Curating Very Large Biomedical Image Datasets For Librarian-In-The-Loop Deep Learning

The Center for Digital Research & Scholarship (CDRS) at Virginia Tech Libraries requests $149,216 from IMLS for a two-year National Leadership Planning Grant to prototype a data curation pipeline for very large biomedical images. Using image data generated from the Focused Ion Beam Scanning Electron Microscopy Collaboration Core (F-SCC) at Yale University School of Medicine, we will explore the problem space of AI-assisted human-in-the-loop segmentation of nanoscale biomedical images. Librarians will work closely with domain scientists to identify and clarify key challenges by more thoroughly replicating, comparing, and extending two complementary research and other related projects. This project will lay the foundation for the implementation project where library technologists and informationists will be able to contribute key expertise to cutting-edge science.

1. Project Justification

1.1 Program Goal and Objectives

This project broadly addresses Goal 3 of the National Leadership Planning Grant (NLG) program, which seeks to enhance access to, preserve, and disseminate information and collections through digital technologies. Specifically, the project focuses on Objective 3.2, which aims to explore innovative approaches to data management, including data curation, based on collaborative efforts between librarians and researchers.

1.2 Statement of Broad Need

The broadly significant need addressed by this project arises from two different sides.

On one hand, academic and research libraries need to play a more active role in cutting edge science in order to gain recognition as a qualified research partner (Lacchia, 2021). Many decades ago the established library service model has already shown signs of disconnect from the patron's fast-changing information needs. Libraries have been eager to explore new roles, new models of operation, and new growth areas (Jaguszewski & Williams, 2013; Gwyer, 2015; Kamposiori, 2017; Meier, 2016; Lewis, 2016; T. Hickerson & Brosz, 2019; Evidence Base, 2021; Ducas et al., 2020; Zhang et al., 2021; Fernández-Marcial & González-Solar, 2021; H. T. Hickerson et al., 2022). Library administrators are under increasing pressure to justify the spending and to articulate more engaging and relevant value propositions (Cox, 2018; Lewis, 2016; Murray & Ireland, 2018; McGinnis et al., 2022). A common vision emerged about 10 years ago (Association of Research Libraries, 2010, 2016), claiming that by 2033 the library should "have shifted from its role as a knowledge service provider … to become a collaborative partner".

Only 10 years left, the ARL 2033 vision is still far from reality. It requires librarians to significantly expedite and deepen research collaborations and generate higher impact (Evidence Base, 2021). The library's traditional collaboration role as a research **supporter** must therefore be upgraded to a **specialist** who directly contributes critical and sought-after resources, skills, and expertise, or even to an idea-generating research **leader** (Robinson-Garcia et al., 2020). Because a regular service provider is

rarely considered an equal partner in research (Bright, 2018; Weng & Murray, 2020), the library's relevance and embeddedness (Dewey, 2004) should be measured less by our physical distance from the collaborators (Carroll et al., 2020; Drewes & Hoffman, 2010), but more by how closely related our substantive research contributions are to the core research idea and its realization. Likewise, the impact librarians can generate is closely related to the impact of the discoveries we facilitate, thus the urgent need for us to engage in cutting-edge science.

On the other hand, the cutting edge science also urgently needs help from informationists and technologists to realize its potential. Data-intensive scientific discovery, also known as the fourth paradigm (Hey, 2009), has now been broadly accepted as an important driver for innovation, yet we cannot expect every neuroscientist and astrophysicist to also become an expert in big data, deep learning, and cloud computing. Task specialization associated with modern science naturally requires more cross-disciplinary collaboration, but currently the most qualified experts in big data, deep learning, and cloud computing are often too deeply engaged in their own domain challenges to have the bandwidth for much broader collaboration. This opens up a narrow window for qualified librarians to play the **specialist** role in scientific collaboration. Fortunately, major tools and technologies used in data-intensive science are widely available and also broadly overlap, at least conceptually, with those used in library big data management. Some of the skills required to build up the groundwork, e.g., labeling, categorization, pattern recognition, and operating image manipulation tools, are not as insurmountable as many would have imagined. It is therefore feasible for adept and versatile library technologists and informationists to cross over into new domains.

