follow the layer midpoint instead of the layer surface, they are not as precise, and they are prohibitively slow for the large CT volumes used in this work.

5.2.1 Quick Segment

A modification to the Canny edge detector-based method was thus designed to target the internal layers with the required precision. The approach is a coarse-to-fine segmentation approach, combining minimal user input with an automated step that fine-tunes the segmentation. The first step uses a graphical user interface called Quick Segment [66], allowing one to manually create a coarse segmentation near the surface of interest. The second step is similar to the aforementioned Canny

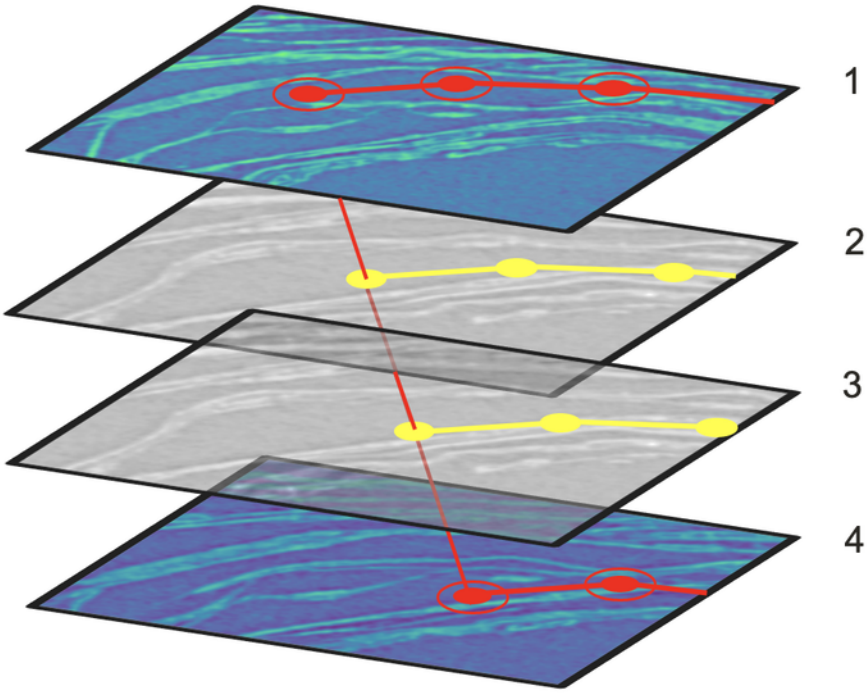


Figure 5.1: Linear interpolation used during manual, coarse stage of Quick Segment. User specifies line segments in selected "keyframe" slices throughout the volume (1 and 4), and lines are interpolated in the slices between (2 and 3). Image from [66].

segmentation method, but projects the rays from the normal vectors of the initial coarse segmentation, rather than from the image boundaries.

Initial coarse segmentation

In the initial coarse stage of this method, a user manually traces a set of connected line segments in a slice of the CT volume. This segmentation relies on targeting layers with some separation, as the user segments not the writing surface itself, but the nearby air gap between the writing surface and the next layer. The graphical interface allows the user to navigate quickly through the volume slices, labeling "keyframe" slices where desired. For all other slices, the points of the line segments are linearly interpolated between the nearest labeled slices. Figure 5.1 visualizes this process.

The user labels as many slices as is necessary to establish a segmentation that accurately remains within the target air gap throughout the region of interest. Once

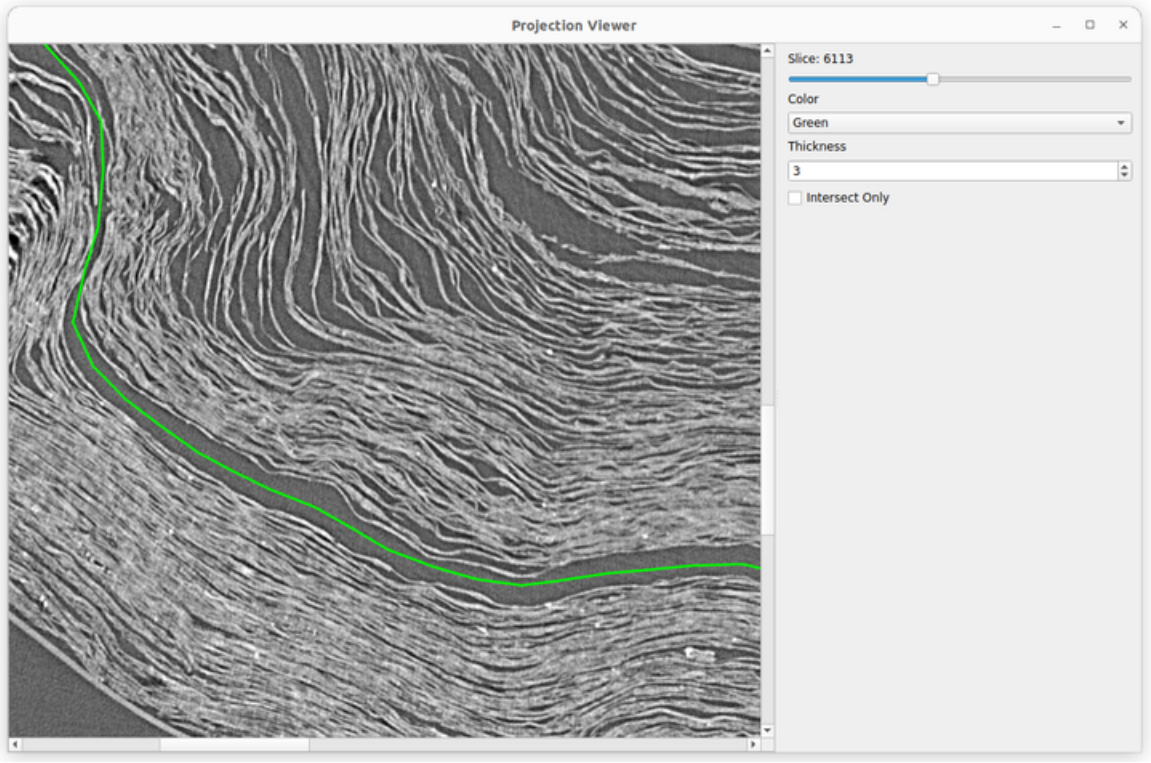


Figure 5.2: Initial manual coarse segmentation follows not the surface of interest but the neighboring gap between layers. Zoomed region of P.Herc.Paris. 3 shown.

completed, the program interpolates the line segments for all segments between the first and last keyframes, and generates a point cloud that can be meshed. Figure 5.2 shows the resulting mesh for an example segmentation in P.Herc.Paris. 3. Where possible, the mesh follows the air gap. In places, where the layers are pressed together, this segmentation cuts into the papyrus surface.

Fine-grained segmentation

The fine-tuning step is a method similar to the Canny edge detector-based method of Section 2.3.3, where the manually segmented mesh is used for the ray origins rather than the image boundary. The Canny edge detector parameters for the dataset are first fine-tuned graphically, in the same way as is done when segmenting the exposed fragment surfaces (Figure 5.3). The selected Canny parameters are typically the same for this internal segmentation as they would be for an exposed surface segmentation

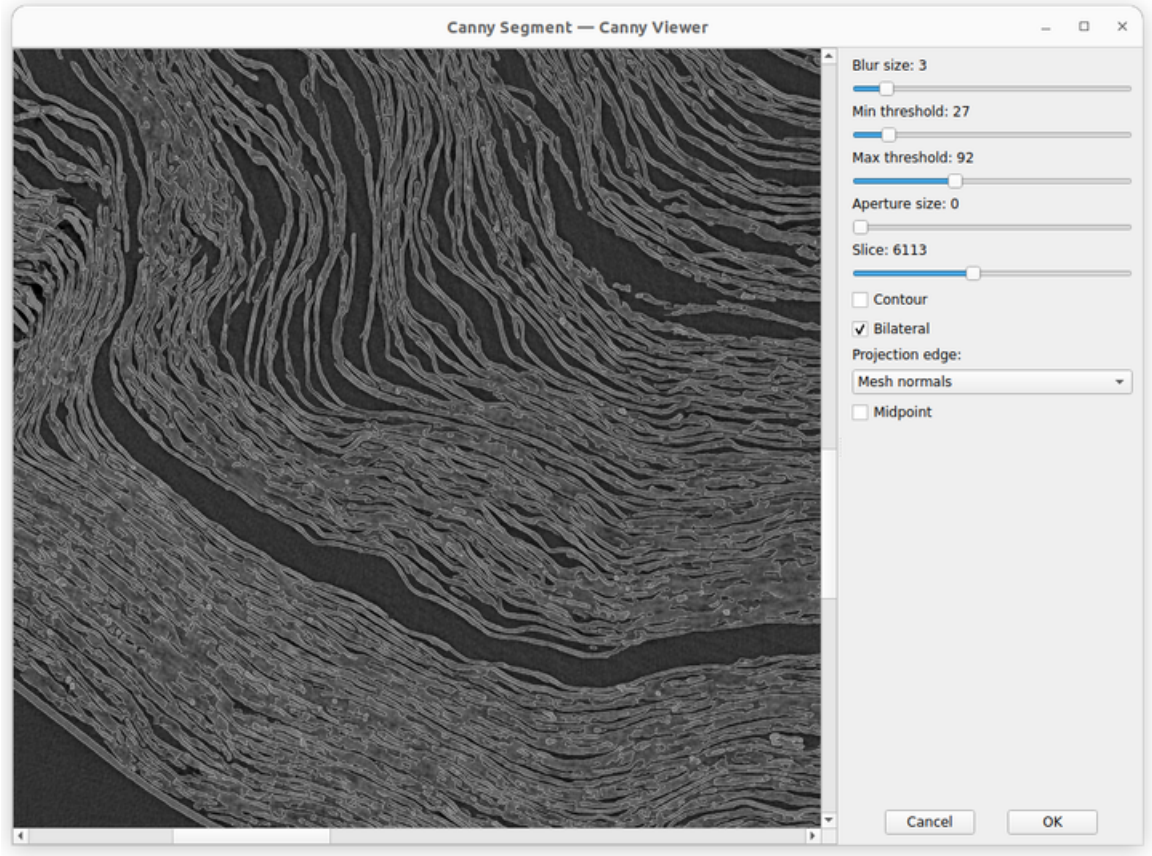


Figure 5.3: Canny edge detector parameters are adjusted until they reliably detect papyrus surfaces near segmentation of interest. Detected edges shown as thin white lines against darkened slice image. Zoomed region of P.Herc.Paris. 3 shown.

of the same object; however, the final contouring step often applied to smooth the detected edges on an exposed surface does not work well in internal areas, so it is not applied.

Next, within each CT slice, rays are projected from the points along the mesh intersection line. The ray direction from each point is set to the mesh surface normal vector at that location. As before, the rays continue until encountering a detected edge, and this intersection point is added to a point cloud. When sampling along a ray, the algorithm moves in 1/2 pixel increments in order to reduce the chance a detected edge is missed due to aliasing. This step is described in Algorithm 4 and visualized in Figure 5.4.

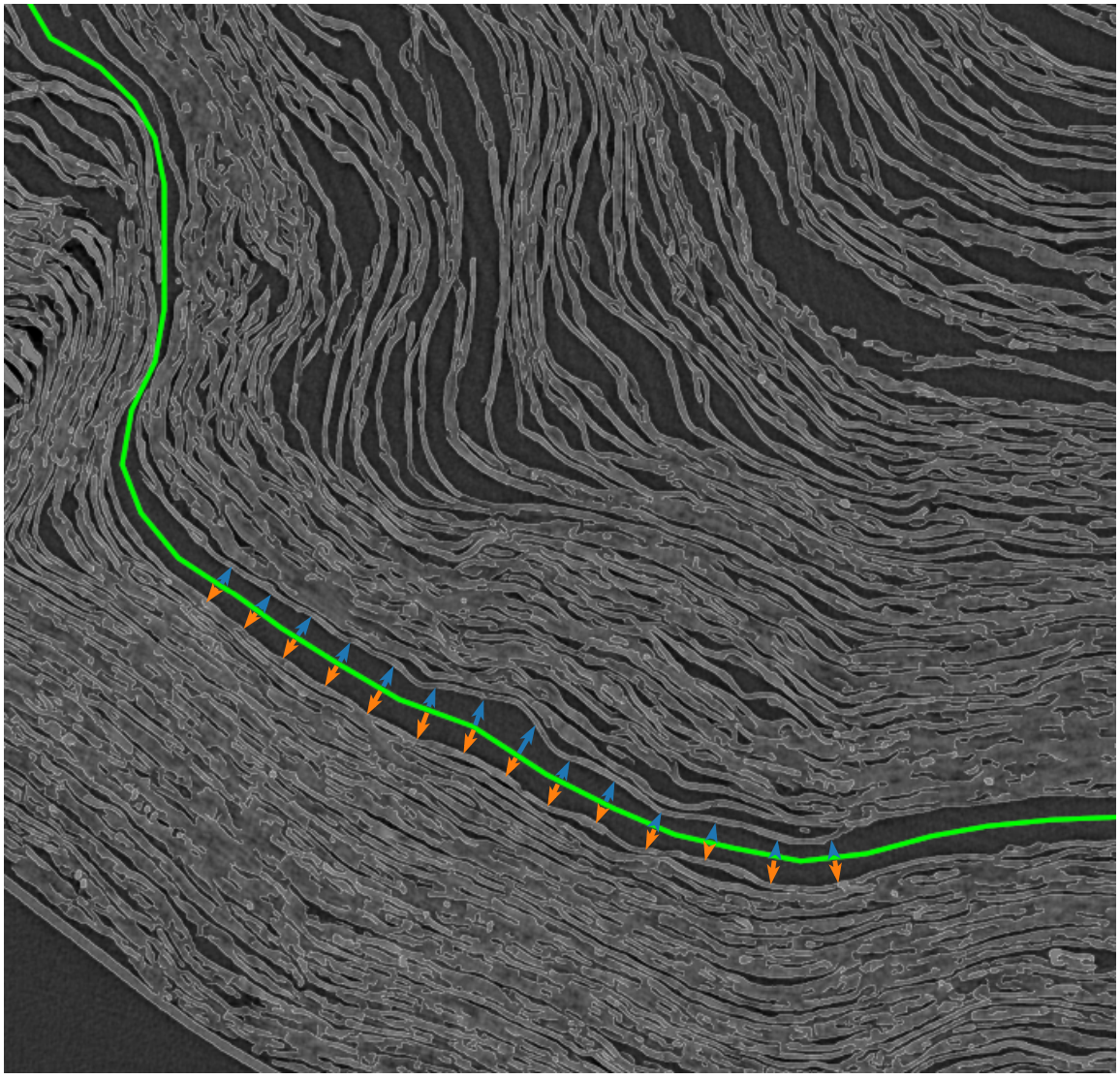


Figure 5.4: Visualization of Algorithm 4. Rays (orange) are projected from mesh intersection (green) along mesh surface normal vectors until encountering detected edges. By optionally inverting surface normal vectors prior to projection (blue), surface on other side of air gap can also be recovered. Zoomed region of P.Herc.Paris. 3 shown.

Algorithm 4 Internal surface segmentation using modified Canny edge detection. Given a volume V, Canny edge detection parameters C, coarse mesh M, and option r to invert surface normal vectors, returns a dense point cloud P with points defining the segmented surface.

procedure INTERNAL-CANNY-SEGMENT(V, C, M)
  $P = \emptyset$ ▷ init. empty point cloud
  for $z_i \in [0, V.\text{slices} - 1]$ **do**
    $I_i = V[z_i]$ ▷ extract slice image
    $E_i = \text{CANNY}(I_i, C)$ ▷ detect edges
    $N_i = \text{MESH-INTERSECTION}(M, z_i)$ ▷ get mesh intersection pixels
    for $\vec{o} \in N_i$ **do** ▷ for each intersection pixel
      $\vec{n} = \text{MESH-SURFACE-NORMAL}(M, \vec{o})$ ▷ get surface normal vector
      if r **then**
        $\vec{n} = -\vec{n}$ ▷ invert normal vector
      $\vec{s} = \vec{o}$ ▷ init. sample point at ray origin
      while $\vec{s}.x < I_i.cols$ and $\vec{s}.y < I_i.rows$ **do** ▷ while inside image
        if $\text{EDGE}(E_i[\vec{s}.y, \vec{s}.x])$ **then**
          $P = P \cup (\vec{s}.x, \vec{s}.y, z_i)$ ▷ add point
          break
        $\vec{s} = \vec{s} + \vec{n}/2$ ▷ follow normal vector
  return P

The result of running this segmentation twice, once with regular surface normal vectors and once with inverted surface normal vectors, is two precise segmentations, one on either side of the air gap (Figure 5.5). Viewing these final meshes alongside the initial manual segmentation shows that the fine-grained meshes capture the detailed fiber structure of the papyrus surfaces (Figure 5.6).

5.2.2 P.Herc.Paris. 3

Continuing the example used to visualize the segmentation method, the meshes generated for P.Herc.Paris. 3 were flattened and textured. At 65 cm^2 each, these two segmentations are by far the largest papyrus sheets yet recovered noninvasively from within a Herculaneum scroll. The resulting texture images are in Figures 5.7 and 5.8.

In both images, large regions of uninterrupted papyrus are visible. Some "islands" ringed in black indicate imperfect areas where the segmentations deviate into the air gap and then into adjacent papyrus layers, but in general the grid-like pattern of

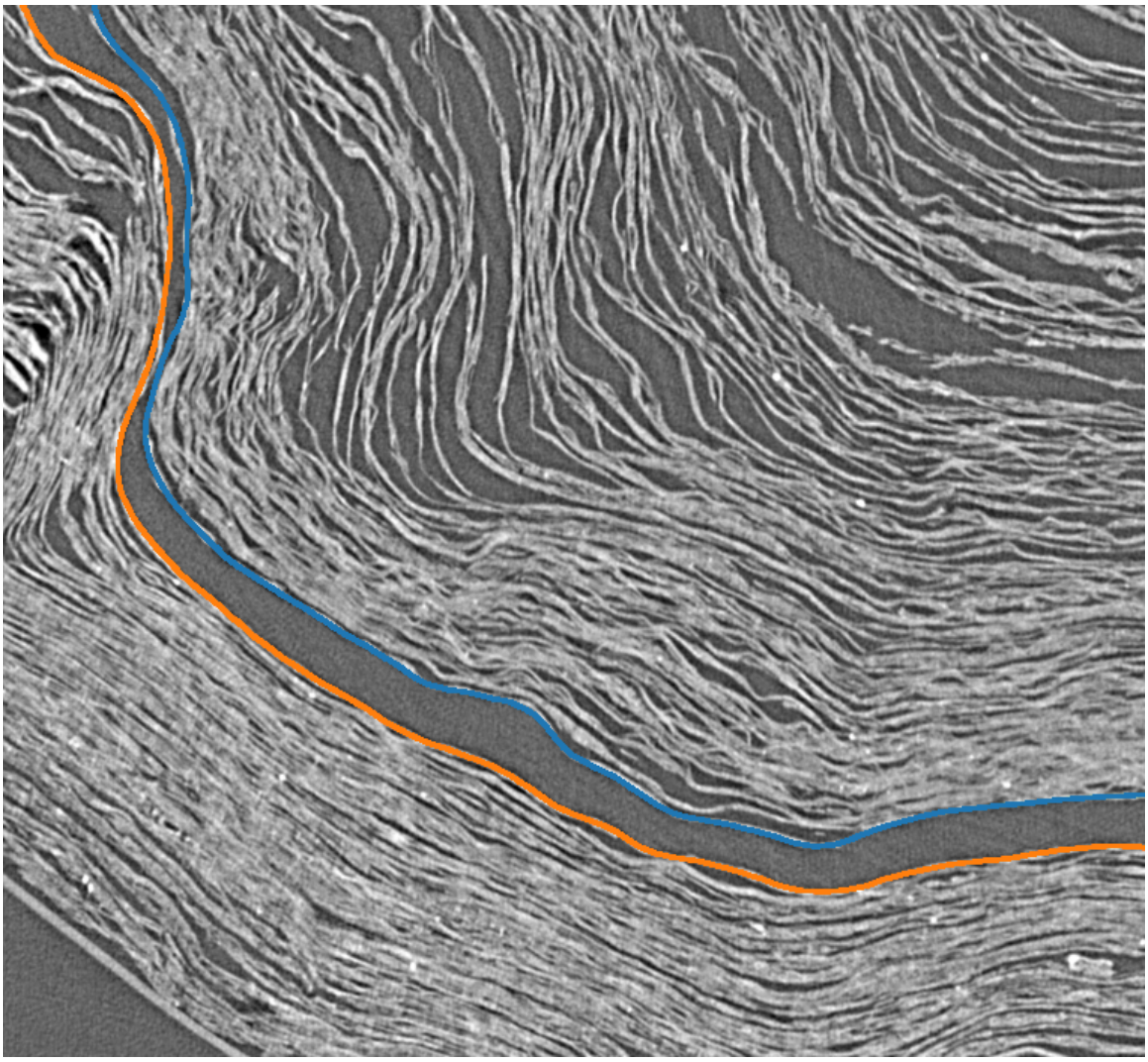
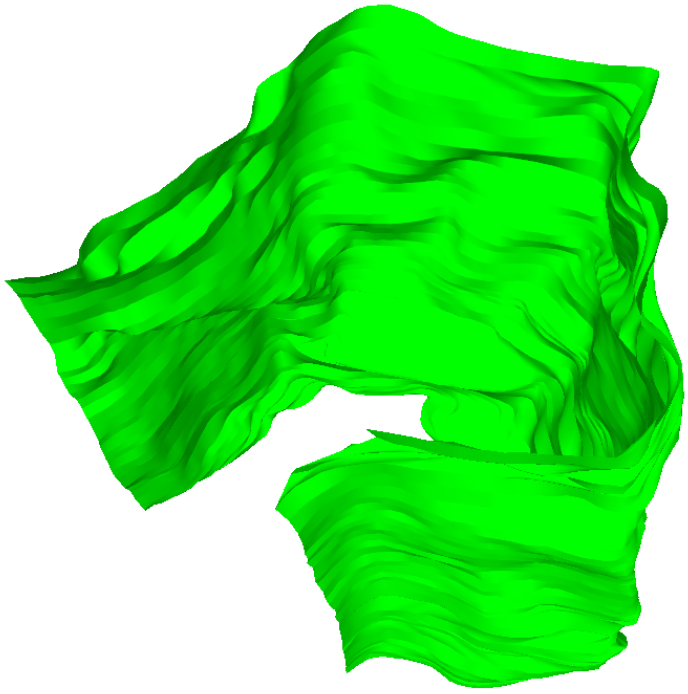
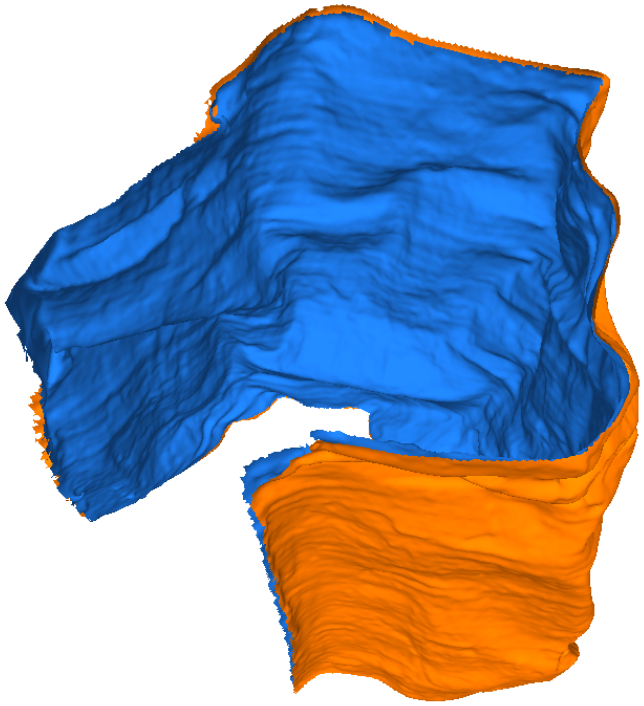


Figure 5.5: Results of Algorithm 4 on zoomed region of P.Herc.Paris. 3. Two segmented surfaces are recovered: internal face (recto) of outer layer (orange) and external face (verso) of inner layer (blue).



(a) Coarse mesh.



(b) Fine-grained meshes.

Figure 5.6: Initial coarse mesh from manual segmentation alongside fine-grained meshes after Canny step. Detailed structure of papyrus fiber surfaces is evident in fine meshes.

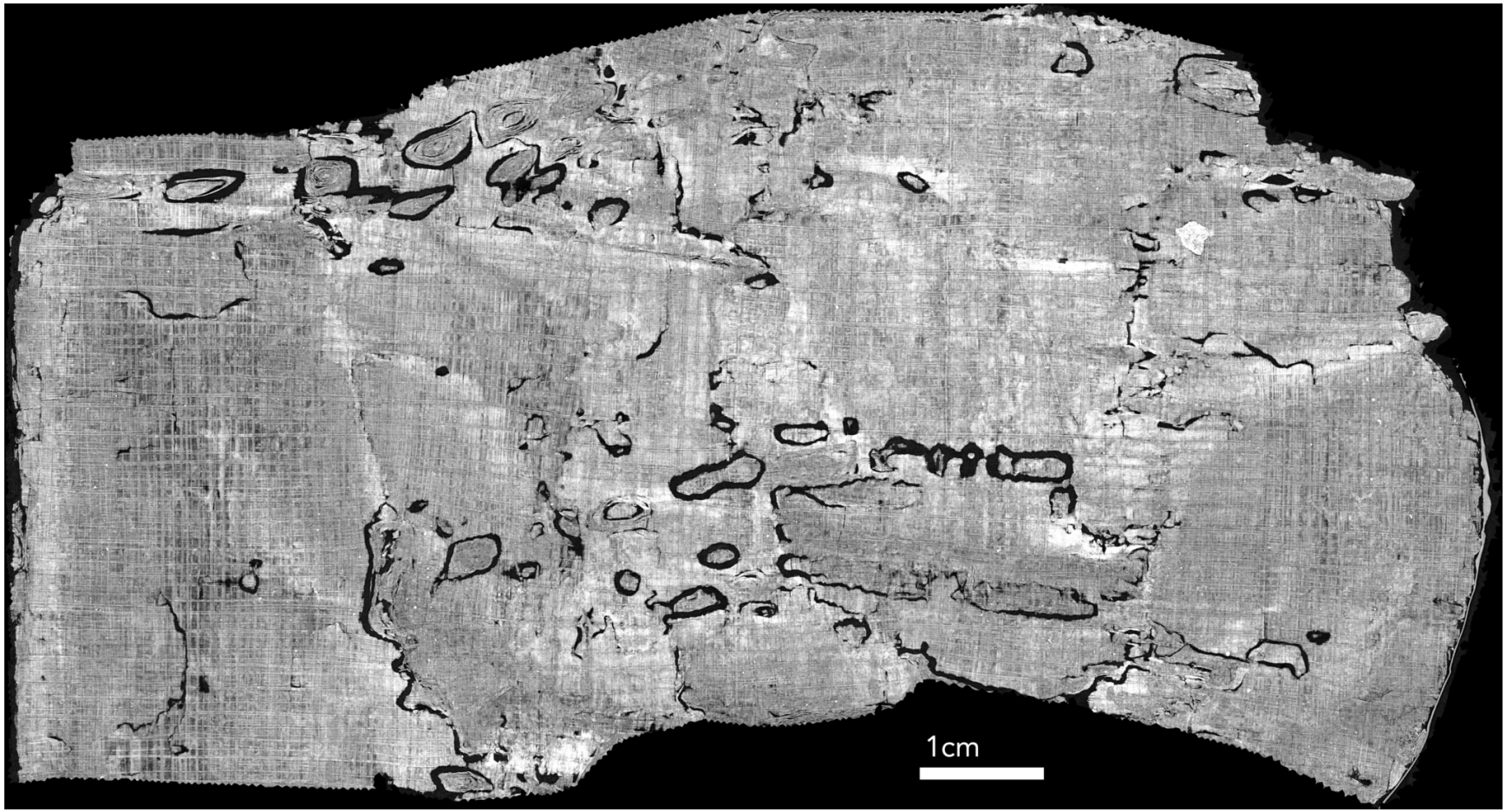


Figure 5.7: Recto segmentation of one wrap of P.Herc.Paris. 3. Large regions of uninterrupted papyrus visible. "Islands" ringed in black indicate imperfect areas where the segmentations passes through air and into adjacent papyrus layers. Papyrological expectations from geometry and location suggest likely writing on surface, in this orientation.

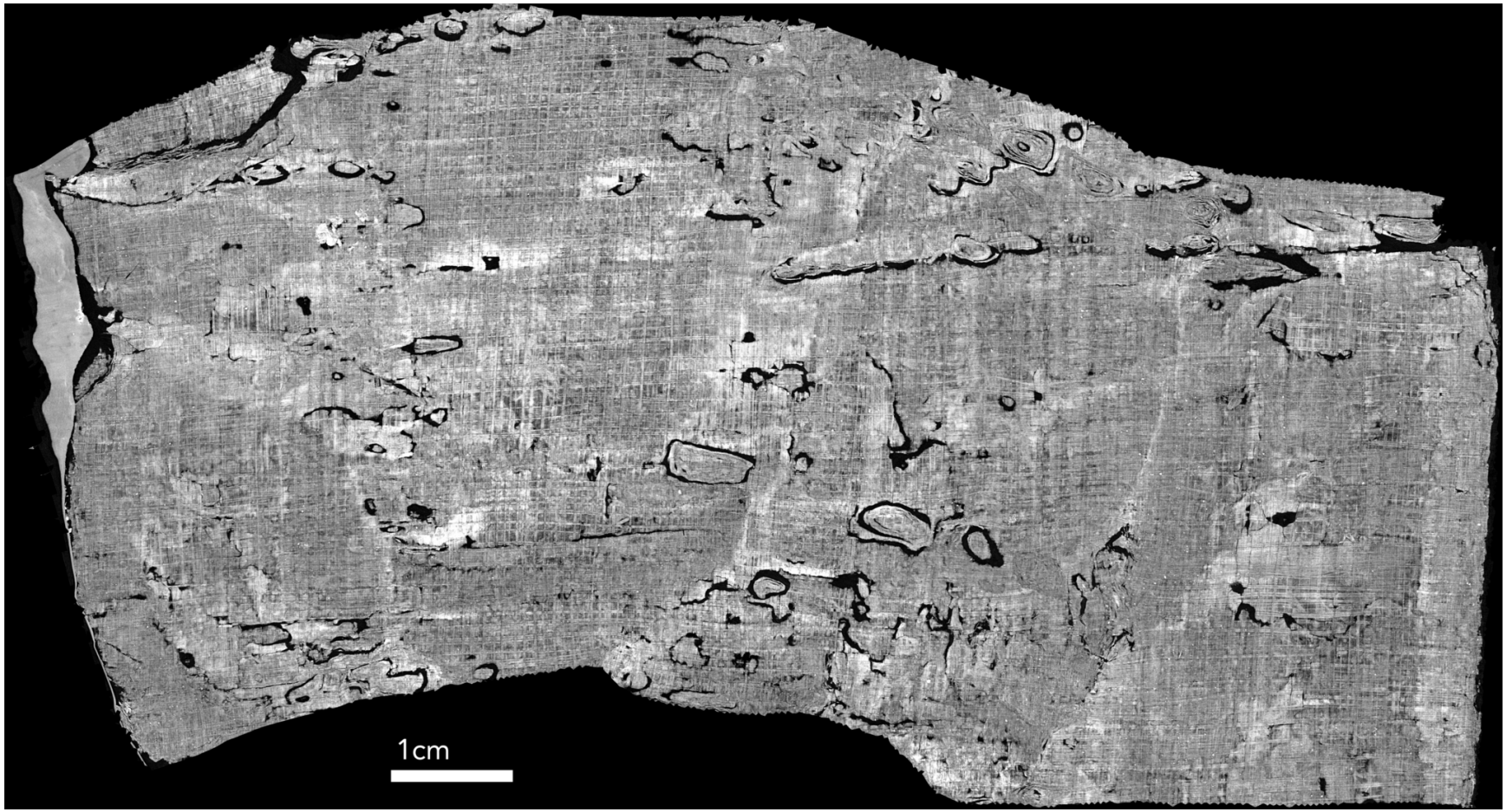


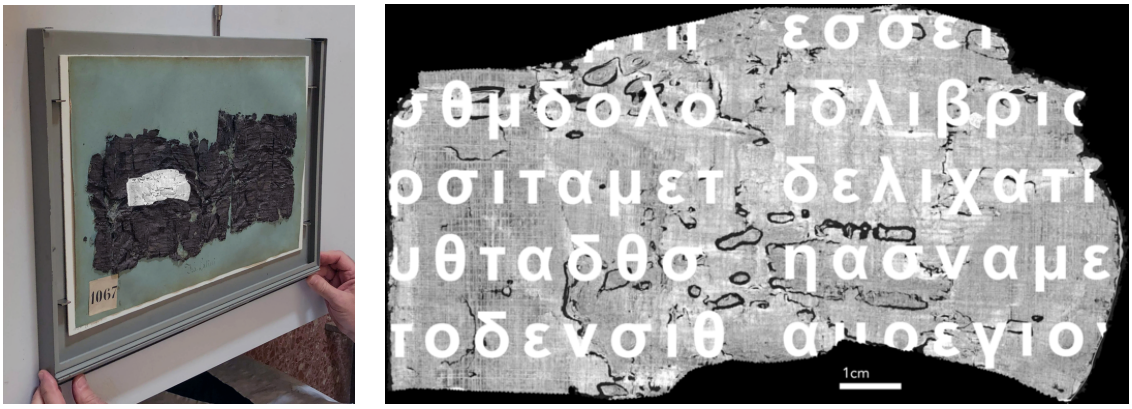
Figure 5.8: Verso segmentation of one wrap of P.Herc.Paris. 3. Large regions of uninterrupted papyrus visible. "Islands" ringed in black indicate imperfect areas where the segmentations passes through air and into adjacent papyrus layers. Papyrological expectations from geometry and location suggest writing unlikely on this surface.

papyrus fibers is evident throughout. Despite the physical deformation applied to the scrolls during their burial (visible in the slice images, or the segmented meshes in Figure 5.6), their internal layers should be in better condition than those of the fragments. Papyrologists use vocabulary inspired by archaeology to describe the pieces of overlying and underlying layers (*sovrapposti* and *sottoposti*) that are commonly encountered on the landscape of a fragment surface, all introduced by the breakages from physical unrolling. The large regions of regular papyrus visible in these texture images seem to confirm the hopes that these breakages are much less common on the better preserved layers of the internal scrolls.

It is expected with Herculanuem papyri that the text is written on the side of the papyrus facing the inside of the scroll, which would suggest the "recto" segmentation is the side where writing would be expected. Other signs also suggest there is likely writing present:

- There are little to no blank or nearly blank scrolls yet encountered in the collection.
- The width of the segmentation far exceeds the expected gap between text columns.
- The height of the segmentation exceeds the expected top and bottom margins of the writing on the scroll. Additionally, the segmentation comes from the middle of the scroll vertically, not the very top or bottom.
- The segmentation does not come from the very beginning (outermost portion) or end (innermost portion) of the scroll, where it is more likely to encounter blank regions.

The segmentation is also shown in the orientation that is expected to be aligned with the text present on the surface. This is deduced based on the expected direction in



(a) (b)

Figure 5.9: Example visualizations illustrating scale of segmented region in Figure 5.7. (a) Shown to scale on typical tray with physically unrolled papyrus. (b) Greek *lorem ipsum* shown at expected scale.

which the scroll is wound. Figure 5.9 visualizes the expected scale of this segmentation relative to the opened scroll trays in the collection, as well as the expected scale of the text on the surface relative to the segmentation.

The "verso" face is not the reverse of the same papyrus region, but is instead from another wrap of the scroll. Text is not expected here, though if it were present, the texture image shows what would likely be the correct orientation.

5.2.3 Fragments

Though they have exposed text on the surface layer, the Herculaneum fragments used in this work are all in reality made up of multiple stacked layers. When broken off from the main scroll body, the fragments typically come off in what can be described as chunks rather than cleanly separated individual layers of papyrus. As a result, the fragments themselves contain some hidden text that has yet to be discovered. The segmentation method developed for the intact scroll layers also works well with some of the scroll fragments.

PHercParis2Fr47 is an ideal candidate for the pursuit of a hidden layer due to the unusually large air gap between layers. The surface layer exhibits convex curvature, while there is another papyrus layer beneath with a concave profile. The reasons

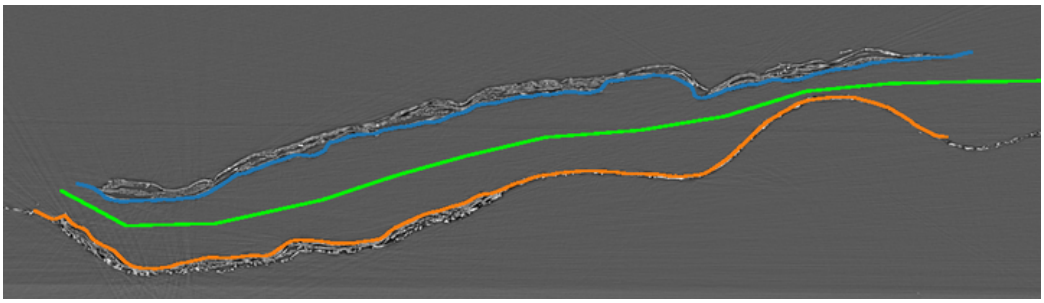
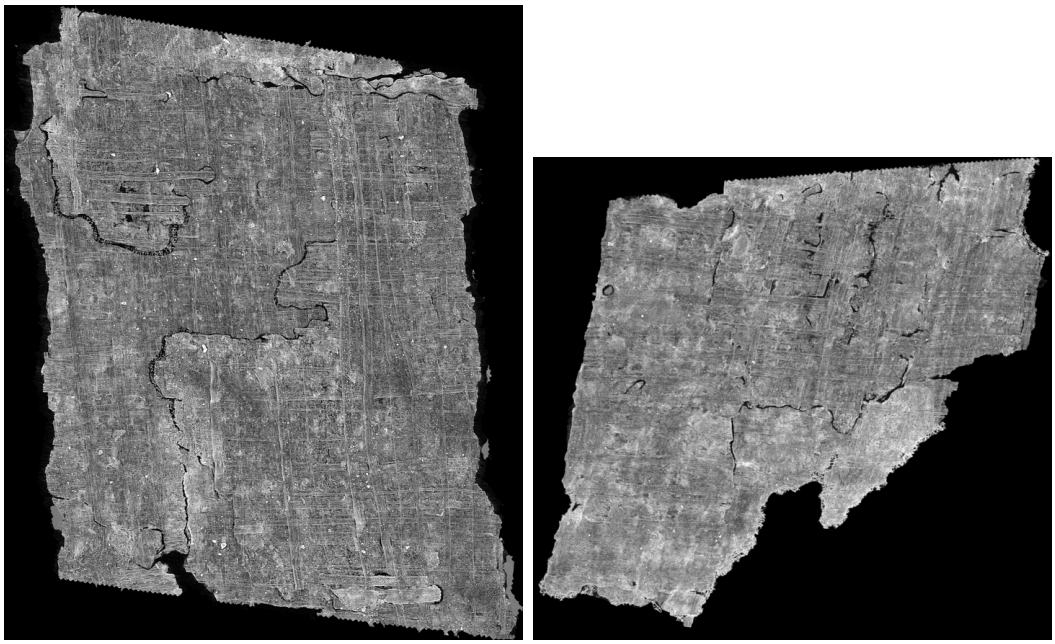


Figure 5.10: Slice from P.Herc.Paris. 2 fr. 47 showing intersections of initial manual mesh (green), recto of hidden layer (orange), and verso of top layer (blue).



(a) Verso, top layer. (b) Recto, hidden layer.

Figure 5.11: Texture images from within P.Herc.Paris. 2 fr. 47.

for this shape are unknown, and are likely some combination of the scroll's original shape, its damage, and the conservation process. Whatever their cause, they create a hidden layer that is easily segmented.

Starting with an initial manual segmentation through the air gap, the recto of the hidden layer and the verso of the top layer were each segmented (Figure 5.10). Their texture images are shown in Figure 5.11.

Though the other fragments do not feature such prominent gaps between layers, there is also occasional separation visible. The most promising layer of each fragment

was also segmented, shown in the next section.

5.3 Applying ink detection

After the segmentation, surface volumes are generated as with the exposed surface layers. A model can now be trained on the surface layer, or on the surface layers of multiple fragments, and inference can be performed on these hidden layers for which there is no ground truth.

5.3.1 First contact

Figure 5.12 shows the first text characters ever revealed noninvasively from Herculaneum papyri. ink-ID was trained on the four fragment surfaces, and then used to generate a binary ink prediction image for the hidden layer of P.Herc.Paris. 2 fr. 47. Though the language is of course Greek, not English, *and* the characters come from different lines, one cannot help but to read it as if it were one word in English: HI! Though small, this first discovery validates an approach long in the making, and promises more to come.

In addition to resembling the script from the fragment surface, these characters pass sanity checks of scale, position, spacing, and orientation. Further, a prediction was also generated for the verso of the top layer, serving as a control as no writing is expected there. Indeed, there is none seen in the prediction image (Figure 5.13).

5.3.2 Other fragments

In the same way, hidden layers from the other fragments were segmented and ink predictions were generated. Figure 5.14 visualizes the ink-ID results on these hidden layers alongside their surface photographs and ink-ID predictions. For each hidden layer, a possible Greek transcription is also provided. Though P.Herc.Paris. 2 fr. 47 has the clearest text, there are signs of text on the other fragments as well. In all cases, where there does seem to be text, the scale, spacing, and orientation align with expectations. There is no indication that ink-ID is generating character-level false positives. The hidden layer predictions are also shown in detail for P.Herc.Paris. 1 fr.

Figure 5.12: The first characters revealed noninvasively from Herculaneum papyri. A greek Eta (H) and Iota (I) are shown. P.Herc.Paris. 2 fr. 47 surface infrared (left) shown to indicate size and vertical position of hidden layer (right). ink-ID trained on top surfaces of all four fragments. ink-ID binary classification output overlaid on texture image. Part of subsurface layer is visible in infrared (red outline).

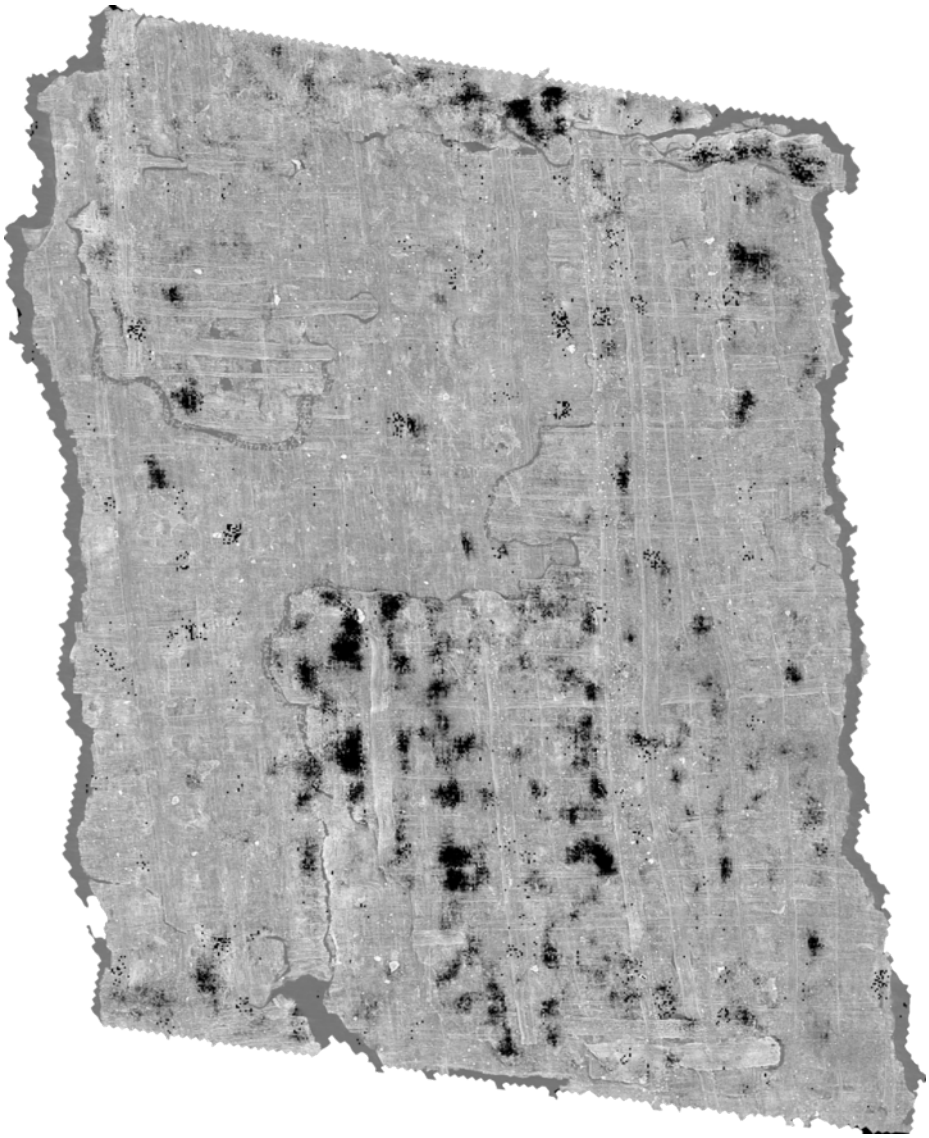


Figure 5.13: ink-ID prediction overlaid on verso of P.Herc.Paris. 2 fr. 47 top layer. No legible characters recovered, as expected.

39 (Figure 5.15) and P.Herc.Paris. 2 fr. 143 (Figure 5.16).

ink-ID likely performs best on the hidden layer of P.Herc.Paris. 2 fr. 47 because of the large air gap that allowed a clean segmentation. Additionally, the subvolumes from that segmentation, having air in their top halves, highly resemble those subvolumes sampled from the training distribution. Where the physical gap between layers is narrower, as with the other fragments, the top of the subvolume can contain not only air, but also some of the underside of the papyrus layer above. This does not resemble the subvolumes from the training set, so creates a distribution shift that may confuse the ink detection model. Some characters do appear despite this challenge, though it is likely they can be improved with better domain adaptation. This is a significantly larger challenge with the intact scrolls, discussed in sections to follow.

5.3.3 Reversing damage

To imagine what one of these surfaces may have looked like 2,000 years ago, an RGB ink-ID model was trained on the fragment surfaces, and used to generate a prediction image for the hidden layer of P.Herc.Paris. 2 fr. 47 (Figure 5.17). The resulting image shows a possible view of this surface before it was damaged during the eruption.

5.3.4 Intact scrolls

Naturally, the next step is to run ink-ID on the segmentations from the intact scroll to see if any text appears. The CT voxel size of the training (3.24 µm) and inference (7.91 µm) images differ, and further it is unknown whether 7.91 µm is sufficient for ink detection. That said, as seen with MS M.910, spatial sampling can enable models trained with one CT resolution to make accurate predictions on another. So there's a chance!

In short, spatial sampling alone is either insufficient to bridge this domain gap, or the CT resolution of P.Herc.Paris. 3 is not quite sufficient to capture the ink signal. An initial experiment used spatial sampling to ensure subvolumes from training and

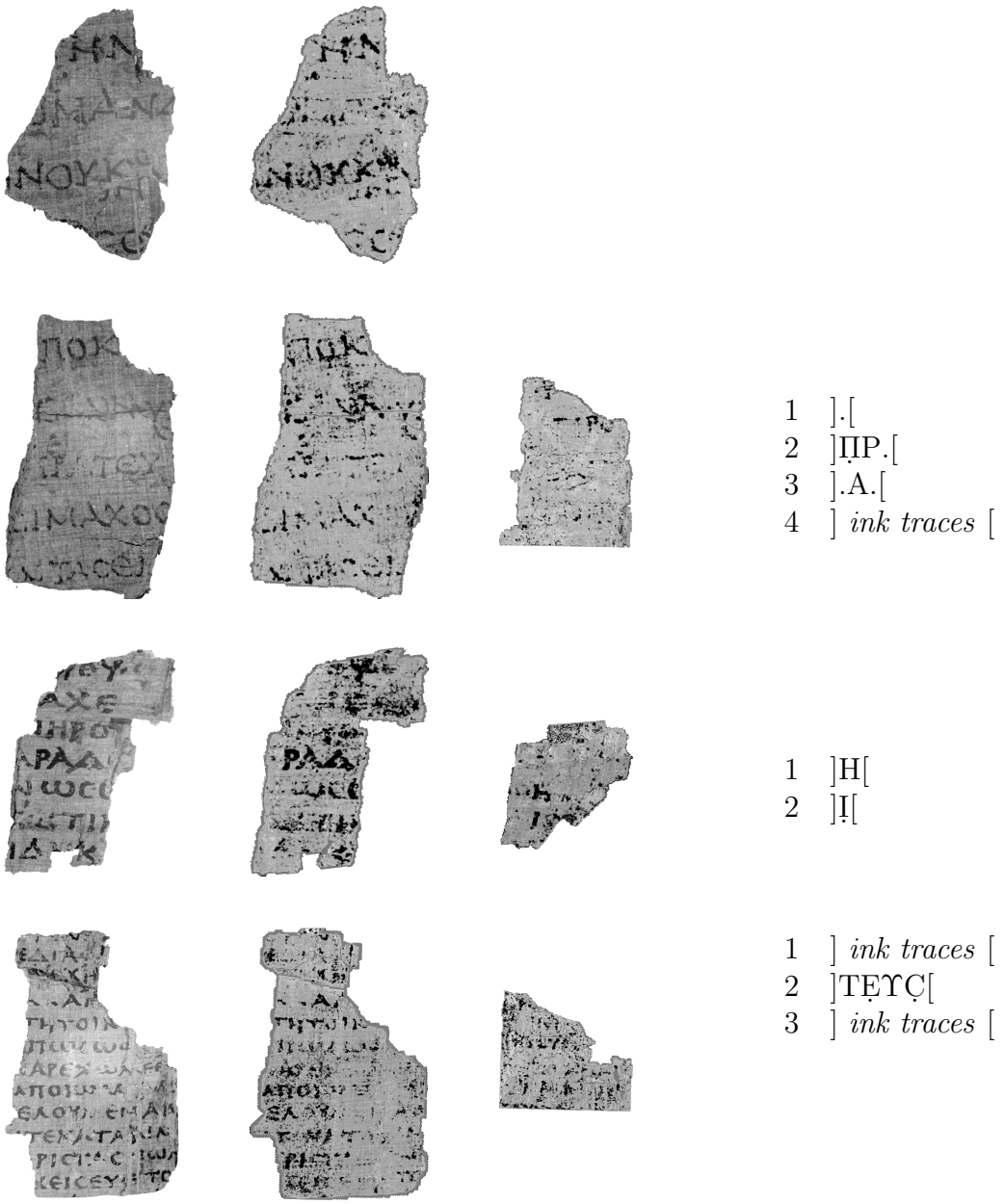


Figure 5.14: Ink detection results for Diamond fragments. (a) Ground truth infrared photographs of fragment surfaces. (b) ink-ID predictions on fragment surfaces. (c) ink-ID predictions on subsurface hidden layers, revealing text that has not been seen in nearly 2,000 years. (d) Possible Greek transcriptions of (c).] and [indicate line beginning and end. Dot indicates indistinct ink traces, underdot indicates uncertain transcription.