Take as an example the data challenge associated with very large biomedical images. Recent breakthroughs (Xu et al., 2017, 2020) in Focused Ion Beam Scanning Electron Microscopy (FIB-SEM) technology have made it possible to reliably acquire nanometer scale 3D imaging from sizable volumetric biomedical samples, each resulting in tens to hundreds of terabytes of raw image data. *Nature* recently named enhanced FIB-SEM one of "Seven technologies to watch in 2023" (Eisenstein, 2023), claiming it is a "quiet revolution" with the potential impact to be "on the same order of magnitude as the effort to map the first human genome". Enhanced FIB-SEM has already been used to generate landmark datasets for neuroscience (Anthes, 2021; Scheffer et al., 2020) and cell biology (Xu et al., 2021) research. C Shan Xu, a primary inventor of enhanced FIB-SEM, is now bringing the invention to enable discoveries in translational and clinical research and has promptly generated stunning images showing how a cancer cell evades attacks from the T cell, illustrated on the cover of *Science* (Andrews, 2022; Gregor et al., 2022; Ritter et al., 2022).

Similar to many other breakthroughs in instrumentation, once the hardware is built and running, FIB-SEM imaging's bottleneck shifts to data, more specifically from data collection and storage to data curation for the primary purpose of extracting insights and knowledge from data. In contrast to other existing biomedical imaging, enhanced FIB-SEM poses an even more formidable challenge on efficiency, timeliness, replicability, and reusability. For example, it has been estimated that FIB-SEM images taken from a single cell may take up to 60 person-years to annotate manually (Heinrich et al., 2021). The very large data volume has rendered comprehensive close reading and manual image annotation impractical. It also takes away from scientists the ability to easily feel the data, because at the resolution this high, all they can see are details without an overview. To make sense of these images, researchers must resort to

artificial intelligence by applying supervised deep learning techniques to train a model from a few small slices of manually annotated images, then use the model to run predictions against unannotated images. Although deep learning algorithms and tools are widely available and routinely used to process other biomedical images, prior successes do not easily transcend to enhanced FIB-SEM. The COSEM project, the first of its kind, only recently published its FIB-SEM images and code, and claimed they have achieved high accuracy and performance in segmenting those images. However, our initial tests showed that the COSEM models could not accurately predict cell organelles from newly acquired FIB-SEM images yet we don't know why. At least for now, it appears to the domain scientists that AI based FIB-SEM image segmentation remains a form of art that depends too heavily on intuitions and serendipity to qualify as skilled trade, let alone predictable technology or even accurate science. But they cannot wait. The newly built enhanced FIB-SEMs at Yale School of Medicine are pumping out image data at a stunning speed, and scientists urgently need help to make sense of them. They are willing to work with anyone who can help, librarians included.

The pressing needs from both sides and the associated sense of urgency have created a perfect moment to loop librarians into cutting-edge science. However, if the history of the library community's encounters with digital publishing and the web is our guide, this moment can evaporate in no time if we don't seize it immediately.

1.3 Target Group and Ultimate Beneficiaries

This project primarily targets scientists who currently develop FIB-SEM technologies and/or use the Focused Ion Beam Scanning Electron Microscopy Collaboration Core (F-SCC) facility at Yale School of Medicine (YSM) to conduct research, starting with Professor C Shan Xu and Professor Angelique Bordey. If we make enough progress, the target group may expand to all clients of F-SCC regardless of their affiliations, and even further expand to scientists using FIB-SEM facilities elsewhere. Two co-PIs working at Virginia Tech Libraries also directly benefit from this project by gaining access to not only the most advanced biomedical imaging facility and its data but more importantly, the large group of scientists using the facility to conduct cutting edge science. Our graduate students and collaborators at Virginia Tech and beyond also directly benefit from such access.