Figure 5.15: Larger view of hidden layer prediction image for P.Herc.Paris. 1 fr. 39.

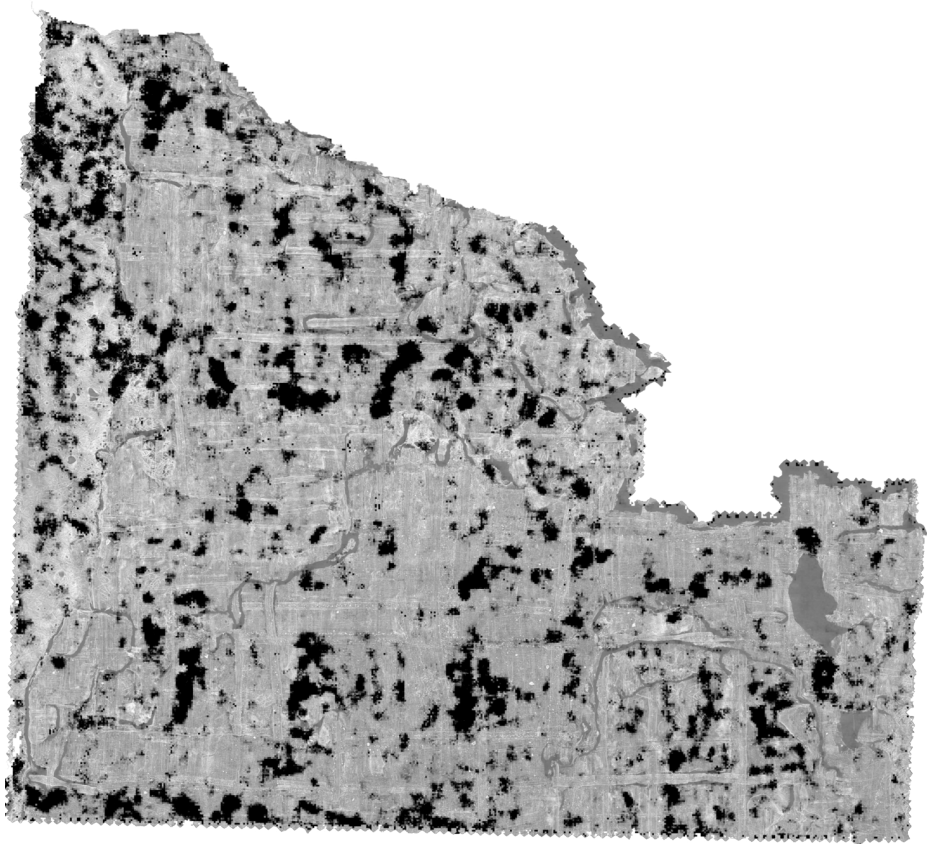


Figure 5.16: Larger view of hidden layer prediction image for P.Herc.Paris. 2 fr. 143.

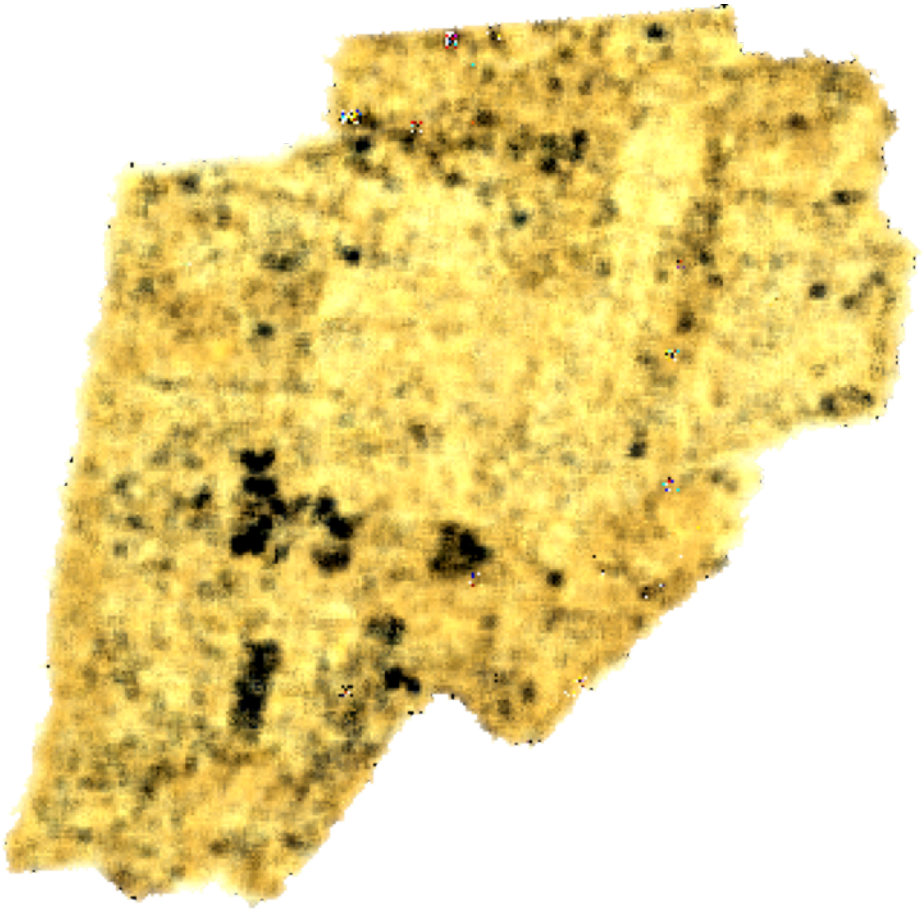


Figure 5.17: ink-ID color prediction image of the hidden layer of P.Herc.Paris. 2 fr. 47, showing what the surface may have looked like 2,000 years ago before its damage.

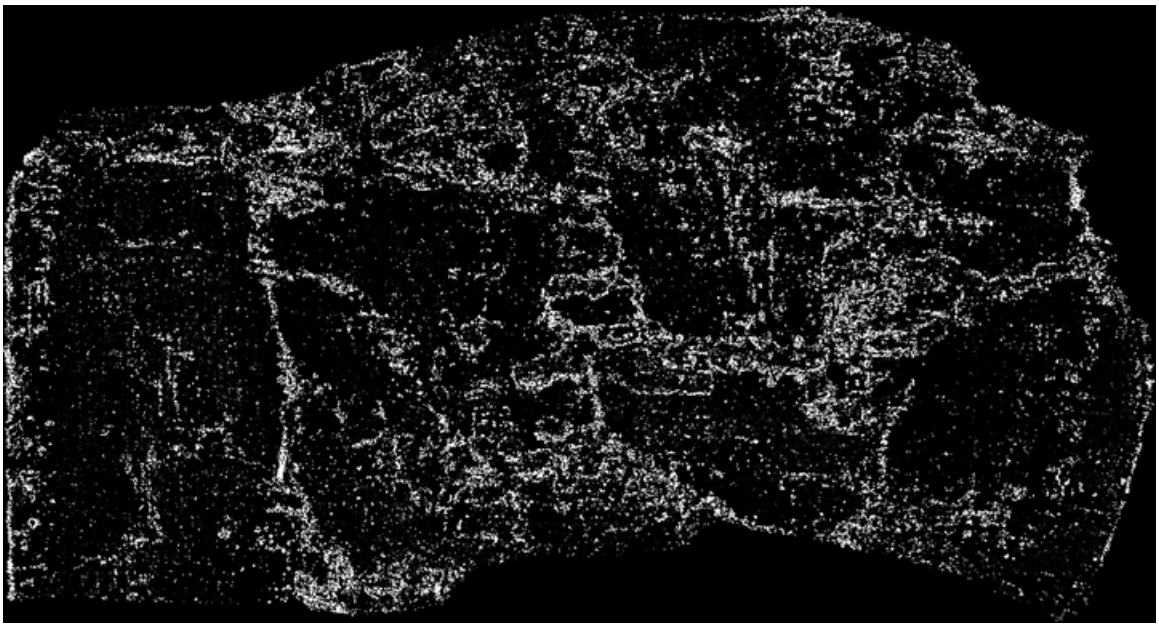


Figure 5.18: Naive application of ink-ID to internal surface of P.Herc.Paris. 3. Trained on fragment surfaces. Spatial sampling used: all subvolumes in training and inference correspond to $77.8 \times 259.2 \times 259.2$ µm. No text evident.

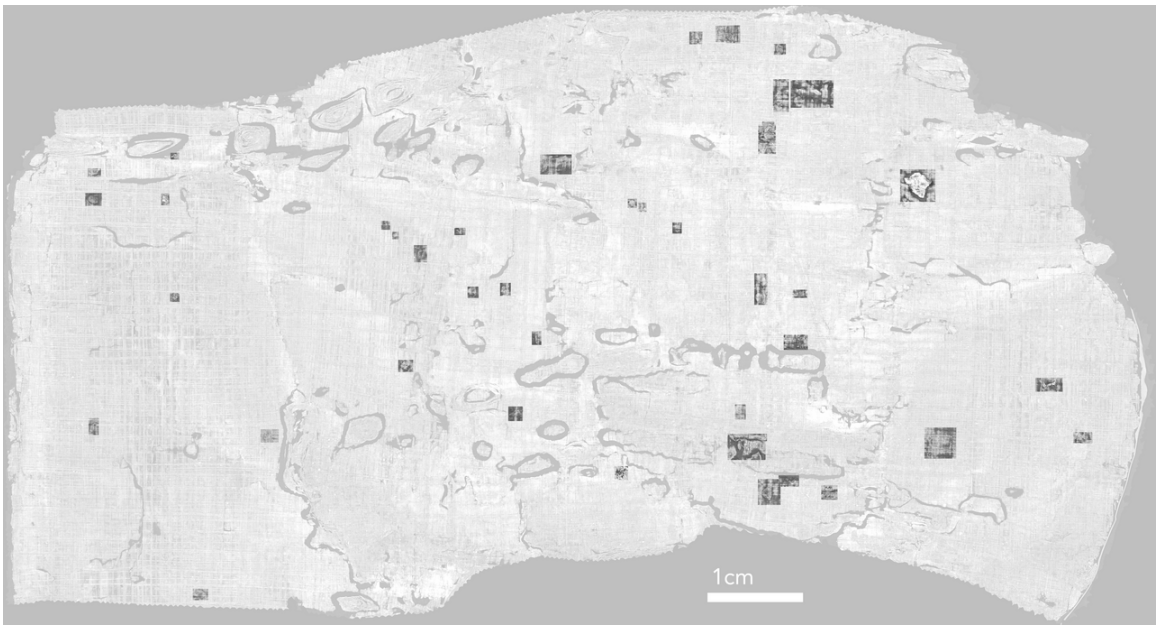
inference had the same physical dimensions, but the resulting prediction does not reveal any patterns of interest (Figure 5.18).

5.3.5 Direct inspection

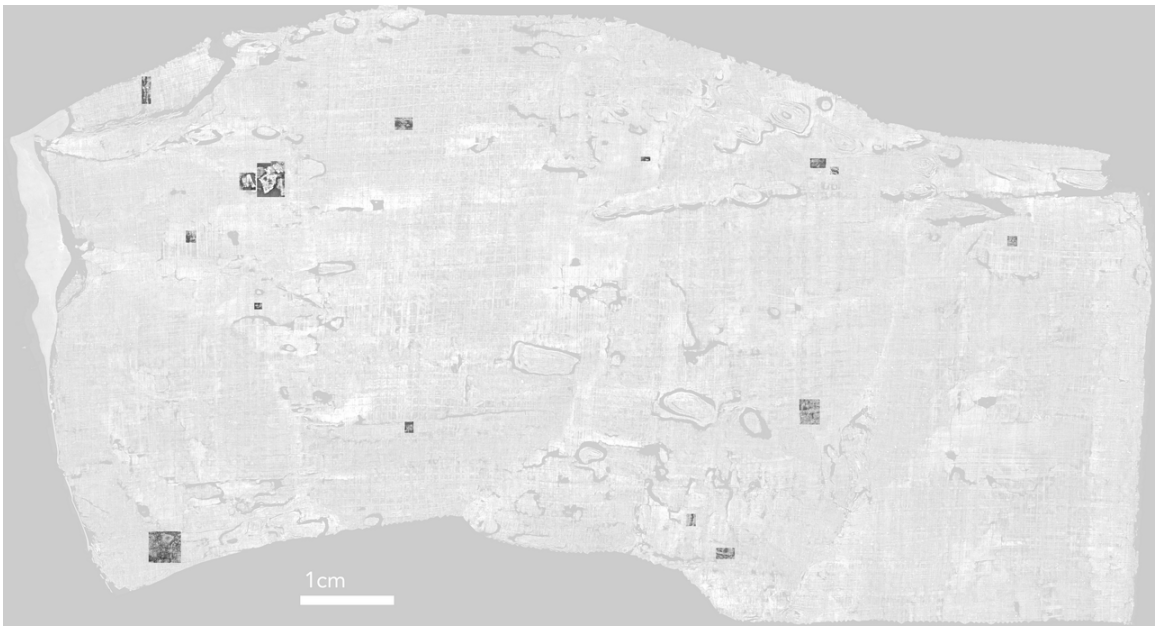
As the ink signal was in rare cases visible in CT on the fragment surfaces, the surface volumes for the large segmentations from P.Herc.Paris. 3 were inspected manually. This data exploration was undertaken seeking a visual intuition for the internal papyrus surfaces of the rolled scrolls. Hopefully, this understanding can help guide the technical steps necessary to overcome the distribution difference. Spots of interest could also indicate where to look first for text in ink-ID prediction images.

Figure 5.19 shows the identified areas of interest on the recto and verso surfaces from within the scroll. Not all identified areas necessarily resemble ink, though some do. Others are interesting features of various kinds. There are more spots of interest on the expected recto than the verso, which is promising.

From the recto, some example regions of various kinds are visualized here. Figure



(a) Recto.



(b) Verso.

Figure 5.19: Manually identified areas of interest on internal wrap from P.Herc.Paris. 3, overlaid on texture images to show their position. Individual areas of interest may come from different depths within the surface volume. Not all identified areas necessarily resemble ink, but may otherwise appear interesting.

5.20 shows the findings that most strongly suggest possible writing. These findings resemble multiple ink strokes, and in at least one case seem to resemble a complete character. Figure 5.21 shows other regions that resemble the end of individual ink strokes. The appearance, scale, and orientation of these strokes are all consistent with expectations. Figure 5.22 shows some other regions of interest, that based on scale, orientation, and appearance, are less likely to be ink. The verso segmentation also contains a number of spots of interest like this that do not quite resemble ink due to the above criteria.

5.4 Domain shift

The absence of revealed text on the hidden wraps from the intact scroll suggests either that the ink signal is not captured in the data, or that the trained model is unable to handle the shift between training and inference distributions. The former is out of scope for this work, which aims to maximize what can be extracted from existing acquired imagery. Further, the occasional visual signs of ink inside the rolled scroll suggest that at least *some* ink signal should be present. This section therefore investigates the distribution shift, attempting to better understand and manage it.

As usual, we begin with visual inspection. One way to observe this shift is to compare the texture images from the fragment surfaces against those from the internal wrap of the intact scroll. The detail images showing spots of interest on the fragment surfaces (Section 3.4) and internal wraps (Section 5.3.5) are one way to do this, but the focus on individual features makes it difficult to gain a broader view.

Figure 5.23 directly compares two 1 cm $\times$ 1 cm squares from the segmented surfaces of P.Herc.Paris. 2 fr. 47 and P.Herc.Paris. 3. The differences in this view are more clear. Some initial observations include the following:

- The image from the higher resolution fragment scan captures sharper details, in line with expectations.
- Both images have had their brightness adjusted for optimal viewing, so any

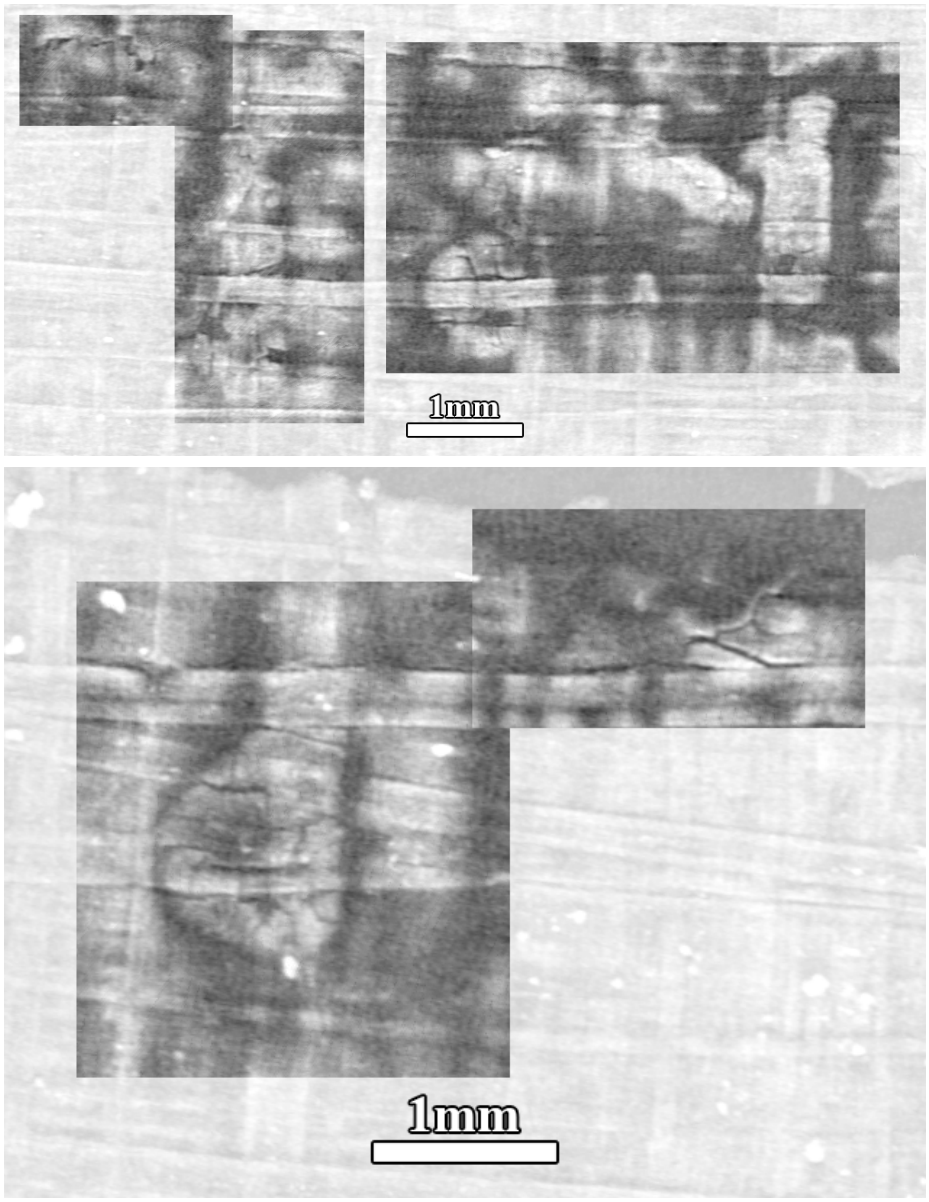


Figure 5.20: P.Herc.Paris. 3 internal layer, recto: the most promising visual examples of regions that could be ink. Scale and orientation consistent with expectations.

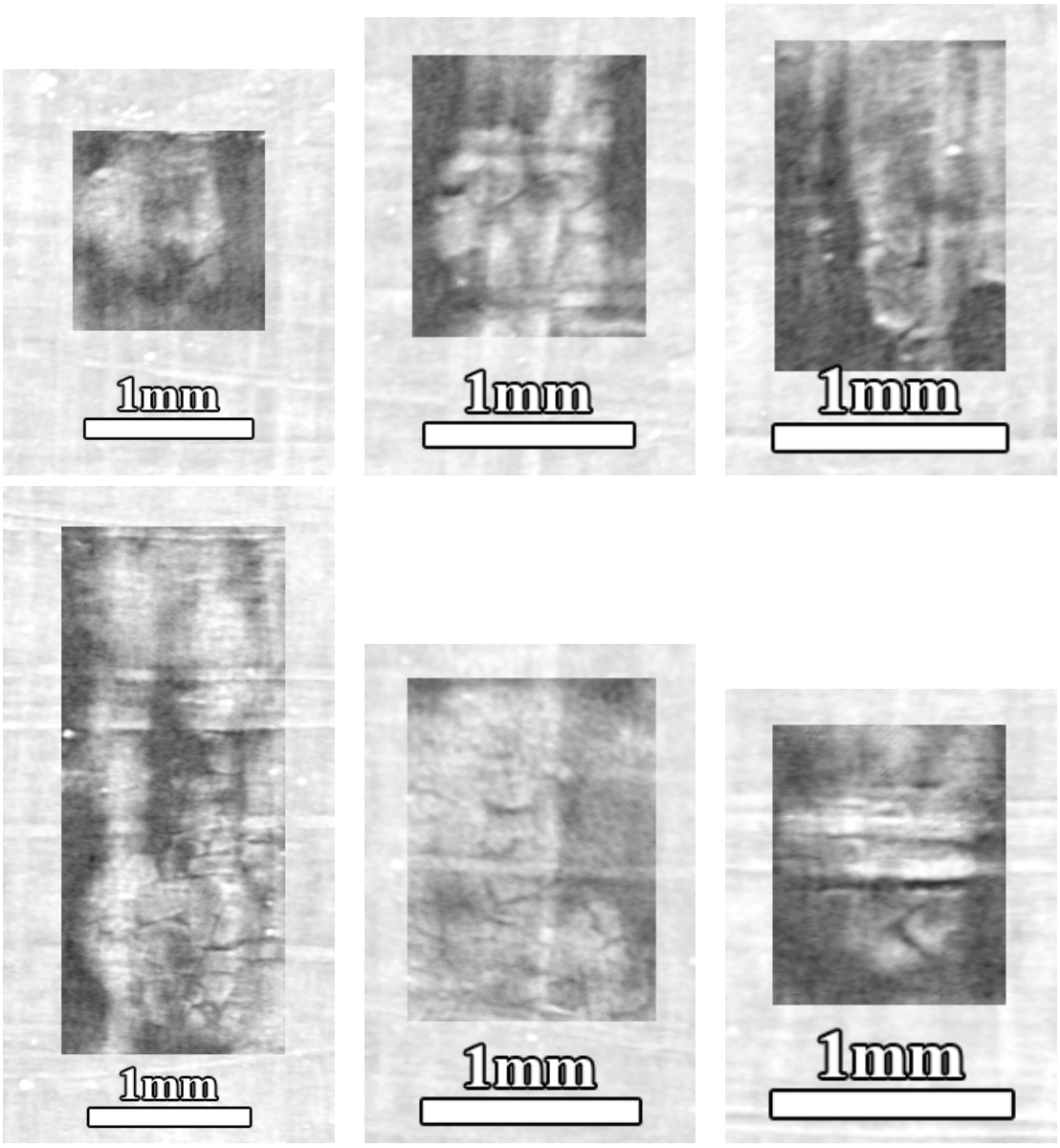


Figure 5.21: P.Herc.Paris. 3 internal layer, recto: other promising visual examples of possible individual ink strokes. Scale and orientation consistent with expectations.

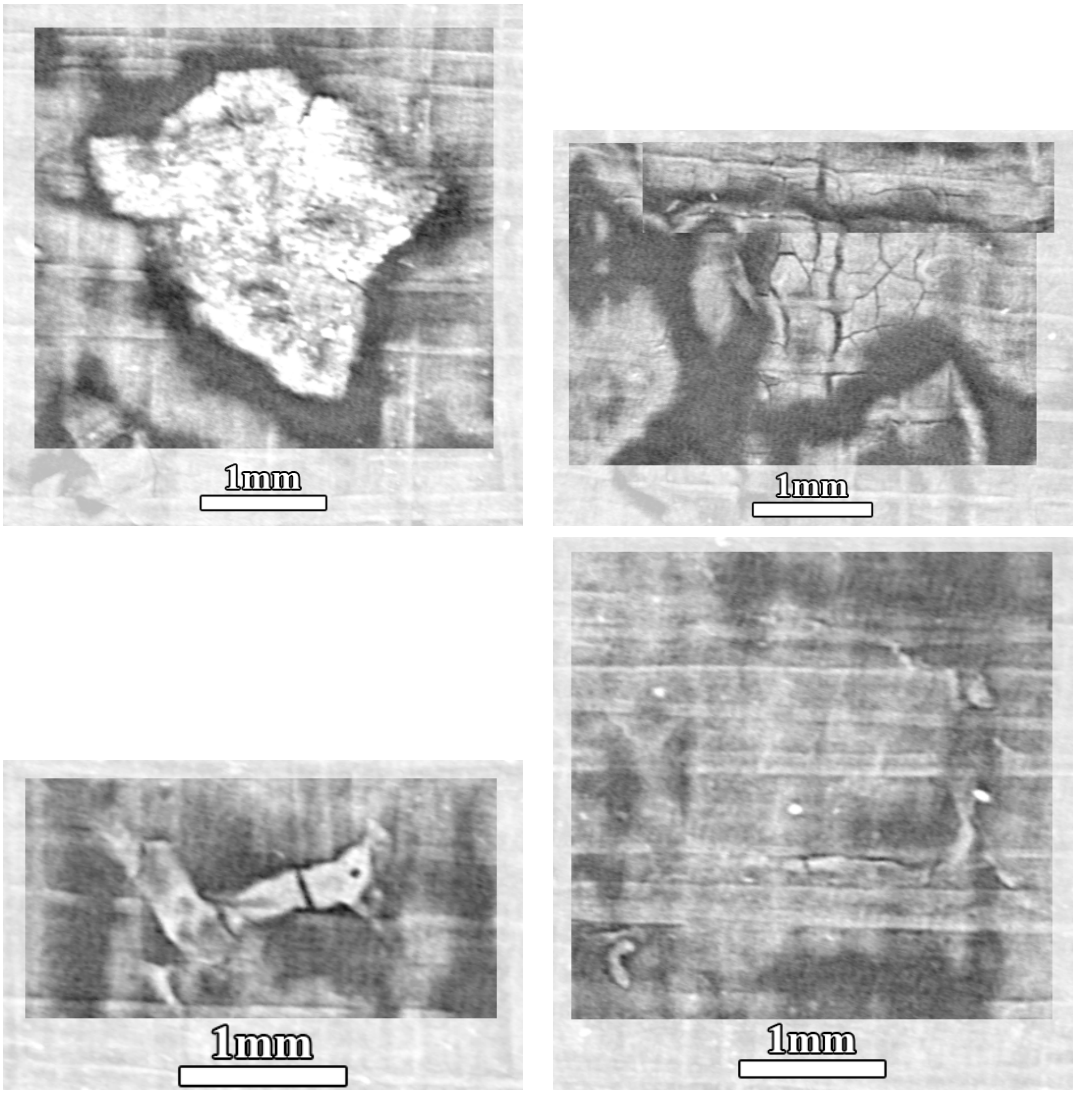


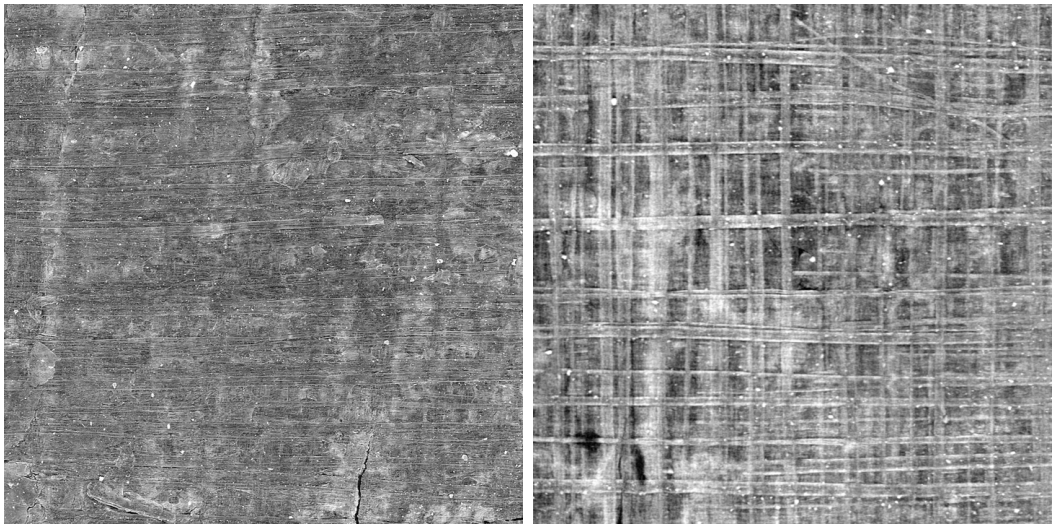
Figure 5.22: P.Herc.Paris. 3 internal layer, recto: example regions of interest that seem less likely to be textual ink based on orientation, scale, or appearance. Verso segmentation has interesting spots similar to these, but very few that resemble ink.

difference in the image intensity *mean* can be ignored for now. But there is also some difference evident in the nature of the image intensity *distribution*. The texture image from the fragment surface captures a flatter gray intensity distribution, while that of the intact scroll segmentation shows more varied intensity.

- Horizontal papyrus fibers are visible in both images, though one has to look closer at the texture image of P.Herc.Paris. 2 fr. 47 to observe them. In both images, they follow roughly similar spacing. *Vertical* fibers, on the other hand, are much more visible in P.Herc.Paris. 3. Papyrological expectations suggest that vertical fibers should correspond to the back (verso) of the writing sheet. The texture images are generated using neighborhoods of the same physical depth, so the baseline expectation would be similar appearances in both texture images. This visual difference suggests either a difference in the physical properties of the respective papyrus sheets imaged, or that there is some blurring with respect to the sheet depth in the CT images of P.Herc.Paris. 3, leading to the visual "bleed-through".

In general, these visual differences suggest that domain adaptation approaches may require more sophistication than spatial sampling and intensity normalization/standardization.

Inspecting subvolumes instead of texture images may help in understanding what exactly is being passed to the models during training and inference. Ultimately these images tell a similar story to the texture images. Figure 5.24 shows representative subvolumes from P.Herc.Paris. 2 fr. 47 and P.Herc.Paris. 3. The oversampling from spatial sampling is evident in P.Herc.Paris. 3, where the larger voxels of the original volume are visible in the subvolume. Particular in the x and y axes, the cross sections also show a distribution difference. The subvolumes are also sampled with the same number of voxels but a larger spatial extent, so the scroll data is sampled "natively"



(a) 1 cm × 1 cm square, P.Herc.Paris. 2 fr. 47 surface (3.24 µm voxel size).

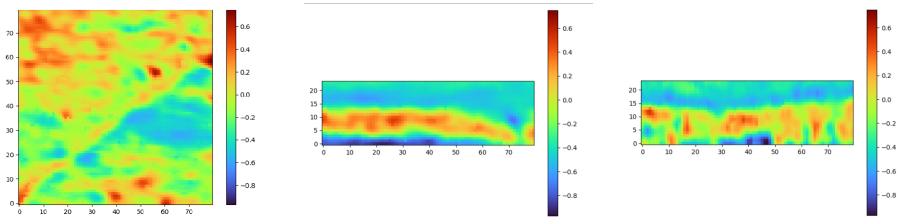
(b) 1 cm × 1 cm square, P.Herc.Paris. 3 surface (7.91 µm voxel size).

Figure 5.23: To-scale visual comparison of papyrus regions from fragment and intact scroll scans, using texture images.

and the fragment data is undersampled (Figure 5.25). This sampling should show similar appearances in both subvolumes, but the fragment subvolume still appears to contain higher frequency details. This suggests the effective resolution of the scroll scan is perhaps less than the voxel size of 7.91 µm.

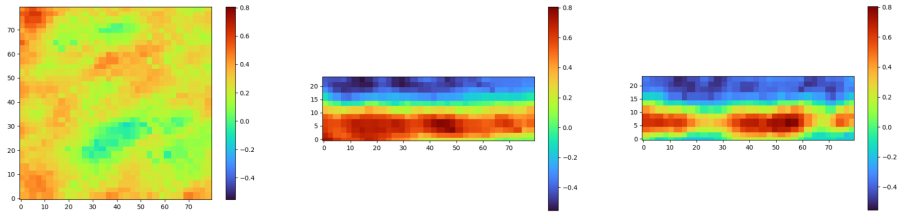
This distribution is also visualized by plotting the histogram of 16 subvolumes sampled from each dataset (Figure 5.26). Differences in the mean intensity as well as the nature of the distribution are visible.

The same subvolume cross sections and histograms (Figure 5.27) are also visualized after the subvolumes are standardized to zero mean and unit variance. The cross sections are not shown, as the automatic windowing for color mapping results in identical visual images even though their intensity scale has changed. The histogram plot demonstrates more clearly that the image intensity distributions are closer, but still vary in shape. In an ink-ID experiment, standardizing subvolumes to zero mean and unit variance during both training and inference is insufficient to bridge the domain shift, as the results to not differ meaningfully from those prior in Figure 5.18.



xy yz xz

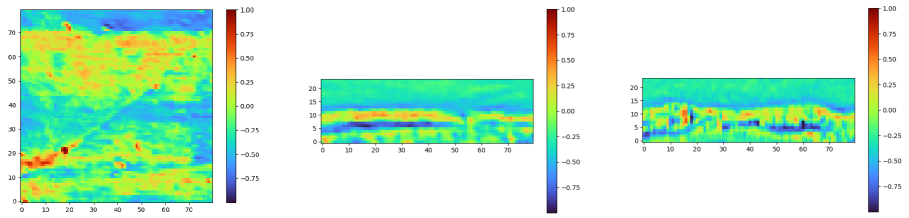
(a) P.Herc.Paris. 2 fr. 47.



xy yz xz

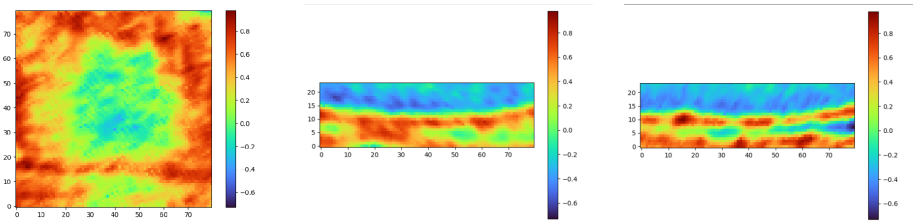
(b) P.Herc.Paris. 3.

Figure 5.24: Visual comparison of raw subvolumes. All are $24 \times 80 \times 80$ voxels and $77.8 \times 259.2 \times 259.2$ µm, "natively" sampling fragment subvolumes in (a) and oversampling scroll subvolumes in (b). Central slice planes shown for each subvolume.



xy yz xz

(a) P.Herc.Paris. 2 fr. 47.



xy yz xz

(b) P.Herc.Paris. 3.

Figure 5.25: Visual comparison of raw subvolumes. All are $24 \times 80 \times 80$ voxels and $189.8 \times 632.8 \times 632.8$ µm, "natively" sampling scroll subvolumes in (b) and undersampling fragment subvolume in (a). Same subvolume origins as Figure 5.24 but with larger spatial extents. Central slice planes shown for each subvolume.

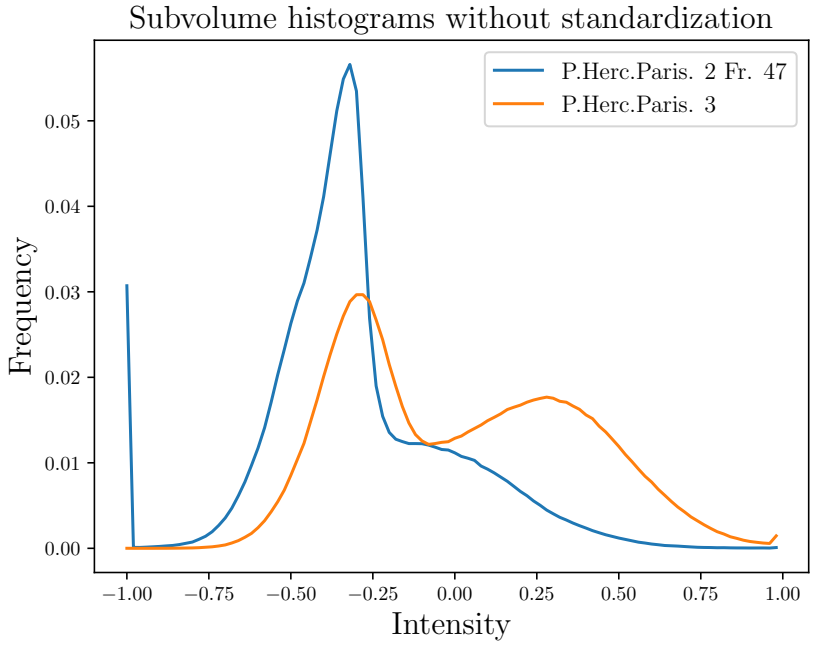


Figure 5.26: Comparing raw subvolume histograms from P.Herc.Paris. 2 fr. 47 and P.Herc.Paris. 3. Histogram across 16 subvolumes from each.

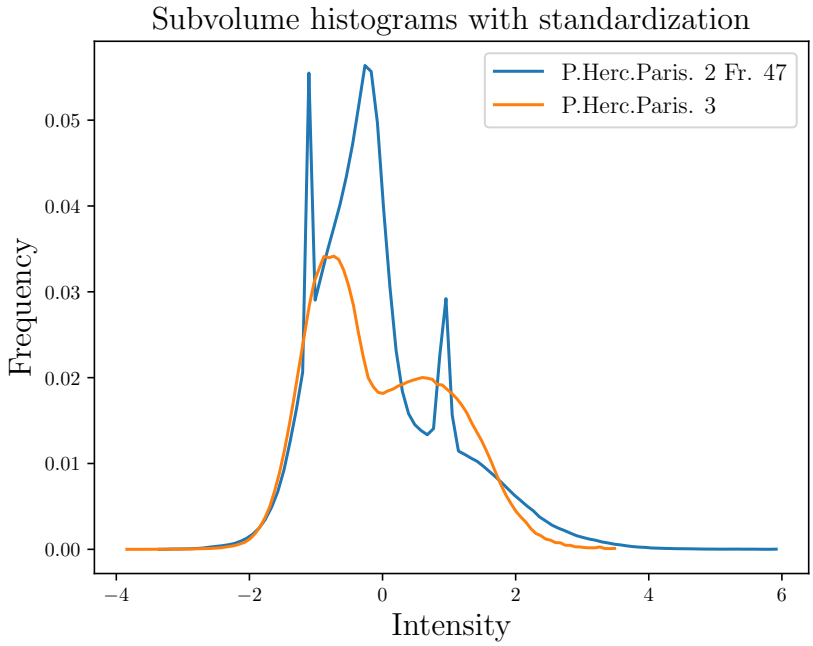


Figure 5.27: Comparing subvolume histograms from P.Herc.Paris. 2 fr. 47 and P.Herc.Paris. 3 after they have been standardized to zero mean and unit variance. Histogram across 16 subvolumes from each.

5.5 Domain transfer

Visual inspection suggests that the distribution shift between fragment and intact scans is more complex than simply a blur or an intensity transformation. Basic attempts to bridge the domain shift using spatial sampling or subvolume standardization have failed to reveal ink on the hidden wraps of the intact scroll, seemingly supporting this hypothesis. Assuming this is true, more advanced domain adaptation methods will be necessary.

Domain *transfer* was chosen as a starting point, in which inputs (subvolumes) from one domain are mapped to resemble the other domain before being passed to the network. Domain transfer is appealing as it only modifies the network inputs: the proven ink-ID model and the training and inference procedures can otherwise be left as-is. Unlike other domain adaptation approaches such as feature alignment, domain transfer also allows visual inspection of the domain adaptation approach, enabling one to build intuition and confidence in how the domain adaptation is behaving.

Related work has used domain adaptation to bridge very similar domain shifts. For example, one work uses a segmentation-enhanced CycleGAN to map volumetric electron microscopy images of neurons to resemble those from another dataset and imaging method [67]. Volumetric segmentation labels are expensive to produce and only existed for the source domain X. Using domain transfer, the authors were able to map volumes from the target domain Y to resemble those from X, enabling a deep learning-based segmentation model to perform well on inputs from both domains.

5.5.1 CycleGANs

The above work used CycleGANs [68], or cycle consistent generative adversarial networks, to perform the domain adaptation. This section describes similar attempts to apply CycleGANs to ink-ID subvolumes, treating those from fragment surfaces as the source domain X and those from the intact scrolls as the target domain Y. As the subvolumes are sampled from different objects, training pairs of $x \sim X$ and

$y \sim Y$ are strictly unpaired. A CycleGAN learns a bidirectional mapping by jointly training two generators: $G : X \to Y$ and $F : Y \to X$. Two discriminators are also trained: D_X which learns to differentiate subvolumes x from translated subvolumes $F(y)$, and D_Y which learns to discriminate y and $G(x)$. Finally, a cycle consistency loss is introduced, encouraging $F(G(x)) \approx x$ and $G(F(y)) \approx y$.

5.5.2 Non-bijection

CycleGANs rely on the assumption that the domain mapping is a bijection, which is possibly violated in this application. As seen in Figure 5.24, target domain Y subvolumes from the lower resolution scroll scans contain less information. $G : X \to Y$ should have no trouble, as it can learn to blur or discretize the more detailed subvolumes from X. But $F : Y \to X$ is being asked to learn a mapping from lower resolution images to higher resolution images: it effectively has to hallucinate details in X that would be sub-voxel in Y. This was deemed an acceptable risk for early CycleGAN experiments for a few reasons.

First, similar domain shifts are present in classic domain adaptation problems that are nonetheless widely used for CycleGANs and related approaches. For example, mapping between horse images and zebra images encounters this issue. Horses typically have more uniform coloring, at least in this toy case, so accurately translating a zebra image to one of a horse can be as simple as painting the coat brown. Mapping from a horse to a zebra is less straightforward, as the generator has to hallucinate stripes where there were none in the source image. This has not prevented the widespread use of this domain transfer task in the literature.

Second, the joint training of bidirectional mappings G and F gives one a choice of which mapping to utilize. Following CycleGAN training, G can be applied to samples from X during ink-ID training, so both training and inference subvolumes resemble those from the target domain Y. Instead, F can also be applied to samples from Y during inference, causing both training and inference subvolumes to resemble those

from the source domain X. If hallucination is feared in one direction, in this case $F : Y \to X$, then G could be used during training to avoid this issue.

Finally, these experiments make the optimistic assumption that the ink signal is ultimately captured in the lower resolution data from Y. Even if sub-voxel details are hallucinated by $G : X \to Y$, it should nonetheless accurately map the lower frequency image features present in Y, hopefully containing sufficient information to capture the ink presence. This optimistic assumption may or may not turn out to be true, but is necessary to enable forward progress.

5.5.3 Initial domain transfer experiments

A number of experiments were conducted to test whether CycleGANs are capable of bridging this domain shift in a way that generates convincing subvolumes. The implementation search space has many variables, some of which are discussed here.

Data generating distribution

The high resolution and unobstructed surfaces of the fragment scans lead to more precise segmentation than is achieved for the internal surfaces from the rolled scrolls. As a result, subvolumes from X more consistently represent the idealized form shown in Figure 2.22a, with the bottom half full of papyrus and the top half full of air. Subvolumes from Y have a more varied appearance. When training a CycleGAN only with subvolumes with precise segmentation from X, the generator $F : Y \to X$ learns only to produce subvolumes with an idealized form, even when the input subvolume $y \sim Y$ has another appearance.

To fix this, X and Y were changed to sample subvolumes from anywhere in their corresponding volumes, rather than only from the segmented surfaces. The CycleGAN thus learned a more general mapping that applied not only to those subvolumes on a papyrus surface, but from anywhere in the volume. This experiment was successful, significantly improving the visual output $x = F(y)$ to more resemble the spatial features of y. This form of subvolume sampling was therefore used in subsequent

CycleGAN experiments.

Subvolume spatial extents

A resolution tradeoff is necessary with the subvolume spatial extents. Experiments with spatial sampling (Section 4.5.2), as well as intuition, suggest the subvolumes during training and inference should represent the same spatial dimensions. As the source and target domains are images of different resolutions, this means either the source domain X has to be undersampled (as in Figure 5.25), or the target domain Y has to be oversampled (as in Figure 5.24). This comes down to whether it is preferable to give the model the "best" data during training or inference. Both were tried, with similar visual results.

Objective functions

The CycleGAN training objective uses image comparison metrics for the cycle-consistency loss, encouraging $F(G(x)) \approx x$ and vice versa. L1 loss (mean absolute error) is typical, and was capable of successfully training CycleGANs with reasonable visual output but without considerable detail.

Other loss functions were tried, including L2 (mean squared error), structural similarity index measure (SSIM) [69], and peak signal to noise ratio (PSNR). SSIM was the only one of these to have a promising effect on the visual output, resulting in generated subvolumes with higher detail. Unfortunately, this came at the expense of the broader image intensity, which is not prioritized by SSIM. Striping was also more evident across the z axis of the generated subvolumes, as the 2D U-net [70] architecture optimized the high-frequency detailed features from each subvolume slice without ensuring the slices were coherent across depth.

Ultimately, L1 loss was preferred for its reliability at producing reasonable output. The detail recovered using SSIM was promising, so perhaps a combined loss function is preferable.

Model architectures

Initial experiments used 2D U-net architectures for the generators and similar architectures for the discriminators. A $24 \times 80 \times 80$ subvolume was treated as a 2D image of size 80×80 with 24 channels. This produced visually plausible output, but striping was often visible in z.

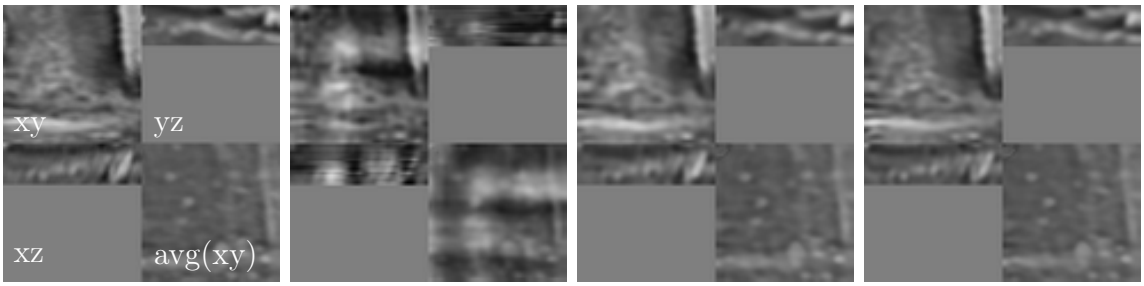
3D convolutional architectures were also implemented for the generators and discriminators. This seems to improve the striping artifact, but did not result in dramatic visual improvements and did sometimes lead to model instability. They were kept, however, as the 3D convolutions are likely a better fit for this 3D spatial data than the 2.5D approach described above.

Balancing adversarial training

The discriminators often acquired low loss values early in CycleGAN training, perhaps suggesting they were consistently outperforming the generators. In an attempt to balance the adversarial training, the discriminators were prevented from updating their weights during training unless their loss exceeded some configurable threshold. The hope was that this would allow the generators to keep up, and the discriminators would only improve periodically, when their performance was bad enough. Multiple thresholds were tested, and the effect on the loss plots for the generators and the discriminators was exactly as hoped, but there was little to no difference in the visual output of the generators.

Contrastive unpaired translation

Contrastive unpaired translation (CUT) [71] was also tried as an alternative to CycleGANs. CUT is unidirectional, so addresses the non-bijective nature of this mapping, and only $G : X \to Y$ must be trained. It also does away with the direct image comparison losses, instead making sure patches from x and $G(x)$ resemble each other in feature space. Finally, CUT is much simpler and faster to train, avoiding the notoriously tricky nature of CycleGAN training due to its 6+ objective functions.



(a) x (b) $G(x)$ (c) $F(G(x))$ (d) $F(x)$

Figure 5.28: Example subvolume passing through a trained CycleGAN. Each subvolume shows three orthogonal slice planes through the center, as well as an average projection through the z axis.

In short, CUT models were also capable of learning visually reasonable mappings, but did not present a notable improvement.

Example outputs

Figure 5.28 visualizes a representative CycleGAN pass given an input subvolume $x \sim X$. The intensity distribution does seem to change in roughly the desired way, and the recovered subvolume after the round trip very much resembles the input ($x \approx F(G(x))$). The identity mapping also performs well, preserving the input ($x \approx F(x)$).

$G(x)$, however, is peculiar. Clearly, some spatial details of the input subvolume x are preserved. Others, though, are hallucinated, and did not exist in x. $F(\cdot)$ has evidently learned to cope with or ignore these hallucinations, as they are no longer visible after mapping back to X. The round trip $F(G(x))$ and cycle-consistency loss are only useful insofar as they encourage a useful $G(\cdot)$, so despite a promising initial appearance, this CycleGAN is visually suspect. Note also some striping evident through z (manifesting in the x and y slices of $G(x)$). This sort of output is characteristic of the CycleGANs trained throughout these experiments.

5.5.4 Applying domain transfer results to ink-ID

Visual inspection of subvolumes was used extensively while testing domain transfer approaches: by comparing X, Y, $G(X)$, $F(Y)$, $F(G(X))$, $G(F(Y))$, $F(X)$, and $G(Y)$, one can gain evaluate to some extent the ability of the generative models to perform

the mapping. This is different from the ultimate objective of leveraging the trained mappings for ink-ID. To test whether the domain transfer models were enabling ink-ID to recover the ink of the intact scrolls, the most promising CycleGAN models were frequently used for ink-ID experiments. This implementation space also has a number of variables, some discussed here.