Ultimate beneficiaries of this project include 1) Virginia Tech Libraries, which through our work can claim to be a more qualified research partner; 2) other projects and groups handling very large images and dataset, who can learn from the discoveries and outcomes of this project; 3) the broader academic and research library community, if they follow our example and similarly build their own expertise and deep partnership; 4) universities and institutes hosting these libraries, who will benefit from a more engaging, creative, and research-oriented library; 5) patients who will benefit from the biomedical discoveries originated from the expedited processing of nanoscale imaging; and 6) the general public and the whole society, who will benefit from accelerated innovation and discovery.

1.4 Related Work and Preliminary Findings

Biomedical imaging is a very crowded field. Deep learning based image segmentation has already been widely applied to magnetic resonance imaging (MRI) and other medical images and is one of the more

mature AI applications. U-Net (Ronneberger et al., 2015a, 2015b) and 3D U-Net (Çiçek et al., 2016), the most recognized and widely used deep learning algorithms for biomedical image segmentation, were developed 7 to 8 years ago. They still maintain a dominant position in the field, largely due to their maturity and superior performance. Software that implements these algorithms is widely available, e.g., via pytorch and tensorflow. Project MONAI integrates many medical AI tools into a suite, on top of which various medical image processing pipelines can be built (Cardoso et al., 2022). However, our initial attempt to use MONAI tools to label a tiny FIB-SEM crop did not work well. Primarily developed for medical use, MONAI currently does not recognize nanoscale cell structures such as organelles.

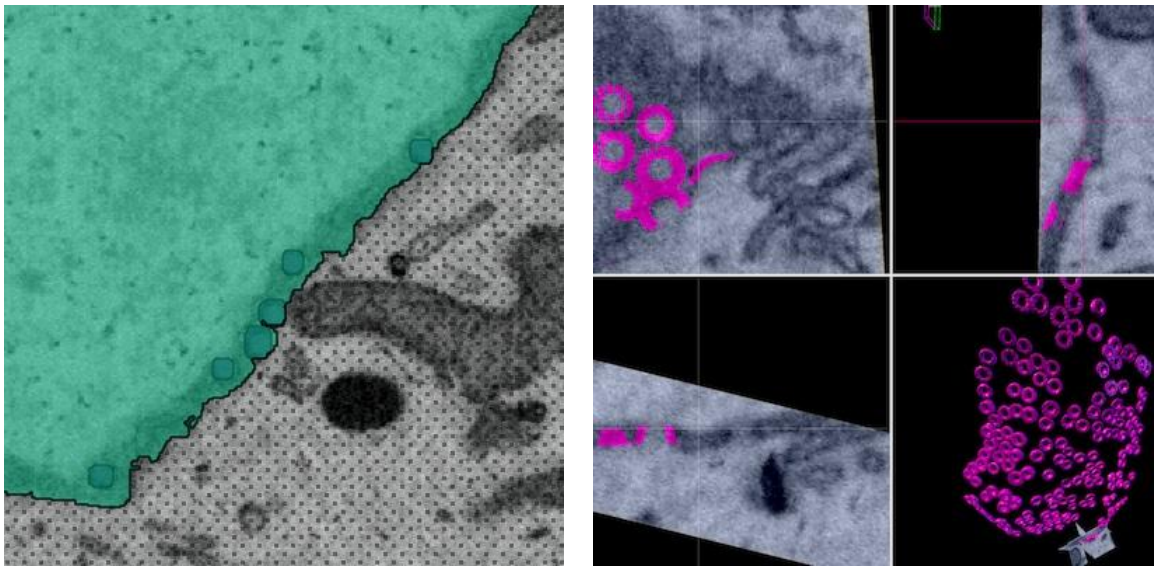
In theory tools used to handle other images should remain effective for FIB-SEM imaging, but the latter poses unique challenges largely due to its sheer size and complexity. Implementing a FIB-SEM segmentation pipeline using tools like MONAI may seem trivial until the raw data arrives in hard drives: many commonly used imaging tools won't even be able to open these tiff images, let alone do anything about them. They are so big that "out of memory" errors are encountered on a daily basis, even on a supercomputer. Very large FIB-SEM datasets, therefore, present an archetypal test case on how to best orchestrate scientists, data curators, cyberinfrastructure, software, and deep learning algorithms to achieve best data-to-insight performance.