Mapping direction

It is unclear a priori which mapping is preferable for ink-ID. By using $F(Y)$ during inference, the model gets to train on unmodified subvolumes from X. This way, the model can access "full-resolution" images during training, likely improving its ink detection at the possible expense of domain adaptation. Using $G(X)$ during training likely worsens the quality of the training data, but increases the odds the model will work well during inference with samples from Y. Both were tried, with similarly unimpressive results.

Identity mappings

The CycleGAN implementation tested also uses an identity loss during training, ensuring $F(x) \approx x$ and $G(y) \approx y$. Let us assume the generators G and F learn usable mappings, but do not perfectly translate a subvolume from one domain to the other, perhaps introducing some artifacts in the process. Based on visual results, this is a reasonable assumption.

$G(X)$ would then not exactly resemble Y, but instead a similar domain $\hat{Y}$ and vice versa. ink-ID trained on samples from $G(X)$ would then not perform well on samples from Y. Instead, might it be helpful to train on $G(X)$ and predict on $G(Y)$? This way, both training and inference data would resemble a common distribution $\hat{Y}$.

This approach was attempted in both directions: training ink-ID on $G(X)$ and predicting on $G(Y)$, and separately, training on $F(X)$ and predicting on $F(Y)$. Neither produced promising ink-ID results.

5.5.5 Discussion

Initial experiments for domain transfer on Herculaneum papyri have been inconclusive. Domain transfer models are capable of visually plausible output, though not yet exceptional, and none of the models so far trained have successfully led to text recovery in the intact scrolls using ink-ID.

There are multiple promising directions for next steps. Visual results suggest domain transfer could likely be improved, perhaps by leveraging additional information during CycleGAN training such as ink labels. Instead of domain transfer on the inputs, domain alignment in feature space [72, 73] could be used to train domain-invariant ink-ID models. Though the samples from Y are generally considered unlabeled, Section 5.3.5 shows some places in the intact scrolls that do strongly suggest ink presence. Perhaps a small set of labels generated from these regions could be used to fine-tune an ink-ID model trained on the fragment surfaces.

Exploring this domain adaptation problem is challenging, in large part due to the lack of labels in the target domain. Relatedly, none of the approaches so far seem to show any ink in the intact scroll segmentations, so it is difficult to compare experimental results and determine the best direction for next steps. Once some signal begins to appear, it will be easier to guide development. For now, this line of inquiry can feel at times like taking shots in the dark. This is reminiscent of early ink-ID experiments on the fragment surfaces, which did eventually converge, a vote in favor of optimism.

5.6 Summary

This chapter has presented a novel segmentation method for 3D surfaces in tomography, combining quick user input with an algorithmic refinement step to produce large, precise segmented surfaces. These segmentations were used with ink-ID to recover noninvasively, for the first time ever, textual characters hidden inside the Herculaneum papyri since A.D. 79. These surfaces can also be rendered in full color,

visually reversing the damaging effects of carbonization and showing the surfaces as they may have appeared 2,000 years ago. As the segmentation method also generates a segmentation for the opposing verso sheet, ink-ID was also used to generate predictions for the verso of the top papyrus layer on one fragment, confirming the expected lack of ink presence.

The segmentation method was applied to the interior wraps of P.Herc.Paris. 3, generating by an order of magnitude the largest sheets of papyrus extracted noninvasively from Herculaneum papyri. An exploratory data analysis of these sheets found multiple spots of interest, including possible ink, and the large regions of "clean" regular papyrus are in better condition than the papyrus of the opened fragments, confirming the hopes that the sheets inside the rolled scrolls are better preserved. An initial investigation was also conducted into the domain shift between the CT images of the scroll fragments and the intact scrolls, along with initial experiments for domain transfer methods.

CHAPTER 6. ABLATION STUDIES

6.1 Introduction

The research contributions of this work have moved forward the effort to noninvasively recover the texts of the Herculaneum scrolls. Where ink was once considered invisible, there is now a pathway to recovering it from X-ray CT. This has been used already to reveal previously hidden characters from the subsurface layers of the scroll fragments. There remains a gap, though, between these exciting results and the ultimate objective of recovering complete texts from the rolled Herculaneum scrolls.

What can we expect moving forward? Is this possible? What level of precision is necessary? What is the ceiling of how well these methods can work?

This chapter investigates these questions with respect to some of the individual pipeline steps. Each of these steps will be probed to build a better understanding: what are the limits, and how can we move towards them? By examining the individual pipeline steps in this way, this chapter paints a picture of where this work is headed, where the boundaries are, and what will be possible. The chapter contributes ablation studies of the following areas:

- **Imaging:** experimental results are shared to inform the imaging requirements for successful ink detection, notably with respect to spatial resolution and incident energy.
- **Inputs:** the optimal inputs are characterized experimentally, including optimal subvolume depth and width, required segmentation precision, and required label alignment.
- **Models:** model architectures and training procedures are explored.
- **Generalization:** experiments probe whether these methods can be expected to perform successfully on the existing scans of the intact Herculaneum scrolls.

6.2 Imaging implications

This work has focused on a computational pipeline capable of augmenting what can be recovered from multidimensional images. In the case of the Herculaneum scrolls, this work has used X-ray micro-CT images, showing that even where contrast does not appear to the eye there is often recoverable ink signal using algorithmic methods.

CT has many configurable parameters, both during the acquisition of raw X-ray projections and the CT reconstruction process that creates volumetric images. As the central thesis of this work suggests the models are optimizing for something that is not readily visible in the original CT images, selecting these imaging parameters at the time of acquisition can feel like a shot in the dark. The datasets in this work sample a few configurations of these parameters. This section will discuss what has been learned about the most important CT parameters, notably resolution and incident energy, and will suggest what future imaging should prioritize for maximal ink recovery.

This work has focused on X-ray CT images because they offer, by far, the highest resolution of object interiors relative to other existing images. Though the ink contrast leaves much to be desired, this work shows it can be nonetheless recovered in many circumstances. Ongoing related work in physics and chemistry is working in parallel to develop imaging methods that will enhance the internal ink contrast. All of this work is highly complementary to the computational pipeline presented here. Each step of the pipeline has room for improvement, and that small wins in individual components accumulate to sizable advancements in ink detection. This is likely true most of all for the imaging method, as it establishes the ceiling under which the remainder of the pipeline operates. Existing literature on this topic suggests that a "silver bullet" imaging method, combining high spatial resolution with high ink contrast, is unlikely in the near future. This is no cause for disappointment: small increases in the captured ink signal, even if they are still below the threshold for

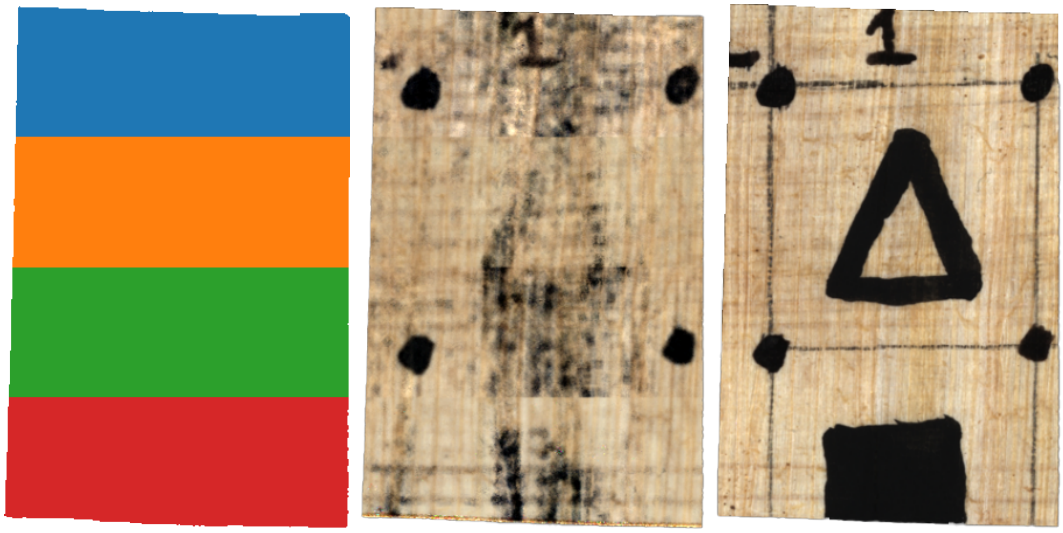
apparent visibility, will have outsize effects on the downstream performance. Whether this comes from phase contrast micro-CT or other imaging methods under study, the possibilities are exciting.

6.2.1 Resolution

In the current paradigm using micro-CT, resolution is very clearly the most important factor in the original images. This is immediately visible in the Carbon Phantom experiments: the thickness of the ink layer, relative to the voxel size of the CT image, determines whether and how well the ink is recoverable. As the ink layer thickness (measured in SEM) slips below the voxel size, the model is no longer capable of detecting the ink, supporting the hypothesis that the model is detecting some textural pattern that must be imaged at sufficient resolution.

The Carbon Phantom, originally imaged at 12 µm voxel size, was later partially imaged on another machine at 8 µm. With the original scan, at best very faint signs of ink were recovered in the first column. This slight resolution improvement considerably improved the ink recovery, shown in Figure 6.1. The Δ visible in the center is faint but legible in the ink-ID prediction image.

Experiments to date with Herculaneum papyri seem to confirm these findings on the real materials. ~ 3 µm is below the expected voxel size threshold, and this seems to be confirmed in the comparable results from the Diamond fragments (3.24 µm) and P.Herc.Paris. Objet 59 (3.37 µm). It is yet unknown if the current highest resolution scans of the intact scrolls, at 7.91 µm, contain enough detail to recover the ink signal. Visual results suggest there are certainly some kinds of inky details captured, at a rate similar to the inky spots observed in the CT images of the fragments, and models are capable of recovering that and much more from the fragment images. Though the domain shift introduces a challenge that has yet to be fully addressed, I predict that these scans will end up containing enough signal to recover meaningful portions of text. The ink signal is highly variable within a scroll and across the collection, so



(a) 4-fold regions. (b) ink-ID predictions. (c) Reference.

Figure 6.1: ink-ID predictions on part of first column of Carbon Phantom when CT scanned at 8 µm voxel size.

these findings will start will small pieces and grow over time.

Simulated effects

Using real data to measure the performance of ink-ID at various imaging resolutions is desirable because the resolution is not simulated; it emerges directly from the physical operation of the scanner. There are confounding variables in that comparison, however, including incident energy among other small changes between scans. Since resolution is such a key factor in the future imaging of these scrolls, this effect was additionally examined via simulation.

Based on the available images to date, this general investigation is, to some extent, dancing around one specific question: is the ink signal captured in the 2019 7.91 µm scans of P.Herc.Paris. 3 and P.Herc.Paris. 4? The domain shift makes this difficult to answer concretely by training ink-ID on the fragments and performing inference on the intact scrolls. There is no ground truth, and it is unclear whether to attribute the lack of revealed text to the image resolution or to the domain shift between training and inference.

Simulated resolution is one way to at least inform this question. Figure 6.2 shows a

direct attempt to evaluate the feasibility of Herculaneum ink detection at 7.91 µm. A typical 8-fold experiment was conducted across the four fragments, this time sampling subvolumes as if they came from a 7.91µm CT volume. Ink detection is marginally worsened as expected, but still performs quite well, clearly recovering text. While not a guarantee of the signal captured in the separate 7.91 µm scans, this is promising.

Further simulated resolution experiments were conducted with the Carbon Phantom. The CT volume was downsampled to various simulated resolutions, and ink-ID was evaluated each time. As before, each of the six columns was treated as a separate 5-fold cross-validation experiment. Figure 6.3 shows the outcome, measured using the area under the receiver operator curve (AUC).

The results show the ink signal is tolerant to more downsampling than anticipated. The leftmost values show performance at the original 12 µm resolution. As column 1 does not recover legible text and column 2 does, the threshold for readability using AUC likely falls between 0.70 and 0.80. Using this measure, columns 3 and above retain readability at simulated resolutions up to 72 µm. As column 3 has an estimated ink thickness of ~ 15 µm, it is pleasantly surprising that the ink signal is still recoverable when the voxel size exceeds that by more than four times. Column 2, with an estimated ink thickness of ~ 10 µm, also retains readability beyond expectations. These results are very promising, in that they suggest ink signal may be recoverable using lower resolution than is currently deemed necessary. They could be interpreted as slightly at odds with the morphological hypothesis, which would have suggested the ink signal would be irrecoverable as the voxel size met and then exceeded the ink layer thickness. More investigations into this area would likely be fruitful.

6.2.2 Incident energy

Second to resolution, incident energy is the largest adjustable parameter during CT image acquisition. Typically, the incident energy is selected deliberately based on the properties of the object or material being scanned. First, the energy must be

(a) 8-fold. (b) ink-ID results.

Figure 6.2: ink-ID results across the four "Diamond fragments" sampling as if they were 7.91 µm.

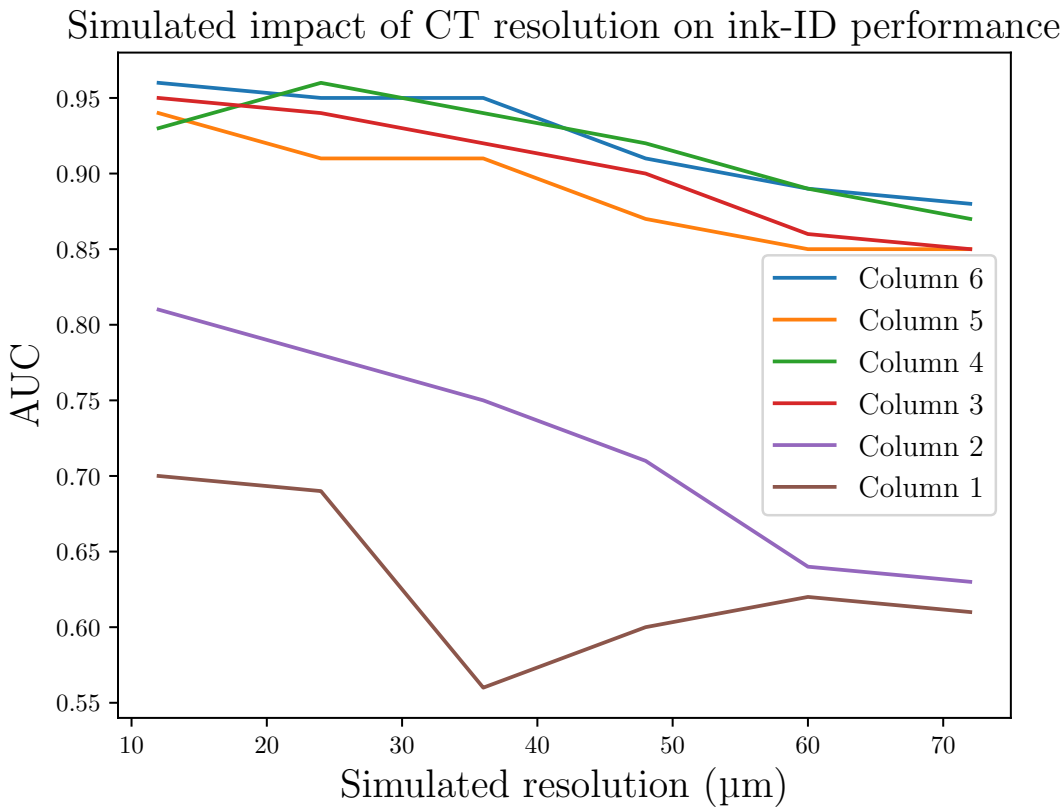


Figure 6.3: Simulated impact of CT resolution on ink-ID performance with Carbon Phantom columns, using volume downsampling. AUC: area under the receiver operator curve.

high enough not to be entirely absorbed by the scanned material. In some cases, with radio-dense materials such as metals, this threshold can be quite high.

Second, it is usually desirable not just to image the internal structure of an object but to image the *contrast* between multiple materials present in the structure. Based on the absorptive edges of the various materials or elements present, it is sometimes possible to choose an incident energy which is heavily absorbed or attenuated by one material but largely transmits through another. This is the case with manuscripts such as the En-Gedi scroll, where the incident X-ray beam is heavily attenuated by iron but transmits more easily through the parchment.

As usual, the Herculaneum scrolls are a challenging case. In addition to the ink and papyrus having very similar chemical composition, the carbon that comprises them has a low atomic weight, making it appear transparent at higher energies. This atomic weight has an absorptive edge in X-ray that is lower than CT sources and detectors can target, ruling out energy approaches that specifically target carbon.

The Herculaneum datasets used in this work have been acquired at three different energies: 22kV (P.Herc.Paris. Objet 59), 54keV (Diamond fragments), and 88keV (Diamond fragments again). 54keV is clearly sufficient for ink detection, as shown in the thorough experimentation on the Diamond fragments. P.Herc.Paris. 1 fr. 34 was separately processed using its 88keV CT images, and performed comparably. It is more challenging to compare the 22kV scan used for P.Herc.Paris. Objet 59, but visually the results also seem comparable. These early results suggest incident energy is nowhere near as important as resolution, but it remains possible that small gains can be achieved by choosing the optimal incident energy. More work could further investigate this, informing future imaging.

There *does* seem to be some difference between the CT images acquired at different incident energies, as one experiment training on 54keV data for P.Herc.Paris. 1 fr. 34 did not perform well on 88keV prediction data. As soon as a small amount of

88keV data was included in the training set, the model performed well, suggesting it needs to have been exposed to at least some samples from the energy distribution used during inference.

In addition to varying peaks of the incident energy distribution, there are also varied distributions that may play a role in enhancing ink detectability. Benchtop X-ray sources typically use an X-ray tube that generates a wide, or polychromatic, X-ray distribution across many energies. The peak of the distribution is noted as the incident energy of the scan (in kV), but is not the only energy present. Synchrotron (particle accelerator) sources are capable of monochromatic beams, where the distribution is much narrower around a single incident energy (measured in keV). Again, early results suggest both are sufficient for ink detection, with no clear winner yet. More work is needed to understand the impact of these changes to the ink detectability.

Future work could also focus on multi-energy scans, where the same object is scanned at multiple energies and the volumes are registered so that the input image can have multiple channels. With machine learning, it is trivial to simply change the number of channels in the input, allowing the model to extract what it can from the data. Adding information by sampling different points along the energy distribution can only help, when searching for subtle signals like carbon ink.

In general, the consistent performance across incident energies supports the morphological hypothesis. These results suggest the signal is not likely to be an intensity-based signal that is only captured within a precise range of incident energies. Instead, it appears to be present across a wide range, suggesting it may be more related to structural features or textural patterns that are present throughout this range.

6.2.3 Windowing

Similarly, an experiment was conducted to coarsely determine where the ink signal might sit within the image intensity range captured in the volumes. In a set of five experiments, the 16-bit image intensities were each time windowed into a different

bucket, clipping those values outside the window. The windows uniformly split the 16-bit range: $[0, (1/5)2^{16}), [(1/5)2^{16}, (2/5)2^{16}), \ldots [(4/5)2^{16}, 2^{16})$. The results (Figure 6.4) suggest that the signal is captured within multiple central intensity windows, indicating it is not tied to a narrow intensity range. This is again consistent with the morphological hypothesis.

6.3 Segmentation

The precision of the segmentation process can significantly impact downstream results. This section measures the impact of imprecision during segmentation on ink detection. Some remaining segmentation challenges are also characterized.

6.3.1 Required precision

Different methods produce segmentations of varying formats and precision. A perfect segmentation would follow the surface of interest with subpixel precision. As this is not easily achieved, it is helpful to know what level of precision is required for ink-ID to perform well.

Jitter is used to experiment with the segmentation accuracy in the local $\vec{n}$ direction. Jitter is implemented as an input parameter to ink classification, which moves the subvolumes origins in the $\vec{n}$ direction before the subvolumes are sampled. Jitter is specified as a number of voxels j along the surface normal $\vec{n}$. The subvolume center $\vec{v}$ is then modified prior to sampling from the volume:

$$\vec{v} = \vec{v} + U(-j, j)\vec{n} \tag{6.1}$$

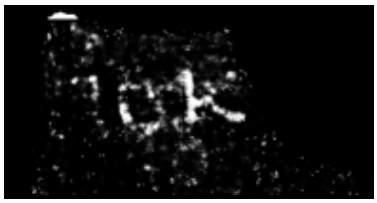
where $U(\cdot)$ is the continuous uniform distribution.

Used only during training, jitter is useful as a form of data augmentation. Figure 6.5 shows that validation loss decreases when $0 < j < d/2$, where d is the depth of the subvolume. In this range, the papyrus surface is likely still captured in the subvolume, but the model is trained on a wider variety of inputs, improving valida-

(a) Regions.

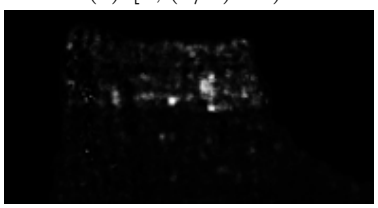
(b) Label.

(c) $[0, (1/5)2^{16})$.



(d) $[(1/5)2^{16}, (2/5)2^{16})$.

(e) $[(2/5)2^{16}, (3/5)2^{16})$.



(f) $[(3/5)2^{16}, (4/5)2^{16})$.

(g) $[(4/5)2^{16}, 2^{16})$.

Figure 6.4: ink-ID results with input volumes clipped to various windows within 16-bit intensity range. Blue: training, orange: prediction. Ink signal clearly exists within multiple central intensity windows, but not all.

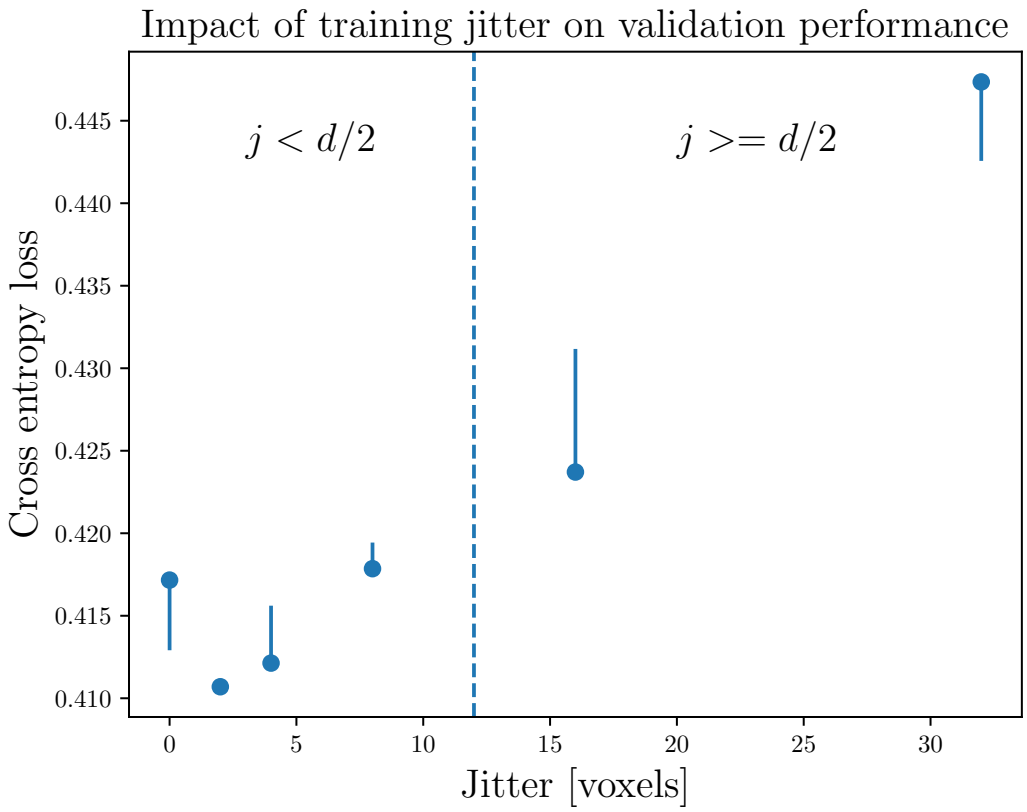


Figure 6.5: Impact of training jitter on validation performance. 2-fold experiments on the surface of P.Herc.Paris. 1 fr. 39 used.

tion performance. As $j >= d/2$, the sampled subvolumes are moved considerably from the initial segmentation, frequently enough that they no longer contain the papyrus surface or ink signal. The performance of the ink classification model decreases accordingly. This suggests that the segmentation needs to be precise enough that subvolumes capture the papyrus surface, which is dependent on subvolume depth. In other words, the segmentation should deviate along $\vec{n}$ no more than $d/2$ voxels from the correct segmentation, where d is the subvolume depth.

6.3.2 Tightly packed layers

The segmentation methods presented in this work are capable of large, precise segmentations from the scroll interiors, but rely on an air gap between layers. This air gap is not always present. Many layers are tightly compressed, and are difficult

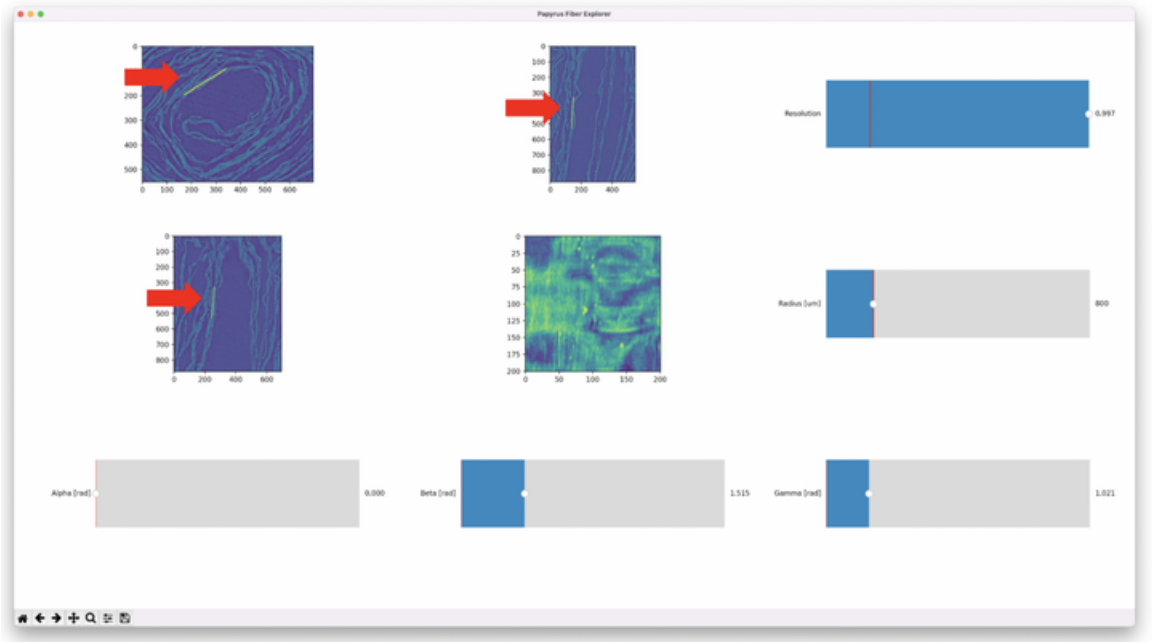


Figure 6.6: GUI developed to explore local nature of papyrus fiber structure in CT volumes. User changes 3D orientation of local slice. When local slice is oriented in-plane with papyrus sheet, grid-like papyrus fiber pattern becomes visible. Red arrows point to local slice intersections visualized in original volume. Center image shows local slice.

to differentiate in the slice view.

Though seemingly hopeless in the slice view, these layers are likely traceable due to the structure of the papyrus fibers in the CT volume. When slicing through a papyrus sheet orthogonally, only the cross section is seen, with minimal structure evident. When slicing in-plane with the sheet, however, there is much more detail present. This is seen in the large segmentations recovered from inside the intact scroll, where the papyrus grid pattern, nowhere to be found in the slice views, is shockingly coherent in the texture images.

A small GUI tool was developed to explore this, allowing the user to manually orient a local slice plane until it is in-plane with the papyrus sheet. When in-plane, the grid-like structure of the papyrus fibers becomes evident in the local slice image. Volume slice views display the intersection of the local slice, and allow the user to visualize the 3D rotation they are configuring.

This explorative tool supports the conclusion from the large texture images recovered from the intact scroll: there is more structure present in the data than is immediately visible in CT slices. By viewing things in the right way, they can be "unlocked," revealing more helpful clues for next steps. I suggest this means the internal wraps of the intact scrolls will be fully segmented, using existing scans, as algorithms improve.

6.4 Label alignment

The supervised methods discussed so far rely on accurate labels of the training fragment surfaces. The labeling hinges on the alignment step, where an image of the surface, typically infrared, is aligned to the texture image generated from virtual unwrapping.

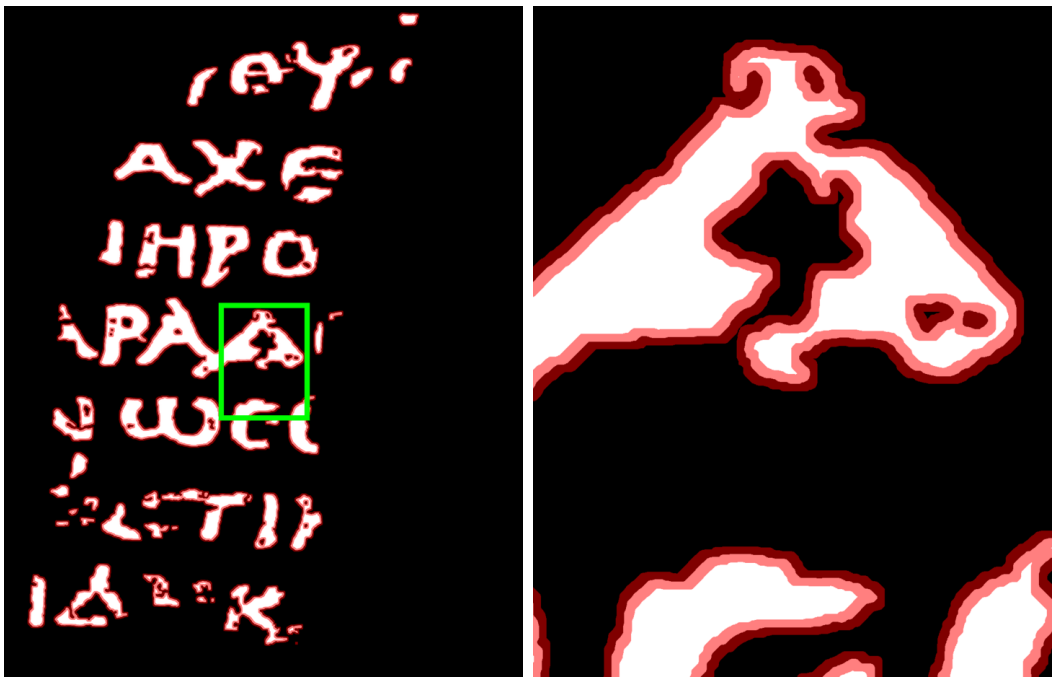
The existing alignment approach primarily uses the papyrus fiber structure in order to find landmark points that can be identified in both images. This process, however, is done manually, and is not pixel perfect. There is surely room for improvement in this step of the pipeline.

6.4.1 Mitigation

Though models can perform well with the existing labels, might their performance improve if noise can be tolerated or reduced?

Particularly for binary ink classification, the pixels nearest the ink/no-ink boundaries are the most likely to be mislabeled. These pixels are the first to be mislabeled if there is small misalignment between the infrared and texture images. These mislabeled pixels may be contributing noise to the dataset.

One approach is to simply remove these boundary pixels from the training set, focusing instead on the pixels that are more likely to have correct labels. The filtering of the so-called "ambiguous" pixels can be performed using edge detection and dilation in the ink label image. Figure 6.7 shows an example of these points highlighted on the ink label image for P.Herc.Paris. 2 fr. 47. The selected points near the label



(a) Ambiguous labels. (b) Detail.

Figure 6.7: "Ambiguous labels" highlighted in red along label boundaries by detecting edges and then dilating. These points not sampled during training.

boundaries are shown in red, and are not sampled during training.

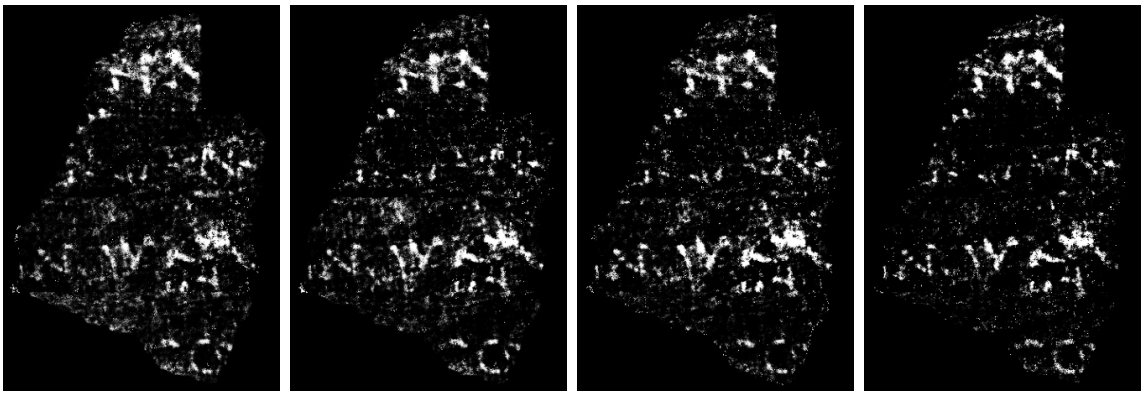
Various dilation radii were tried, to see how far from the label boundaries it was helpful to remove points from the training set. Figure 6.8 visualizes the results on four separate 4-fold jobs using P.Herc.Paris. 1 fr. 34. Up to a dilation radius of 16 pixels, filtering out the "ambiguous" points seems to reduce false positives and sharpen the results, though the change is slight. Beyond 16 pixels, performance drops off quickly when the dilation size approaches the ink stroke width (not pictured).

These results are slight, but would support the idea that there is slight misalignment in the label images, and that it can be helpful to ignore those labels most likely to be noisy, but that this is not critical for successful ink detection.

6.4.2 Improvements

There are multiple ways the label alignment might be improved in future work. One direction would be to use traditional image registration pipelines, but with more

(a) 4-fold. (b) Label.



(c) Baseline. (d) Radius: 4. (e) Radius: 8. (f) Radius: 16.

Figure 6.8: ink-ID results on P.Herc.Paris. 1 fr. 34 using varying radii (in pixels) for dilation step when filtering out "ambiguous" ink labels.

deliberate methods tailored specifically to these images. For example, whether using a feature- or intensity-based approach, the regions containing ink could be excluded from the iterative metric calculation. This would be easily done, particularly as these regions have already been masked when creating the binary ink label image. By dilating these masks, the ink could reliably be excluded, forcing the registration process to focus only on the papyrus fiber structure during the alignment.

The images could also be preprocessed using various filters, to align their representations as much as possible before registration begins. One example of this would be to find a suitable edge detector, and apply it to both images, aiming to highlight the edges of the papyrus fibers from either representation. Registration could then be performed on the outputs of the edge detector, where the two images would have been transformed into a more common domain. It may also make sense to apply the Fourier transform to both images, converting their representation to the frequency domain. Here, the frequencies making up the fiber structure could be isolated, where others could be masked out, reducing the images to a more basic "fiber pattern only" image. The resulting images could be registered directly in the Fourier domain, or could be translated back to the spatial domain before registration.

Learned approaches could also be explored as alternatives that may simplify this process. The difference between the features present in the infrared and texture images have been discussed at length, and are varied and complex. Might a learned model be able to navigate these differences more fluently than a hand-implemented approach? Related work has had promising results in similar areas, for instance the registration of different images of paintings using only the detailed crack structure of the paint [54]. Elsewhere in computer vision, there are methods for incorporating a learned alignment step to the image inputs, for instance allowing small alignment corrections to be learned for NeRF input images [74]. It may also be possible to train paired or unpaired image-to-image translation models between the infrared and

texture images. If a model can faithfully produce one from the other, for instance a texture image from an infrared photo, then the produced texture image could be registered to the real texture image using traditional registration methods.

When considering improvements to the label alignment, it is helpful to evaluate the ceiling of model performance if label alignment can be perfected. Section 4.6 shows that, with perfect label alignment, ink-ID models are capable of remarkably sharp prediction images. This suggests that through registration improvements or by building models that are more robust to label misalignment, very sharp images will one day be produced of the interior wraps of the Herculaneum scrolls, revealing their text in high detail.

6.5 Spatial support

As it dynamically samples subvolumes from the CT volumes during training and inference, ink-ID allows easy experimentation with the subvolume dimensions. What is the optimal size and shape for a subvolume in order to capture the ink presence? What does ink-ID need to see in order to determine whether or not there is ink present? In addition to improving ink detection results, this should inform our knowledge of what ink-ID is learning. All experiments in this section use column 2 of Carbon Phantom, isolating the central Θ character as the validation set and using the other four characters as the training set.

First, cube-shaped subvolumes of varying sizes were tried to get a general feel for the spatial extents necessary for ink prediction. For each size, the experimental configuration described above was run, and the validation results were recorded. The resulting performance plot across subvolume sizes is shown in Figure 6.9. Unsurprisingly, larger subvolumes include more spatial context, and lead to improved ink-ID performance fairly consistently.

This same experiment was repeated, this time isolating the width or depth of the subvolume to gauge their impact on ink detection. The plot across width is shown in

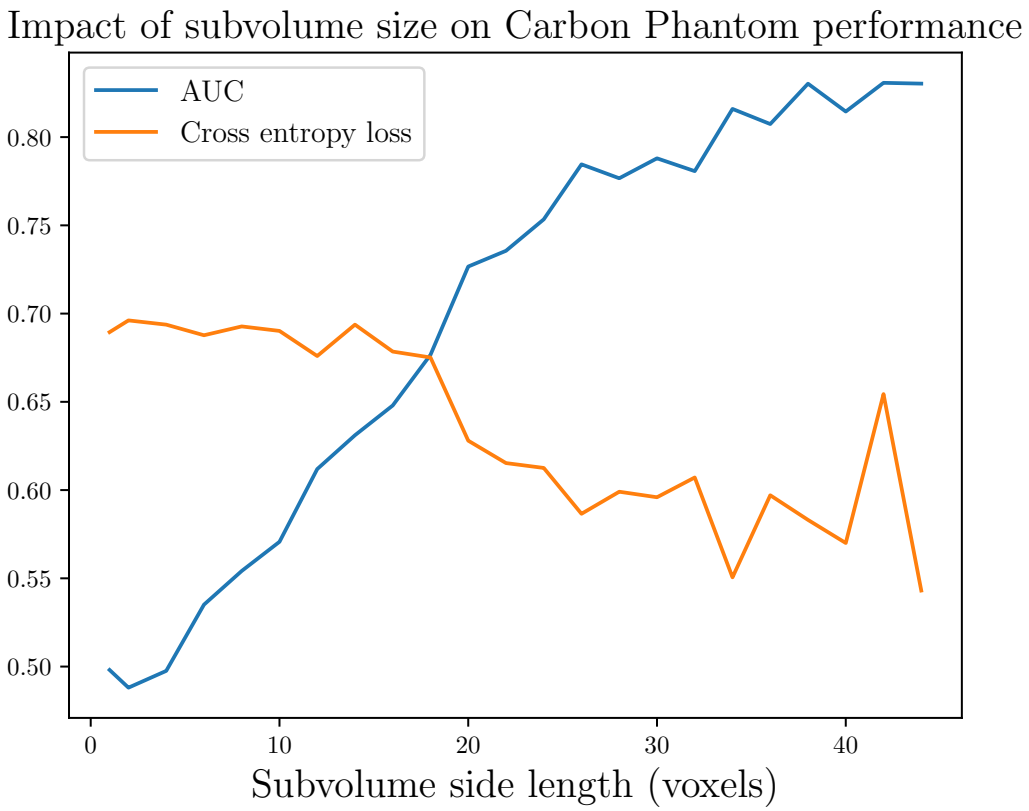


Figure 6.9: ink-ID performance on Carbon Phantom, column 2, with cube-shaped subvolumes of varying sizes. Larger subvolumes include more spatial context and improve ink detection.

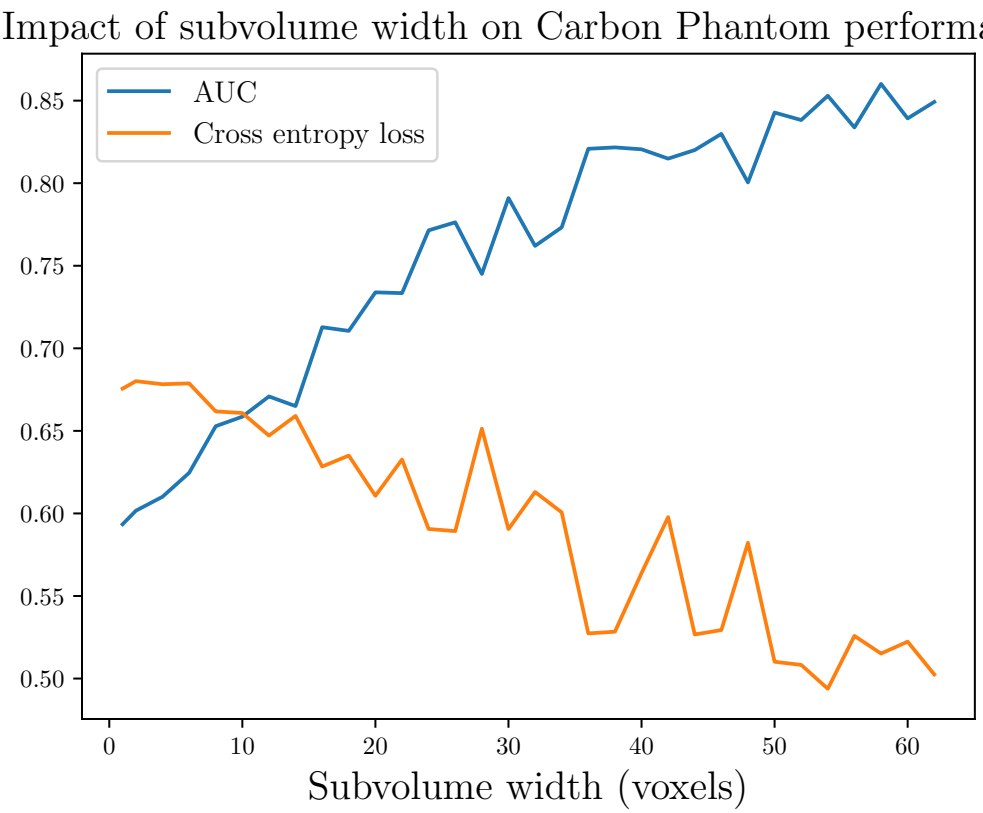


Figure 6.10: ink-ID performance on Carbon Phantom, column 2, adjusting subvolume widths. Depth held constant at 24 voxels. Wider subvolumes lead to improved ink detection and do not appear to have plateaued.

Figure 6.10, and depth in Figure 6.11.

Both wider and deeper subvolumes improve ink-ID performance, but depth appears bounded while width does not seem to have plateaued. Intuitively, this makes sense and aligns with the morphological hypothesis. It is necessary to capture enough depth that the surface and the ink are reliably captured, but additional depth does not help, moving out of the papyrus layer and into less relevant features. Width, on the other hand, continues to add spatial context that may be useful for ink detection. Particularly if the model is detecting textural patterns, larger spatial contexts would be helpful for observing these. These findings suggest, wider, shallower "pancake" subvolumes are optimal.

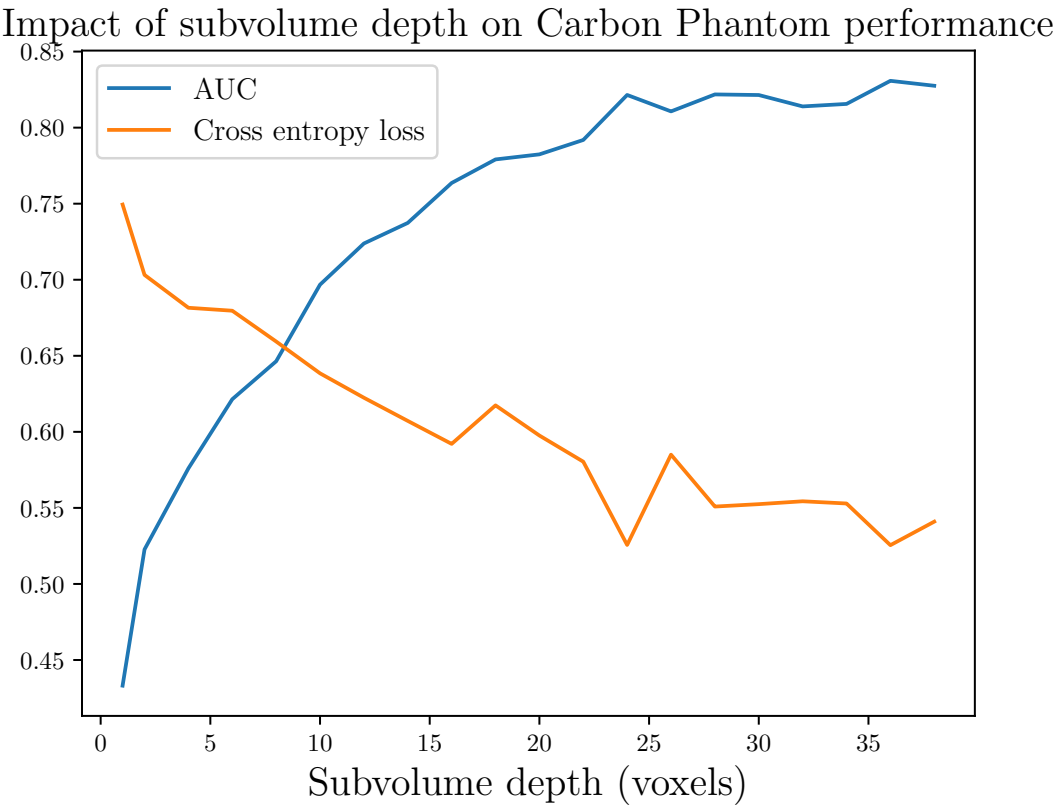


Figure 6.11: ink-ID performance on Carbon Phantom, column 2, adjusting subvolume depths. Width held constant at 48 voxels. Deeper subvolumes lead to improved ink detection, but only to some extent.

As discussed in Chapter 7, these findings were validated and extended during the Vesuvius Challenge. Though the datasets differ, the implications seem to apply to both in broad strokes. For the Ink Detection Progress Prize, the highest-performing team took this concept to the extreme: their model inputs were 16 voxels deep and 1024 voxels wide!

6.6 Model and training procedure

Many tweaks to the model and training procedure were attempted in the search for improved ink detection. A handful are listed here:

- Balanced sampling to address the ink/no-ink class imbalance. This changed the false positive rate in the prediction images, but did not improve readability.
- Autoencoder pretraining before ink-ID training. An attempt to have more expressive models that learned a wider variety of image features. No measurable impact.
- Multi-task learning with an autoencoding task in addition to ink classification. An attempt to learn more expressive models, and to smooth the learning curves. Successfully smoothed learning, but did not lead to improved results after convergence.
- Various changes to the convolutional model architecture. Minimal impact.

The mediocre results from these experiments suggest that sophisticated model and training procedures are not yet the bottleneck for ink detection. As this is a new method with relatively small and probably noisy label sets, it is perhaps not surprising that the largest strides to date have come from improvements to the data processing pipeline rather than machine learning techniques. The Ink Detection Progress Prize from the Vesuvius Challenge seems also to have largely confirmed this, with winning teams reporting that data sampling and augmentation were major components of their submissions.

6.7 Nature of dataset

In addition to using higher resolution CT images, the proof of concept experiments with scroll fragments have had another advantage over the ultimate task of reading the intact scrolls. Almost all experiments shown so far have split the fragment surfaces, training on part of one surface while predicting on another part. Even in 4-fragment experiments, 8-fold cross-validation was used. This was done because it was desirable to see an upper bound on what ink-ID could be expected to achieve, but it is not realistic for the intact scrolls. For the intact scroll case, the model will not have been trained on any labeled data from the inference scroll. Is ink-ID capable of bridging this domain shift?

One experiment changed the standard 8-fold configuration to a 4-fold experiment across entire fragments. This experiment is designed to test whether ink-ID can generalize to CT volumes it has never seen. The results are shown in Figure 6.12. ink-ID still performs strongly, suggesting this is not a concern.

In addition to predicting on CT volumes it has never seen, ink-ID will have to generate predictions for *scrolls* it has never seen. Different scrolls may have slightly different ink compositions, ink thicknesses, and so on. Is this possible? As the four fragments used in these experiments come from two scrolls (two fragments each), it is possible to test this. Figure 6.13 shows the results of training a models in a 2-fold experiment, where one model trained on the fragments from P.Herc.Paris. 1 and the other trained on those from P.Herc.Paris. 2. Again, ink-ID seems content to handle this shift, even though these two scrolls have a visibly different hand in their reference images.

These experiments cannot offer any guarantees for the intact scrolls, about which little is known. Though ink-ID was able to handle this particular domain shift, there is a chance the ink in the intact scrolls that have been imaged is different enough to cause a problem. This seems unlikely, though, considering the rather consistent

(a) 4-fold. (b) ink-ID results.

Figure 6.12: ink-ID results across the four "Diamond fragments" using 4-fold cross-validation.

(a) 2-fold. (b) ink-ID results.

Figure 6.13: ink-ID results across the four "Diamond fragments" using 2-fold cross-validation.

performance so far across various fragments from a number of different scrolls. In general, more and more *varied* training data is the strongest preventative measure to avoid domain transfer problems between scrolls.

6.8 Summary

This chapter has explored the limits of the various pipeline components, deepening our understanding of what factors enable successful ink detection and what should be prioritized in future work.

As far as imaging is concerned, resolution is currently the single most important factor. For the Herculaneum scrolls in particular, the morphological hypothesis as well as empirical evidence suggest that ink detection in rolled scrolls is just on the threshold of feasibility. Based on the resolution that is possible with current instruments, it is preferable when imaging a Herculaneum scroll to prioritize the highest achievable resolution (smallest voxel size) above all else. Experiments simulating lower resolution with downsampling seem to suggest the ink signal is often recoverable at surprisingly low resolutions, but more work is needed to confirm that this maps to actual scans. Incident energy may also turn out to be a meaningful factor, but has not led to sizable changes in ink detectability with existing data.