Enhanced FIB-SEM was initially used by neuroscientists to understand the brain. The intertwining of the neurons made image processing very challenging and eventually compelled the scientists to collaborate with Google Research to develop proprietary solutions for "connectomics"(Scheffer et al., 2020). When the FIB-SEM was applied to cell biology research, data processing posed a different set of problems. The COSEM project, also known as OpenOrganelle (Heinrich et al., 2021), is to date the only successful attempt to segment all cell organelles from FIB-SEM images, and has been highlighted in the *Nature* article as a major achievement (Eisenstein, 2023). The COSEM team has developed many tools and utilities to handle very large image sets reformatted in an obscure data format called N5 (stands for not HDF5) and also made their code, data, and documentation openly available. COSEM abides by all the principles and best practices set forth by the FAIR principles (Wilkinson et al., 2016), but that does not make it any easier to reproduce or reuse. Despite the fact that COSEM data has been openly available for almost 2 years, insiders are not aware of any other independent attempt to reproduce its results. COSEM insiders also described a seemingly chaotic real-world data flow, drastically different from the established data lifecycle (Higgins, 2008). This made the librarian's imagination of orderly conducted science sound naive. Indeed, scientists may directly jump from any one stage of the data curation lifecycle to any other stage based on needs and circumstances.

Our preliminary attempt to reproduce the COSEM paper on clusters at Virginia Tech's Advanced Research Computing (ARC) took more than 6 months to reach a point of no errors, but we are still not sure if we have done everything correctly. According to the COSEM paper, one round of training could take at least 7 to 16 months to complete on their GPU powered computer cluster, and they seemed to have exclusive access to that cluster. Although our NVIDIA A100 powered cluster at ARC is faster than what COSEM originally used, ours is a shared one that we must yield to other users at least every 3 days. We also used VT Libraries' smaller GPU cluster for debugging and testing. Instead of trying to run the training from scratch, we decided to use their published checkpoints to run predictions on images cropped from COSEM data not previously used for training. The results were not very good, suggesting the

published checkpoints may lead to overfitting. We then ran predictions on FIB-SEM images acquired elsewhere, the results were even worse to the point of unrecognizable. This seems to defeat the purpose of the COSEM project. If previously trained checkpoints cannot be used to correctly segment new images, why would anyone want to run training for such a long time? The training performance seems too poor for practical use.

In retrospect, COSEM's overall objectives could be too ambitious. It may have suffered the same naivety of the DCC Data Lifecycle Model by assuming domain scientists follow a premeditated plan to do research and would give data scientists an extended period of time to process the full set of data to produce the best overall results in one attempt, e.g., predict every single cell organelle in high confidence, then never need to revisit the raw data again. The data pipeline as published did not seem to accommodate trial-and-error, considering the training had to be such an expensive endeavor. However, what really happened during our initial discussion with Professor Bordey, a neuroscientist at Yale School of Medicine studying epileptic seizure, was that her research started from several straightforward yet specific questions that had to be answered first. She was only interested in a few nuclei, and only needed to know the total number and the density of the nuclear pores (see Figure 1). Until she knows the answers to those questions, she is not sure what her next set of questions would be. And if the already imaged sample does not help to answer her questions, she may have to run another sample. This trial-and-error cycle makes the collaboration and the data pipeline a lot more interactive and iterative than the COSEM paper presented.