Improved algorithms will be necessary for the complete segmentation of intact scrolls. The slice views showing tightly compressed layers make this look impossible, but texture images from within the rolled scrolls show that much more detail is captured in the images when it is visualized in the right way. This suggests that complete scroll segmentation will likely be achieved using existing images.

These experiments have also helped develop an understanding for what it is ink-ID is detecting. Testing various subvolume sizes, as well as a host of other experiments, seems to support the morphological hypothesis that ink-ID is detecting some textural difference between ink regions and those without ink.

Finally, some feasibility studies were conducted to evaluate the basic expectations

when applying ink-ID to existing images of rolled scrolls. Though the imaging distribution domain shift makes this a significant challenge, ink-ID does at least seem capable of generalizing across different scrolls, and of detecting ink at ~ 8 µm voxel size. Though this provides no guarantees, it is very promising. As the pipeline components are refined, ink detection will improve, strongly suggesting optimism as we look forward with this work.

CHAPTER 7. THE VESUVIUS CHALLENGE

7.1 Introduction

In early 2023, we were presented with a unique opportunity to open this research problem to the global community. Nat Friedman, Silicon Valley investor and entrepreneur, had the idea to launch a research contest for which he and Daniel Gross funded an initial prize pool. The contest would use the datasets established in this work, inviting others to reproduce, validate, and contribute to the various pipeline components. The algorithms developed in this work were also released as the benchmark method.

The Vesuvius Challenge launched on March 15, 2023 (Figure 7.1). The website (`https://scrollprize.org`) contains prize information, dataset downloads, and tutorials and visualizations to introduce new contestants to this work. Parts of two intact scroll scans were released, as well as three of the four Diamond fragments and their associated label images. The contest is centered around a grand prize, the objective of which is to read an intact scroll, and there are also a number of smaller "progress prizes" awarded in order to guide the community towards this ultimate goal. The prize pool started at $250,000, and grew within days of the launch to $1,000,000+ as more sponsors contributed. As of this writing, the Vesuvius Challenge is ongoing, and it is too early to offer broad conclusions about the outcomes.

This chapter will instead discuss the contest specifically in relation to the work presented in this dissertation. The Vesuvius Challenge has provided the unique opportunity to have this work tested, validated, reproduced, visualized, disseminated, and extended. This chapter will briefly overview some of the ways this has manifested to date.

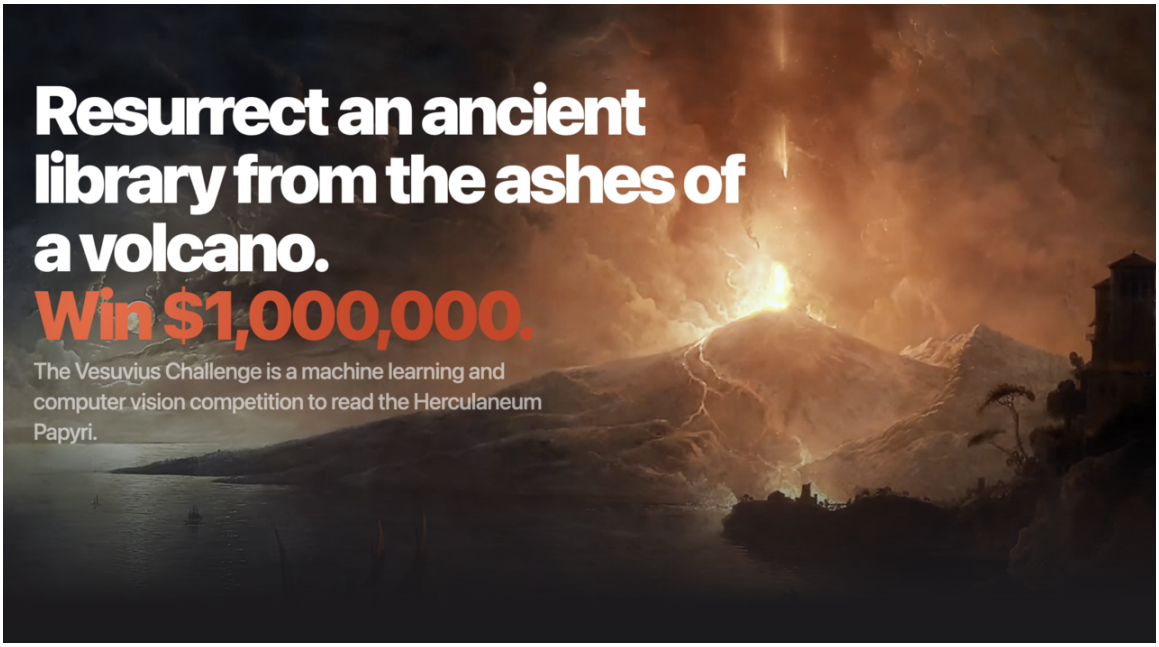
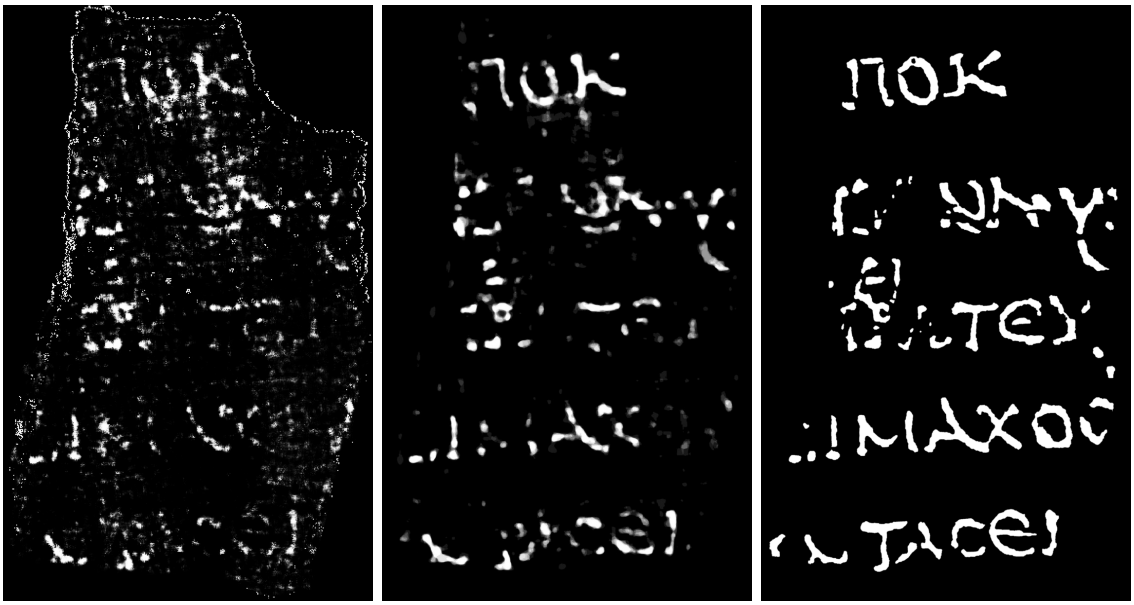


Figure 7.1: `scrollprize.org` landing page.

7.2 Ink Detection Progress Prize

As a result of the work in this dissertation, the baseline was robust enough and the dataset mature enough to support an organized ink detection competition. The Vesuvius Challenge therefore launched with an Ink Detection Progress Prize hosted on Kaggle (`https://kaggle.com`) to encourage optimization of the ink detection subproblem. P.Herc.Paris. 1 fr. 39 was reserved as a held-out evaluation fragment, and the others were released as training data. Submissions were evaluated on P.Herc.Paris. 1 fr. 39 using the binary classification ink detection task, with F0.5 as the scoring metric. Contestants were provided with ink-ID as a reference implementation, but were encouraged to develop their own methods.

The Ink Detection Progress Prize closed a few weeks ago, with submissions from 1,249 teams. The results are a resounding confirmation of the detectability of Herculaneum ink in CT. Figure 7.2 roughly visualizes the ink detection achieved by the top competitors, alongside the ink-ID baseline and the ink label image. These results confirm those of ink-ID and show further convergence towards the ground truth,



(a) ink-ID baseline. (b) Top Kaggle results. (c) Ink label.

Figure 7.2: State of the art ink detection from Kaggle ink detection progress prize. Submissions are binary ink classification on surface of P.Herc.Paris. 1 fr. 39. Top ten submissions averaged together. Baseline ink-ID results also shown.

validating that much of the ink signal is recoverable.

Additionally, as the entries must be open sourced to be eligible for prizes, this progress prize has resulted in a new set of methods and experiments that extend and inform ink detection. There have been informative findings regarding the data inputs as well as the model architectures.

Many teams studied the optimal CT input dimensions for ink detection. Regarding subvolume depth, this work presented in Section 6.5 that increased depth aids ink detection up to a point, before it plateaus. Subvolume width, on the other hand, continues to improve ink detection much longer. The experiments in this work did not seem to find the bound to this increase in size. This suggests a flatter "pancake" shaped subvolume is optimal, supporting the morphological hypothesis as wide regions of support allow models to observe textural surroundings rather than only local intensity shifts.

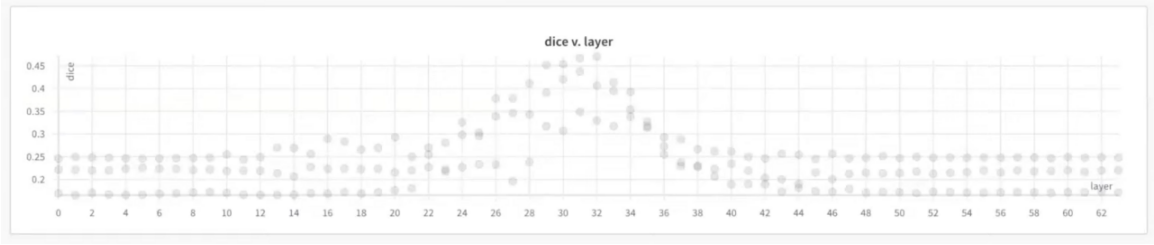


Figure 7.3: Predictive power of each surface volume layer, illustrating where ink signal exists along surface volume depth. For each of 65 surface volume layers, model trained on that single layer from two fragments and validated with third fragment. Dice loss shown. Top: validation on P.Herc.Paris. 2 fr. 47, middle: validation on P.Herc.Paris. 2 fr. 143, bottom: validation on P.Herc.Paris. 1 fr. 34. Image used with permission from winning Kaggle team "ryches."

The investigations shared by top Kaggle teams reproduce these findings. The top team took the "pancake subvolume" idea to the extreme, sharing that the inputs to their ink detection model were $16 \times 1024 \times 1024$ voxels. Figure 7.3 illustrates one of their experiments that motivated these final subvolume dimensions. This experiment illustrates the predictive power of each layer in the surface volume, effectively plotting where the ink signal resides within the depth of the surface volume. The results show that the ink signal commonly resides primarily within the central 16 layers, corresponding to a thickness of roughly 52 µm. This signal thickness likely incorporates both the physical thickness of the ink layer and the small segmentation errors introduced by Canny edge detection.

Competitive Kaggle entries largely used model architectures that allowed for 2D spatial context, such as U-nets [70] and related architectures. The dimensionality of the individual layers varied, with 3D and 2D convolutional approaches, as well as many that combined both with 2.5D approaches. For example, some entries used 3D convolutions before applying max pooling across the subvolume depth to reduce to 2D outputs.

Like ink-ID, these models take 3D subvolumes as input, but unlike the ink-ID architectures shown so far, they output 2D image patches of ink predictions instead

of a single binary prediction for each subvolume. As expected, this enables models to leverage the spatial dependencies inherent in the ink labeling: for example, those label pixels adjacent to an ink pixel are more likely to be ink than those adjacent to a not-ink pixel. The resulting prediction images (shown in Figure 7.2) therefore contain less noise.

Many other findings resulted from the Ink Detection Progress Prize. For example, multiple competitors found small refinements to the data labels that they shared and I incorporated into the dataset. The immediate next steps from this progress prize are exciting: first, a more complete review of the methods used by top competitors will be conducted, and many of them will be incorporated into ink-ID. Second, these models can be used to generate non-binary prediction images. The Kaggle submissions were necessarily thresholded and binarized for scoring, but for pure visualization this is not necessary, and more ink signal may become visible. Finally, it will be possible to extend the winning methods to predict in infrared instead of binary classification, further refining the multimodal results shown in Chapter 4.

7.3 Segmentation

Other prizes have encouraged progress within the segmentation subproblem, both using existing tools to perform more segmentation, and developing new tools to accelerate this work going forward. Figure 7.4 shows the total area in cm^2 that has been segmented by the Vesuvius Challenge community. The large spike near 4/30/2023 represents the two segmentations from this work, shown in Figures 5.7 and 5.8, which were released to the community. While these are still the largest individual segmentations, the total segmented area from inside the rolled scrolls has more than tripled this result. There have been many open source tool submissions to create and visualize the segmentations as well. From the segmentations produced, exploratory data analyses similar to those carried out in this work are already uncovering many interesting findings.

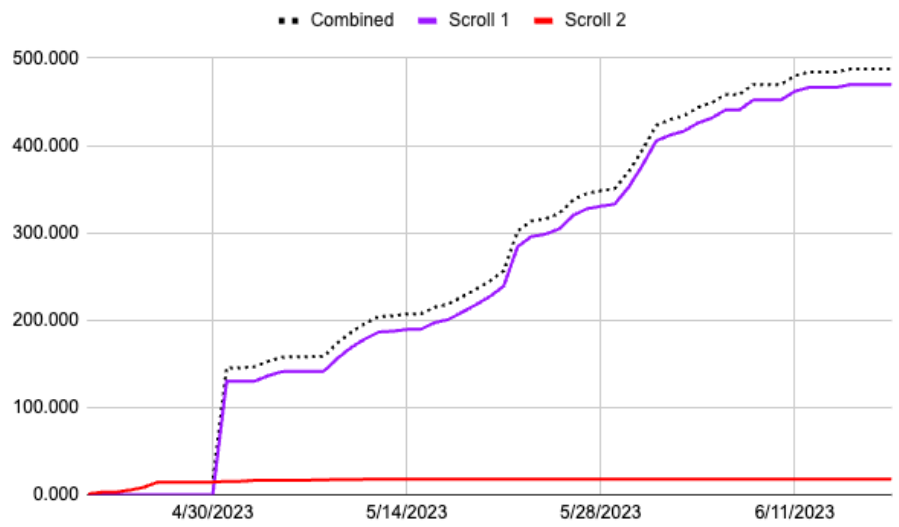


Figure 7.4: Total segmented area from intact scrolls by Vesuvius Challenge community over time, in cm^2. Large spike near 4/30/2023 is two large segmentations from this work, shown in Figures 5.7 and 5.8. While still the largest individual segmentations, total segmented area has now more than tripled this result.

7.4 Summary

It is too early to comment on the outcome of the grand prize, which runs through the end of the year. It is already clear, however, that the Vesuvius Challenge has enabled the widespread dissemination and validation of the principal findings of this dissertation. Ink detection has been thoroughly reproduced, confirming the presence of some detectable signal in X-ray micro-CT that captures the carbon ink of the Herculaneum papyri, despite its tendency to elude visual observation. This result was enabled as a result of the open-source nature of this work, releasing code and datasets and inviting the world to contribute.

There has also been broad interest generated in this dataset and research problem, building a community that will help sustain forward progress. A Vesuvius Challenge Discord server has 1,305 members at the time of writing, working together daily to advance this research problem. There have been efforts to make proxy scrolls and carbonize them, shared datasets contributed, helpful pull requests to our software

tools, and too many fun quotes to include. Someone bought a surplus micro-CT scanner and is fixing it up at home, sharing ongoing project updates and intending to scan proxy scrolls to learn more about the nature of carbon ink in CT.

I am grateful to Nat Friedman and the Vesuvius Challenge team he assembled, which has already fulfilled its mission of accelerating progress with months remaining in the challenge. Well outside of my comfort zone, this unique experience has also pushed me in my own work, and I am excited to see where it takes us.

CHAPTER 8. DISCUSSION

This chapter will discuss the current status and outlook of the effort to read the Herculaneum scrolls, list desired next steps, provide broader reflections, and conclude with a summary of the contributions in this work.

8.1 Status and outlook

Prior to this work, the idea that machine learning could detect the "invisible" carbon ink of the Herculaneum papyri was just that: an idea. Very early experiments had seemed to suggest there might be something detectable, but there was no apparatus to validate the concept. As those tools have been developed, ink detection has consistently shown promise, and has improved as the various stages of the pipeline are refined.

Today, with sufficient imaging resolution, ink is recoverable with an accuracy that generates usable text for scholars and does not introduce meaningful false positives. Already, a handful of characters have emerged from the hidden layers of the scroll fragments, previously hidden for 2,000 years. It took multiple years to build up the framework necessary to make these findings, and the revealed characters have all come from the last few months at an accelerating pace. Using existing images and techniques, more characters will follow in the near future. It is likely that this research problem is at an inflection point, suggesting that acceleration will continue.

The CT images used and released through this work are incredibly rich datasets that will yield many more findings. With focused work using these images, it will be possible to better understand the optimal resolution and incident energy necessary for ink recovery in the intact scrolls. This situation will only improve as more CT images are acquired, diversifying the training dataset and teaching us more about this collection.

The domain adaptation problem between fragment scans and scroll scans is a chal-

lenge, but experimental evidence so far suggests it should be possible to extract at least some text from these scans. As this begins to happen, collaboration is necessary to support scholarly needs when processing born-digital images [75].

The large physically unrolled scroll sections, glued to backing boards in trays, also contain hidden layers of text. As these trays are the subject of active scholarly work based on their exposed surfaces, these hidden layers are of particular interest to some scholars. Unfortunately, the large width and high aspect ratio of these trays make them a challenge to image using conventional CT. I expect that the texts of the rolled scrolls will be more sooner recovered than these hidden layers. There is promise, though, in the development of new imaging methods for these trays, for example approaches based on tomosynthesis [76].

In general, every finding of this work suggests the upper bound for ink detection has not yet been encountered.

8.2 Future work

Future work in the following areas is likely to further improve the results of this dissertation.

8.2.1 Evaluation

This work has simultaneously developed a dataset, a baseline method for the dataset, and evaluation techniques for the method. As the work progressed, much was learned and improved in each of these areas. As a result, the evaluation methods shifted considerably throughout the course of the work, starting purely with visual evaluation and then evolving through a number of quantitative metrics.

Taken as a whole, the evaluation of various experiments therefore lacks consistency. Ablation experiments were performed across a variety of datasets using different metrics, so are difficult to compare directly. Now that the method has converged enough that the output of quantitative metrics informs readability, it will be very helpful to develop a consistent train/test data split with consistent metrics across experiments.

8.2.2 Visualization

Consistently, visualization tools have been worthwhile time investments with these datasets. I developed a number of visualization tools for this work, which together have enabled and considerably aided the data processing pipeline, but there remains room for improvement. The existing tools are separate, are launched with command line parameters, and leave out some exciting visualization ideas. For instance, a consolidated GUI showing CT volume views alongside segmented and flattened surfaces could be very helpful. One could perform segmentation, immediately see the effect on the texture image, and perhaps even see the output of an ink detection model directly on the segmented mesh. Possibly, one could further adjust the segmentation from the flattened mesh view, pushing and pulling mesh points to move them in 3D.

It would also be helpful to have more integrated viewers that display all segmentations for a given object in their 3D context, or that allow one to navigate between multiple objects without closing a window and running another command. It would also increase engagement to have these tools be more readily accessible to a broad user base, for instance by developing web interfaces that pull data from remote servers and do not require installs or large data downloads.

8.2.3 Volume management

There are some engineering improvements to ink-ID and related tools that would greatly accelerate various steps. One would be to store CT volumes not as full-width slices, but instead as 3D chunks. ink-ID and other tools could fetch only the necessary chunks, instead of fetching full slices even if only a small cropped region of the slice was necessary. This would greatly reduce RAM requirements, which are presently still prohibitively large.

These chunks could be stored at multiple resolutions, so the full resolution images are fetched only when necessary, while graphical tools for instance could most of the time use lower resolution images without any user impact. Storing a multiresolution

set of chunked volumes at successively halved resolutions would less than double the disk space of the volumes while dramatically speeding up or enabling other scenarios. There are available libraries designed for this purpose [77, 78]. ink-ID implemented some of these in initial experiments that introduced prohibitive time performance penalties but were otherwise successful. This is an engineering problem that should be tractable going forward.

Some volumetric computation, for instance the generation of texture images, could also be accelerated by utilizing the graphics processing unit (GPU). One library and web interface does this in addition to using chunked volumes, and is a great example of where this work could go in the future [79].

8.2.4 2D labels

The geometric framework behind ink-ID was originally developed to link 2D and 3D spaces using the per-pixel map. Subvolumes were sampled directly from the original CT volume during training and inference. This approach is flexible, but also makes it challenging to use image-to-image models: individual points on the PPM map to 3D, but there is no straightforward way to sample PPM image patches from the 3D volume. ink-ID predictions therefore resemble "pointillism," generating separate predictions for individual neighboring pixels in the prediction image.

Surface volumes were largely developed to get around the system memory bottleneck at runtime, but also introduce an opportunity to discard the PPM and use 2D $\rightarrow$ 2D image-to-image methods on the surface volumes and label images. Such methods could leverage a wide literature of image-to-image models, and would factor spatial dependencies into their predictions. Image-to-image models using input and output image patches are the natural next step, using model architectures such as U-net [70] and related approaches.

8.2.5 Vesuvius Challenge

Fortunately, every one of these areas for future work has been identified and advanced by the Vesuvius Challenge community.

The Kaggle Ink Detection Progress Prize required the formalization of train and test sets, isolating P.Herc.Paris. 1 fr. 39 as the test set and using the other three Diamond fragments as the training set. F0.5 was also established as the primary metric. Now that there are established results using this experimental configuration, I suggest future work could use the same setup for more directly comparable evaluations.

Multiple community members have been working on visualization and volume management solutions, and the top Ink Detection Progress Prize teams consistently used image-to-image architectures such as U-nets in their implementations.

8.2.6 Tearing it down

This work has built up a fairly involved software pipeline of successive steps. Multiple steps require manual user input, and most are tailored to specific datasets either in their design or using deliberate parameters during training or inference. Looking around at other fields today, one often sees multiple pipeline components consolidated into a single, more powerful machine learning step. Hand-coded photogrammetry pipelines are being replaced by learned NeRFs [80], sophisticated natural language processing tools are being replaced by powerful generative models [81], and so on. Might it be possible to do something similar with the pipeline presented in this work?

Being so close to this pipeline, I am likely not the best candidate for this way of thinking, but I will nonetheless speculate. It seems possible that segmentation in particular could be simplified, absorbed into ink detection, or eliminated. One approach would be to train an ink-ID model with random subvolume orientations instead of using the surface normal vectors. The resulting model could tolerate any input orientation and still accurately report the presence of ink at the subvolume center. This model could then be applied to seed points directly from the inference

volume, perhaps sampled using 3D edge detectors, without the explicit need for a fully meshed segmentation.

Another approach to the same idea would be to use the 2D $\rightarrow$ 3D mapping from the PPM to "push" the ink-ID surface labels back into the volume, generating volumetric ink labels. A volumetric 3D $\rightarrow$ 3D ink-ID model could then be trained directly in the volume space, sampling the same input subvolumes from the CT volume and the label volume. The volumetric outputs at inference could be accumulated across a large volume, highlighting ink spots in 3D without any segmentation step. Successful volumetric ink labels require that the depth or thickness of the ink signal is known, but this is being achieved as shown in Figure 7.3.

These approaches are appealing not only because they might improve performance, but because they could simplify the pipeline, removing some of the small errors that accumulate through the data processing steps. That said, simple is of course not always easy.

8.3 Reflections

In this section I offer some broader reflections about my experience and this work.

8.3.1 Application to other domains

This work has been highly focused on one applied research problem, but I believe the implications are far-reaching. To my knowledge, it is not widely accepted that multidimensional images capture significantly more than humans are capable of seeing. Machine learning approaches using high-dimensional images are still largely trained on features that can be labeled by human annotation. The idea of registering images of different modalities to achieve labeling and subsequent model-based detection seems broadly powerful.

Medical imaging is one domain in which some version of this idea seems to have growing acceptance. For example, breast cancer screening datasets now use not only human annotated mammograms, but also ground truth labels using later screenings

and biopsies [82–84]. These datasets capture some positive labels which are missed by human experts in the mammogram labeling, as they have not yet progressed enough to be detectable by eye. As with ink-ID, it is shown that models are often capable of detecting these early and subtle signals that evade human eyes.

One other example domain is archaeology, using remote sensing to identify points of interest on the ground. Often this work will use spectral image stacks with human annotated labels identifying the points of interest. There are points of interest, however, that humans cannot identify visually in the spectral images, despite being able to identify them in person from the ground. As a result, the trained models do not learn to identify these more subtle locations. It seems possible that some of these are *captured* in the spectral image data, just not in ways that are readily apparent to a human visualizing the high-dimensional spectral data. If ground truth can be generated using site visits, these labels could be added to the dataset, and it seems likely that models could then learn to identify more subtle points of interest than before.

It is clear that in some domains, computational approaches are capable of detecting signals in high-dimensional images that cannot be seen by humans via direct visual inspection. There is some question of whether or not this is surprising. On the one hand, the human visual apparatus is a highly optimized pattern recognition system. Using slice views or other visualizations, humans can look at the exact same voxels that are input to ink-ID or other models. This would suggest we should be able to see anything in the image that the algorithm can see.

With available visualization methods, ink-ID conclusively shows otherwise. I suggest this is not actually very surprising. Though the human visual system is highly optimized, it is optimized for very specific tasks. Humans are experts at processing two-dimensional images in the visible wavelengths. But major compromises are made any time we attempt to visualize high-dimensional images such as 3D CT scans or

multispectral 2D images. We are reduced to methods that necessarily remove information: slice views, volume rendering, false coloring, principal component analysis and so on. These are all powerful tools, but none allow us to fully digest the original input image. Humans are simply not wired to take in a complete view of a volumetric image. Computational approaches are not limited by this wiring, and may have the advantage.

8.3.2 General

The methods presented in this dissertation are the result of chasing a single experimental pathway, and have yielded exciting results. This is likely not the only viable pathway, though, and I am eager to see how this area develops.

Naturally, there have been cycles in this work of feeling stuck and breaking through. I have found that the largest breakthroughs, which from my perspective were surface volumes and the coarse-to-fine segmentation approach, were simple combinations of existing components. It simply takes time for these ideas to crystallize. Right now, I do not feel stuck, but believe the opposite: too many ideas, not enough time!

It is tremendously helpful to spend time with the scholars who study this material, and usually not in ways that one expects. I believe the discipline should prioritize the collocation of multidisciplinary teams for weeks or months at a time, even without a clear agenda. Similarly, one learns every time they interact with the material in person. Notably, the scroll fragments are shockingly small in person, and provide a great perspective of the minute scale of the pixels and voxels labored over in the datasets.

In general, it has been the experience of a lifetime to work on this project, and I feel very fortunate and grateful for the opportunity. I will never forget being the first person to see Greek characters that had been hidden since a human hand wrote them 2,000 years ago!

8.4 Summary

This dissertation has presented the following research contributions:

- **Geometric framework:** A novel geometric framework, combining 3D and 2D images, enables the creation of first of their kind labeled datasets for supervised learning.

- **ink-ID:** A reference implementation is presented of the machine learning component, combining dynamic data generation, model architectures, training and inference routines, and dataset management and visualization.

- **Ink detection results:** Experimental results show conclusively that the carbon ink of the Herculaneum papyri can be recovered from X-ray micro-CT alone (Figures 3.39 - 3.42).

- **Evaluation:** Visual, quantitative, and papyrological evaluation methods are presented to validate the ink detection approach using ground truth images.

- **Exploratory data analysis:** Using the developed tools, visual results are presented to indicate that carbon ink in CT is more complex than previously thought. In rare cases, it is visible directly.

- **Multimodal transformations:** A powerful generalization of ink-ID is presented that allows models to learn transformations between image modalities, for example producing simulated color photographs from CT inputs.

- **Segmentation methods:** A novel coarse-to-fine segmentation method generates large, precise scroll segmentations with minimal user input. Using this method, the largest papyrus sheets ever recovered noninvasively from Herculaneum papyri are presented (Figures Figures 5.7 and 5.8).

- **Hidden text:** For the first time ever, text characters from the hidden papyrus layers of the Herculaneum scrolls are revealed noninvasively, not seen in 2,000 years until now (Figures 5.12, 5.15, and 5.16).

- **Domain shift:** The domain shift or generalization gap between fragment and intact scroll scans is characterized, informing next steps for ink detection on the intact Herculaneum scrolls.

- **Ablation:** An ablation study evaluated the various pipeline components, probing their limitations to suggest where future work would be best targeted.

- **Data release:** Via the Vesuvius Challenge, the data and methods developed in this work have been released and put in the hands of an active user community, already accelerating progress and moving this field forward. The research contributions above enabled the creation of this dataset, which is now mature enough to support this level of engagement and scrutiny.

To conclude, this work has shown that CT images of the Herculaneum papyri capture more subtle patterns than was previously thought. Using a geometric pipeline and machine learning, it is possible to noninvasively recover the presence of carbon ink from these scans, revealing hidden texts that have been unseen for the last 2,000 years. Generally, this work suggests that high-dimensional images such as volumetric CT or multispectral photographs capture more than the human visual system can readily process. By leveraging aligned labels achieved using other image modalities, trained models can detect patterns previously considered invisible.

This work has shown that even though computational approaches can detect carbon ink in CT, it is often not visible to human eyes. Much of this work has therefore investigated the question of *what* the model is detecting. These experiments so far support what I call the *morphological hypothesis*: that the presence of ink is captured in subtle, three-dimensional textural patterns. Many experiments were designed to

probe the boundaries of this ink signal, and their results have yielded many findings: the ink signal is at most around 50 µm deep, but wider regions of support are helpful in order to provide spatial context; the ink signal is present at similar levels using different incident CT energies, on both monochromatic and polychromatic beams; and spatial resolution of the CT images is critically important in order to capture the requisite level of detail.

The findings in this work are based on a specific pipeline, developed in order to pursue the first viable path to ink detection. Looking forward, at a minimum, each of these pipeline components has room for improvement that will reduce the overall error and will improve the ink detection results. More likely, other pipelines will be developed that restructure or combine the existing steps into simpler machine learning-based steps. Regardless of the specific technical implementations, I predict that future developments to this work will enable the complete recovery of texts from the rolled Herculaneum scrolls. The resulting images, produced noninvasively, will be more complete and in better condition than the physically unrolled fragments that are currently the subject of active textual scholarship.

REFERENCES

[1] D. Sider, *The library of the Villa dei Papiri at Herculaneum*. J. Paul Getty Museum, Getty Publications, 2005.

[2] K. Kleve, "Phoenix from the ashes: Lucretius and Ennius in Herculaneum," 1991.

[3] W. B. Seales and Y. Lin, "Digital restoration: Principles and approaches," in *Proceedings of DELOS/NSF Workshop on Multimedia in Digital Libraries*, 2003.

[4] W. Seales, "Reading the Invisible Library: A Retrospective," in *Modern Alchemy: New Technology for Museum Collections*, C. Brune and C. Foutch, Eds., Gilcrease Museum, Tulsa, OK, 2017.

[5] W. B. Seales and Y. Lin, "Digital restoration using volumetric scanning," in *Digital Libraries, 2004. Proceedings of the 2004 Joint ACM/IEEE Conference on*, IEEE, 2004, pp. 117–124.

[6] Y. Lin and W. B. Seales, "Opaque document imaging: Building images of inaccessible texts," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, IEEE, vol. 1, 2005, pp. 662–669.

[7] Y. Lin, "Physically-based digital restoration using volumetric scanning," Ph.D. dissertation, University of Kentucky, 2007, ISBN: 978-0-549-43331-6.

[8] D. Stromer *et al.*, "Non-destructive Digitization of Soiled Historical Chinese Bamboo Scrolls," in *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, IEEE, 2018, pp. 55–60.

[9] D. Stromer, V. Christlein, X. Huang, P. Zippert, T. Hausotte, and A. Maier, "Virtual cleaning and unwrapping of non-invasively digitized soiled bamboo scrolls," *Scientific reports*, vol. 9, no. 1, p. 2311, 2019.

[10] D. Stromer, V. Christlein, T. Schön, W. Holub, and A. Maier, "Browsing Through Closed Books: Evaluation of Preprocessing Methods for Page Extraction of a 3-D CT Book Volume," in *IOP Conference Series: Materials Science and Engineering*, IOP Publishing, vol. 229, 2017, p. 012 005.

[11] D. Stromer *et al.*, "Browsing through sealed historical manuscripts by using 3-D computed tomography with low-brilliance X-ray sources," *Scientific reports*, vol. 8, no. 1, pp. 1–10, 2018. DOI: 10.1038/s41598-018-33685-4. [Online]. Available: https://doi.org/10.1038/s41598-018-33685-4.

[12] D. Allegra, E. Ciliberto, P. Ciliberto, G. Petrillo, F. Stanco, and C. Trombatore, "X-ray computed tomography for virtually unrolling damaged papyri," *Applied Physics A*, vol. 122, no. 3, p. 256, 2016.

[13] D. Baum *et al.*, "Revealing hidden text in rolled and folded papyri," *Applied Physics A*, vol. 123, no. 3, p. 171, 2017.

[14] C. S. Parker, S. Parsons, J. Bandy, C. Chapman, F. Coppens, and W. B. Seales, "From invisibility to readability: recovering the ink of Herculaneum," *PloS one*, vol. 14, no. 5, e0215775, 2019.

[15] G. J. Tserevelakis, M. Tsagkaraki, P. Siozos, and G. Zacharakis, "Uncovering the hidden content of layered documents by means of photoacoustic imaging," *Strain*, e12289, 2018.

[16] R. Baumann, D. C. Porter, and W. B. Seales, "The use of micro-ct in the study of archaeological artifacts," in *9th International Conference on NDT of Art*, 2008, pp. 1–9. eprint: `https://www.ndt.net/article/art2008/papers/244Seales.pdf`.

[17] P. C. Dilley, C. Chapman, C. S. Parker, and W. B. Seales, "The X-Ray Micro-CT of a Full Parchment Codex to Recover Hidden Text: Morgan Library M.910, an Early Coptic Acts of the Apostles Manuscript," *Manuscript Studies: A Journal of the Schoenberg Institute for Manuscript Studies*, vol. 7, no. 1, pp. 162–174, 2022.

[18] O. Samko, Y.-K. Lai, D. Marshall, and P. L. Rosin, "Virtual unrolling and information recovery from scanned scrolled historical documents," *Pattern Recognition*, vol. 47, no. 1, pp. 248–259, 2014.

[19] C. Liu, P. L. Rosin, Y.-K. Lai, and W. Hu, "Robust Virtual Unrolling of Historical Parchment XMT Images," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 1914–1926, 2018.

[20] P. L. Rosin *et al.*, "Virtual Recovery of Content from X-Ray Micro-Tomography Scans of Damaged Historic Scrolls," *Scientific reports*, vol. 8, no. 1, p. 11 901, 2018.

[21] W. B. Seales, C. S. Parker, M. Segal, E. Tov, P. Shor, and Y. Porath, "From damage to discovery via virtual unwrapping: Reading the scroll from En-Gedi," *Science Advances*, vol. 2, no. 9, 2016. DOI: `10.1126/sciadv.1601247`. eprint: `http://advances.sciencemag.org/content/2/9/e1601247.full.pdf`. [Online]. Available: `http://advances.sciencemag.org/content/2/9/e1601247`.

[22] H.-E. Mahnke *et al.*, "Virtual unfolding of folded papyri," *Journal of Cultural Heritage*, vol. 41, pp. 264–269, 2020, ISSN: 1296-2074. DOI: `10.1016/j.culher.2019.07.007`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S1296207419301670`.

[23] J. Dambrogio *et al.*, "Unlocking history through automated virtual unfolding of sealed documents imaged by X-ray microtomography," *Nature Communications*, vol. 12, no. 1, p. 1184, Mar. 2021, ISSN: 2041-1723. DOI: `10.1038/s41467-021-21326-w`. [Online]. Available: `https://doi.org/10.1038/s41467-021-21326-w`.

[24] F. Albertin, "X-ray tomography for manuscripts," *Umanistica Digitale*, no. 12, pp. 39–64, 2022.

[25] G. H. Barfod, J. M. Larsen, A. Lichtenberger, and R. Raja, "Revealing text in a complexly rolled silver scroll from Jerash with computed tomography and advanced imaging software," *Scientific reports*, vol. 5, p. 17 765, 2015.

[26] D. Baum, F. Herter, J. M. Larsen, A. Lichtenberger, and R. Raja, "Revisiting the Jerash Silver Scroll: a new visual data analysis approach," *Digital Applications in Archaeology and Cultural Heritage*, e00186, 2021, ISSN: 2212-0548. DOI: https://doi.org/10.1016/j.daach.2021.e00186. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2212054821000151.

[27] B. Wilster-Hansen, D. C. Mannes, K. L. Holmqvist, K. Ødeby, and H. Kutzke, "Virtual unwrapping of the BISPEGATA amulet, a multiple folded medieval lead amulet, by using neutron tomography," *Archaeometry*, Nov. 2021. DOI: https://doi.org/10.1111/arcm.12734. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/arcm.12734.

[28] S. Stabile *et al.*, "A computational platform for the virtual unfolding of Herculaneum Papyri," *Scientific Reports*, vol. 11, no. 1, pp. 1–11, 2021.

[29] M. Segal, E. Tov, W. B. Seales, C. S. Parker, P. Shor, and Y. Porath, "An Early Leviticus Scroll from En-Gedi: Preliminary Publication," *Textus*, vol. 26, no. 1, pp. 29–58, 2016. DOI: 10.1163/2589255X-02601004.

[30] S. Parsons, K. Gessel, C. Parker, and W. Seales, "Deep Learning for More Expressive Virtual Unwrapping," in *Proceedings of the 25th International Conference on Cultural Heritage and New Technologies 2020.*, W. Börner, H. Rohland, C. Kral-Börner, and L. Karner, Eds., Nov. 2020, pp. 203–207. DOI: 10.11588/propylaeum.1045.c14501. [Online]. Available: https://doi.org/10.11588/propylaeum.1045.c14501.

[31] S. W. Booras and D. M. Chabries, "The Herculaneum Scrolls," in *PICS*, 2001, pp. 215–218.

[32] G. A. Ware, D. M. Chabries, R. W. Christiansen, and C. E. Martin, "Multispectral document enhancement: Ancient carbonized scrolls," in *IGARSS 2000. IEEE 2000 International Geoscience and Remote Sensing Symposium. Taking the Pulse of the Planet: The Role of Remote Sensing in Managing the Environment. Proceedings (Cat. No. 00CH37120)*, IEEE, vol. 6, 2000, pp. 2486–2488.

[33] A. Tournié *et al.*, "Ancient Greek text concealed on the back of unrolled papyrus revealed through shortwave-infrared hyperspectral imaging," *Science Advances*, vol. 5, no. 10, eaav8936, 2019.

[34] W. B. Seales and D. Delattre, "Virtual unrolling of carbonized Herculaneum scrolls: Research Status (2007–2012)," *Cronache Ercolanesi*, vol. 43, pp. 191–208, 2013.

[35] E. Brun *et al.*, "Revealing metallic ink in Herculaneum papyri," *Proceedings of the National Academy of Sciences*, vol. 113, no. 14, pp. 3751–3754, 2016.

[36] P. Tack *et al.*, "Tracking ink composition on Herculaneum papyrus scrolls: quantification and speciation of lead by X-ray based techniques and Monte Carlo simulations," *Scientific reports*, vol. 6, no. 1, pp. 1–7, 2016.

[37] O. Bonnerot, G. Del Mastro, J. Hammerstaedt, V. Mocella, and I. Rabin, "XRF ink analysis of some Herculaneum papyri," *Zeitschrift für Papyrologie und Epigraphik*, pp. 50–52, 2020.

[38] A. Gibson *et al.*, "An assessment of multimodal imaging of subsurface text in mummy cartonnage using surrogate papyrus phantoms," *Heritage Science*, vol. 6, no. 1, p. 7, 2018.

[39] T. Christiansen *et al.*, "Insights into the composition of ancient Egyptian red and black inks on papyri achieved by synchrotron-based microanalyses," *Proceedings of the National Academy of Sciences*, vol. 117, no. 45, pp. 27 825–27 835, 2020. DOI: 10.1073/pnas.2004534117. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2004534117. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.2004534117.

[40] P.-O. Autran *et al.*, "Revealing the Nature of Black Pigments Used on Ancient Egyptian Papyri from Champollion Collection," *Analytical Chemistry*, vol. 93, no. 2, pp. 1135–1142, 2021. DOI: 10.1021/acs.analchem.0c04178. eprint: https://doi.org/10.1021/acs.analchem.0c04178. [Online]. Available: https://doi.org/10.1021/acs.analchem.0c04178.

[41] B. Seales, "Lire sans détruire les papyrus carbonisés d'Herculanum," *Comptes rendus des séances de l'Académie des Inscriptions et Belles-Lettres*, vol. 153, no. 153, pp. 907–923, 2009. eprint: https://www.persee.fr/doc/crai_0065-0536_2009_num_153_2_92557.

[42] W. B. Seales, J. Griffioen, R. Baumann, and M. Field, "Analysis of Herculaneum papyri with x-ray computed tomography," in *International Conference on nondestructive investigations and microanalysis for the diagnostics and conservation of cultural and environmental heritage*, 2011.

[43] V. Mocella, E. Brun, C. Ferrero, and D. Delattre, "Revealing letters in rolled Herculaneum papyri by X-ray phase-contrast imaging," *Nature communications*, vol. 6, 2015.

[44] I. Bukreeva *et al.*, "Virtual unrolling and deciphering of Herculaneum papyri by X-ray phase-contrast tomography," *Scientific reports*, vol. 6, p. 27 227, 2016.

[45] I. Bukreeva, M. Alessandrelli, V. Formoso, G. Ranocchia, and A. Cedola, "Investigating Herculaneum papyri: An innovative 3D approach for the virtual unfolding of the rolls," *arXiv preprint arXiv:1706.09883*, 2017.

[46] S. Parsons, C. S. Parker, C. Chapman, M. Hayashida, and W. B. Seales, "EduceLab-Scrolls: Verifiable Recovery of Text from Herculaneum Papyri using X-ray CT," *arXiv preprint arXiv:2304.02084*, 2023.

[47] J. Canny, "A computational approach to edge detection," *IEEE Transactions on pattern analysis and machine intelligence*, pp. 679–698, 1986.

[48] P. Cignoni, M. Callieri, M. Corsini, M. Dellepiane, F. Ganovelli, and G. Ranzuglia, "MeshLab: an Open-Source Mesh Processing Tool," in *Eurographics Italian Chapter Conference*, V. Scarano, R. D. Chiara, and U. Erra, Eds., The Eurographics Association, 2008, ISBN: 978-3-905673-68-5.

[49] M. Kazhdan and H. Hoppe, "Screened poisson surface reconstruction," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 3, p. 29, 2013.

[50] P. Cignoni, C. Rocchini, and R. Scopigno, "Metro: measuring error on simplified surfaces," in *Computer Graphics Forum*, Blackwell Publishers, vol. 17, 1998, pp. 167–174.

[51] A. Sheffer and E. de Sturler, "Parameterization of faceted surfaces for meshing using angle-based flattening," *Engineering with computers*, vol. 17, no. 3, p. 326, 2001.

[52] B. Lévy, S. Petitjean, N. Ray, and J. Maillot, "Least squares conformal maps for automatic texture atlas generation," *ACM transactions on graphics (TOG)*, vol. 21, no. 3, pp. 362–371, 2002.

[53] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: A survey," *IEEE transactions on medical imaging*, vol. 32, no. 7, pp. 1153–1190, 2013.

[54] A. Sindel, A. Maier, and V. Christlein, "CraquelureNet: matching the crack structure in historical paintings for multi-modal image registration," in *2021 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2021, pp. 994–998.

[55] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[56] S. Behnel, R. Bradshaw, C. Citro, L. Dalcin, D. S. Seljebotn, and K. Smith, "Cython: The best of both worlds," *Computing in Science & Engineering*, vol. 13, no. 2, pp. 31–39, 2010.

[57] M. Abadi *et al.*, "Tensorflow: a system for large-scale machine learning.," in *Osdi*, Savannah, GA, USA, vol. 16, 2016, pp. 265–283.

[58] K. Fukushima, "Neocognitron: A hierarchical neural network capable of visual pattern recognition," *Neural networks*, vol. 1, no. 2, pp. 119–130, 1988.

[59] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[60] I. Cohen *et al.*, "Pearson correlation coefficient," *Noise reduction in speech processing*, pp. 1–4, 2009.

[61] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE intelligent systems*, vol. 24, no. 2, pp. 8–12, 2009.

[62] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 843–852.

[63] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[64] J. Schindelin *et al.*, "Fiji: an open-source platform for biological-image analysis," *Nature methods*, vol. 9, no. 7, pp. 676–682, 2012.

[65] E. D. Foster and A. Deardorff, "Open science framework (OSF)," *Journal of the Medical Library Association: JMLA*, vol. 105, no. 2, p. 203, 2017.

[66] B. Athie Teruel, S. Chapman, S. Parsons, C. S. Parker, and W. B. Seales, "Quick Segment: A Coarse-to-Fine Segmentation Method for Virtual Unwrapping," 2023.

[67] M. Januszewski and V. Jain, "Segmentation-Enhanced CycleGAN," *bioRxiv*, 2019. DOI: 10.1101/548081. eprint: https://www.biorxiv.org/content/early/2019/02/13/548081.full.pdf. [Online]. Available: https://www.biorxiv.org/content/early/2019/02/13/548081.

[68] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[69] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[70] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, Springer, 2015, pp. 234–241.

[71] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, Springer, 2020, pp. 319–345.

[72] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," *arXiv preprint arXiv:1412.3474*, 2014.

[73] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[74] Y. Jiang *et al.*, "AligNeRF: High-Fidelity Neural Radiance Fields via Alignment-Aware Training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 46–55.

[75] J. H. Brusuelas, "Scholarly Editing and AI: Machine Predicted Text and Herculaneum Papyri," *magazén*, 2021.

[76] L. T. Niklason, L. E. Niklason, and D. B. Kopans, *Tomosynthesis system for breast imaging*, US Patent 5,872,828, Feb. 1999.

[77] A. Miles *et al.*, *zarr-developers/zarr-python: v2.4.0*, version v2.4.0, Jan. 2020. DOI: 10.5281/zenodo.3773450. [Online]. Available: https://doi.org/10.5281/zenodo.3773450.

[78] J. Maitin-Shepard and L. Leavitt. "TensorStore for High-Performance, Scalable Array Storage." (2022), [Online]. Available: https://ai.googleblog.com/2022/09/tensorstore-for-high-performance.html (visited on 07/05/2023).

[79] J. Maitin-Shepard *et al.*, "Neuroglancer," *github.com/google/neuroglancer, Retrieved*, pp. 04–30, 2021.

[80] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[81] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding with unsupervised learning," 2018.

[82] S. M. McKinney *et al.*, "International evaluation of an AI system for breast cancer screening," *Nature*, vol. 577, no. 7788, pp. 89–94, 2020.

[83] E. D. Pisano *et al.*, "Diagnostic performance of digital versus film mammography for breast-cancer screening," *New England Journal of Medicine*, vol. 353, no. 17, pp. 1773–1783, 2005.

[84] D. M. Ikeda, R. L. Birdwell, K. F. O'Shaughnessy, E. A. Sickles, and R. J. Brenner, "Computer-aided detection output on 172 subtle findings on normal mammograms previously obtained in women with breast cancer detected at follow-up screening mammography," *Radiology*, vol. 230, no. 3, pp. 811–819, 2004.

VITA

Stephen Parsons
Lexington, KY

Education

B.S. in Computer Science, University of Kentucky. May, 2016.

B.S. in International Studies, University of Kentucky. May, 2016.

Professional positions held

2018 – 2019: Staff researcher, Digital Restoration Initiative, University of Kentucky.

2016 – 2017: Product Manager, Microsoft.

2015: Associate Product Manager intern, Google.

Scholastic and professional honors

2020 – 2023: National Science Foundation Graduate Research Fellowship

Publications

Parsons, S., Parker, C. S., Chapman, C., Hayashida, M., & Seales, W. B. (2023). EduceLab-Scrolls: Verifiable Recovery of Text from Herculaneum Papyri using X-ray CT. *arXiv preprint arXiv:2304.02084*.

Parker, C. S., Parsons, S., Bandy, J., Chapman, C., Coppens, F., & Seales, W. B. (2019). From invisibility to readability: recovering the ink of Herculaneum. *PloS one, 14*(5), e0215775.

Parsons, S., Chappell, J., Parker, C. S., & Seales, W. B. (2021, March). Machine learning infrastructure on the frontier of virtual unwrapping. In *International Symposium on Grids & Clouds 2021. 22-26 March 2021. Academia Sinica Computing Centre (ASGC)* (Vol. 15).

Parsons, S., Gessel, K., Parker, C., & Seales, W. (2020). Deep Learning for More Expressive Virtual Unwrapping. In *Proceedings of the 25th International Conference on Cultural Heritage and New Technologies* (pp. 203-207).

Chapman, C., Parker, S., Parsons, S., & Seales, W. B. (2020, December). Using METS to Express Digital Provenance for Complex Digital Objects. In *Research Conference on Metadata and Semantics Research* (pp. 143-154). Cham: Springer International Publishing.

Chapman, C. Y., Parker, C. S., Bertelsman, A., Gessel, K., Hatch, H., Seevers, K., Brusuelas, J., Parsons, S., & Seales, W. B. (2021). The Digital Compilation and Restoration of Herculaneum Fragment P. Herc. 118. *Manuscript Studies: A Journal of the Schoenberg Institute for Manuscript Studies, 6*(1), 1-32.

Parsons, S., Parker, C. S., & Seales, W. B. (2017). The St. Chad Gospels: Diachronic Manuscript Registration and Visualization. *Manuscript Studies: A Journal of the Schoenberg Institute for Manuscript Studies, 2*(2), 483-498.

Parsons, S., & Salehi, S. A. (2022, October). Probability Distribution Calculations with Stochastic Circuits. In *2022 56th Asilomar Conference on Signals, Systems, and Computers* (pp. 1-5). IEEE.