(a) (b)

Figure 1. Labeling Nuclear Pores of a Mouse Brain Neuron (a) Label one image layer at a time using Arivis Vision4D, an expensive yet clunky cloud- and web-based commercial tool (b) 3D labeling using Paintera, an open source tool

We thought of modifying COSEM code to run training on a single cell organelle label (e.g., the outside of the pore) at a time instead of the default setting of training for all organelle labels together, but that proved to be more troublesome than we had time for. Our attempts on COSEM training also turned out to be very

time-consuming and converged very slowly. In the end we developed our own code based on Tensorflow for training and prediction, then labeled a handful of thin slices cropped from different sides of a nucleus as the ground truth. We ran the training for a few days at a time, then added some new ground truth data, trained some more, until the evaluation results no longer significantly improved with much more training iterations. We then ran predictions on the whole nucleus to identify the pores, see Figure 2 and this video. The prediction looks promising and has correctly identified many pores on the nuclear envelope, but there are large, contiguous areas that should have been densely covered with pores (we did not know this until consulting Professor Bordey) but were not identified (false negatives). Then there are quite some pores incorrectly identified off the envelope (false positives). We then noticed that false negatives are mostly concentrated on one orientation of the sample, causing us to posit that it might be the shading effects of the microscopy's backlight, even if we know very little about the physics and the mechanism of FIB-SEM. But if our guesses were correct, and if all image processing techniques have been exhausted, a possible remedy for the problem could be to rerun the imaging using a different sample but with different backlight settings, which could better capture nuclear pores. We reached out to Prof. Xu, who did not support our conjecture, but nevertheless suggested ways to verify. In this way, we started to engage in scientific discussions librarians would otherwise not be involved in, and we truly sensed the comradery of common curiosity in searching for answers.

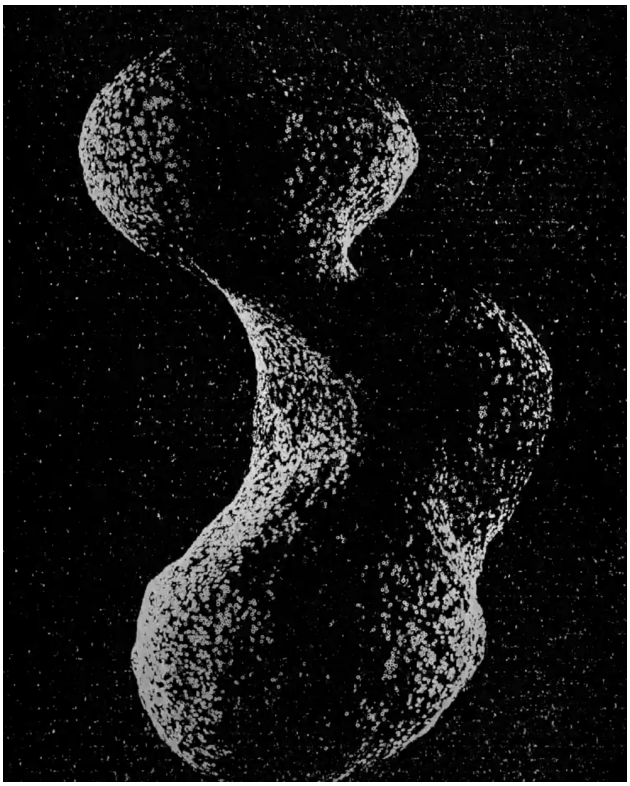


Figure 2. Predicting Nuclear Pores of a Mouse Brain Neuron, also see here for the video of its 3D rendition

In Oct 2022, our collaborators at Yale brought to our attention a BioRXiv preprint (Gallusser, Maltese, Caprio, et al., 2022), published two months later in JCB (Gallusser, Maltese, Di Caprio, et al., 2022), where the authors took a similar approach as we did and reported a significant performance boost against

COSEM, which we had also experienced. They also trained for one label at a time, then applied the so-called human-in-the-loop techniques (Holzinger, 2016), which applied human intuitions to add new ground truth data to augment previously trained models. While the paper's conclusions seemed reaffirming and encouraging, we still hesitated to draw too broad a conclusion beyond its context, fearing we might have not given the COSEM paper a fair hearing. Afterall, deep learning involves too many variables and moving parts that an apple-to-apple comparison is often difficult to make.

At this stage, we have tried many different things. But we did not have high confidence that our experiments were thorough or our results were reliable, but they did have generated more questions than answers. We need a more thorough planning project to focus our future work on directions that can make the most impact. Although our ultimate goal is to develop a well coordinated pipeline for semi-automatic segmentation of very large FIB-SEM images, we first need to explore our problem space more thoroughly in order to gain a good sense of what is possibl and what is not.

This proposal also draws insights from our prior IMLS funded project curating very large research datasets. Our past experience has shown that 1) data curators/librarians should be deployed in the big data pipeline as early as possible, even at the stage of physically acquiring data (Xie et al., 2015), because knowledge in data acquisition often affords pertinent opportunities to optimize the data pipeline, as illustrated by our previous conjecture on microscopy backlight . 2) Data curation should be driven primarily by data use and reuse (Xie et al., 2015), which closely aligns librarians/data curators with domain scientists. Long-term preservation and AI research and development are better performed as a side effect of data use and reuse. 3) The efficiency, cost, and performance of extracting insights from data are often the critical success factors for data curation and are closely associated with both the data format (Wang & Xie, 2020) and the cyberinfrastructure options and choices (Xie & Fox, 2017). Trial-and-error, experimenting and benchmarking are often the more effective way to achieve balanced results, therefore this planning proposal.

2. Project Work Plan

Our ultimate goal is to develop a comprehensive pipeline for semi-automated, human-in-the-loop image segmentation of FIB-SEM images. This project serves as its planning phase, in which critical aspects of the pipeline are clearly identified and potential contradictions surfaced from past research can be clarified. The project work revolves around five components of the pipeline: data collection, annotation, training, augmentation, and performance.

2.1 Project Activities

Project Task 1 Data Collection (Lead: PI Xie): We will contract with F-SCC at YSM to acquire a new FIB-SEM dataset from scratch. Today, most FIB-SEM images openly available were generated 3-6 years ago or even earlier at HHMI Janelia Research Campus by C Shan Xu on the so-called FIB-SEM 1.0 machines. Xu's team is now building a new generation of FIB-SEM at YSM, potentially capturing images with different features. As discussed before regarding the nuclear pore false negatives, without in-depth knowledge and experience with the machines and the sample preparation methods, we image processors often lack the necessary intuition to expedite the training process. The remedy is to be fully engaged in

the data collection, at least once. In addition, our preliminary work indicates that models trained from COSEM do not seem to accurately predict organelles from images captured in a different batch, and there may exist various artifacts such as backlight and shades that could interfere with AI-assisted segmentation. To make a fair comparison, it is often necessary to create a new, reference dataset. The new dataset collected from F-SCC will serve as the reference data. The F-SCC at YSM will closely coordinate with the VT team in establishing the scientific goals for data acquisition and preparing data curators with sufficient background knowledge in biology, physics, and chemistry to understand the process of sample preparation, milling, and imaging.

Project Task 2 Annotation (Lead: PI Xie and co-PI Chen) aims to establish a reliable and efficient image annotation environment capable of handling high resolution 3D images to create ground truth training datasets. We will evaluate Paintera and MONAI on 1) VT Libraries' GPU cluster and 2) Amazon Web Service (AWS). We exclude many commercial products, e.g., Arivis Vision4D, not only because they can be very expensive but also because they tend to lock in user data. In other words, if we annotate images in their environment, we are also expected to run training in their environment, which is not ideal. Fortunately both Paintera and MONAI are open source software. MONAI may also be run on ARC's Open OnDemand service, which is also widely available on many other supercomputer centers. We will also teach students to use them and collect their reviews on usability.

Project Task 3 Training (Lead: PI Xie and co-PI Chen) aims to experiment with 3D UNet based FIB-SEM image segmentation. We will continue our attempts to replicate and then interrogate the COSEM paper (Heinrich et al., 2021) and its associated software, but using new data collected in Task 1 and then annotated in Task 2. We intend to verify if our prior doubts on COSEM rationale is justified. If yes, if our own Tensorflow based implementation compares favorably against the PyTorch based implementation (Gallusser, Maltese, Di Caprio, et al., 2022). We will use ARC's Nvidia DGX A100 cluster and/or the Amazon cloud for this work. Both platforms are widely available at many research-intensive university campuses for various deep learning projects.

Project Task 4 Augmentation (Lead: PI Xie and co-PI Chen) aims to experiment with human-in-the-loop enhancements by replicating various image pre-processing and transfer learning procedures described in (Gallusser, Maltese, Di Caprio, et al., 2022) using the newly acquired data. An already-trained model for one label will be fine-tuned with additional annotations to speed up the training.

Pre-processing can significantly impact the model's performance. We will experiment with various pre-processing techniques to determine their impact on the model's performance, including image denoising, normalization, contrast enhancement, and histogram equalization. The model's results will be compared with and without pre-processing to determine the impact of each pre-processing technique on the model's performance.

Data augmentation is another technique that can improve the performance of deep learning models. This project will use data augmentation techniques such as random rotations, flips, and translations to increase the number of images available for training. The additional images generated through data augmentation will help prevent overfitting and improve the model's generalization.

Project Task 5 Benchmarking (Lead: PI Xie and co-PI Chen) Using the same dataset, image format, and computing facility, we will parse, verify, and compare the previously published performance claims. It is important to repeat the experiments on randomized sample regions to eliminate the survivorship bias in the benchmarking. We will also evaluate the performance differences between different image labels (e.g., mitochondria vs. golgi vs. ER, etc.), cyberinfrastructure choices (public cloud vs. institutionally shared GPU cluster vs. designated small GPU cluster), image formats (tiff vs. H5 vs. Zarr vs. Parquet), and various curator intervention points. When benchmarking on the public cloud, we will gather cost information.

2.2 Time, Financial, Personnel, and Other Resources

Time: All project activities will be spread over a 2-year period. For the detailed project timeline please refer to the Schedule of Completion.

Budget: Our request to IMLS include the following direct cost breakdowns: $53,917 wages and fringe benefits (including $36,231 for 1.5 GRA-year counted as student support), $25,000 for acquiring one FIB-SEM dataset at F-SCC at YSM, and $22,948 tuition remission for the GRA which is also counted as student support. After applying Virginia Tech's federally negotiated indirect cost rate of 60%, our total IMLS request is $149,216.

Personnel: Project staff include PI Zhiwu Xie, who will lead and manage the grant, and Co-PI Yinlin Chen. The project will also hire a graduate research assistant for a period of one and a half years. The research assistant is expected to come from the VT Computer Science Department. Under the direction of PI and co-PI, the GRA will perform duties such as data collection, data preprocessing, annotation, development and implementation of supervision and machine learning workflows, algorithm tuning, and assistance in writing reports and research publications.

An advisory committee is established to advise this project. Members will participate in quarterly one-hour meetings with the project team throughout the performance period. These meetings will allow the VT project team to obtain feedback on project progress, discuss challenges and potential solutions, and review preliminary research findings, white paper drafts, and publication drafts. Advisory committee members include (alphabetically by last name):

- Angelique Bordey, Rothberg Professor of Neurosurgery; Co Vice Chair of Research, Neurosurgery, Yale School of Medicine
- C. Shan Xu, Harvey and Kate Cushing Professor of Cellular & Molecular Physiology, Yale School of Medicine
- Edward A. Fox, Professor, Department of Computer Science, Virginia Tech
- Martin Klein, Scientist & Team Lead of Research & Prototyping, Los Alamos National Laboratory Research Library
- Nicholas Polys, Director of Visualization, Advanced Research Computing, Virginia Tech

Major Computing Resources: DGX Nodes and A100 Node at TinkerCliffs cluster and Open OnDemand, both from VT Advanced Research Computing; Amazon Web Services

Instrumentation Facility: Focused Ion Beam Scanning Electron Microscopy Collaboration Core (F-SCC) at Yale University School of Medicine

2.3 Dissemination Plan

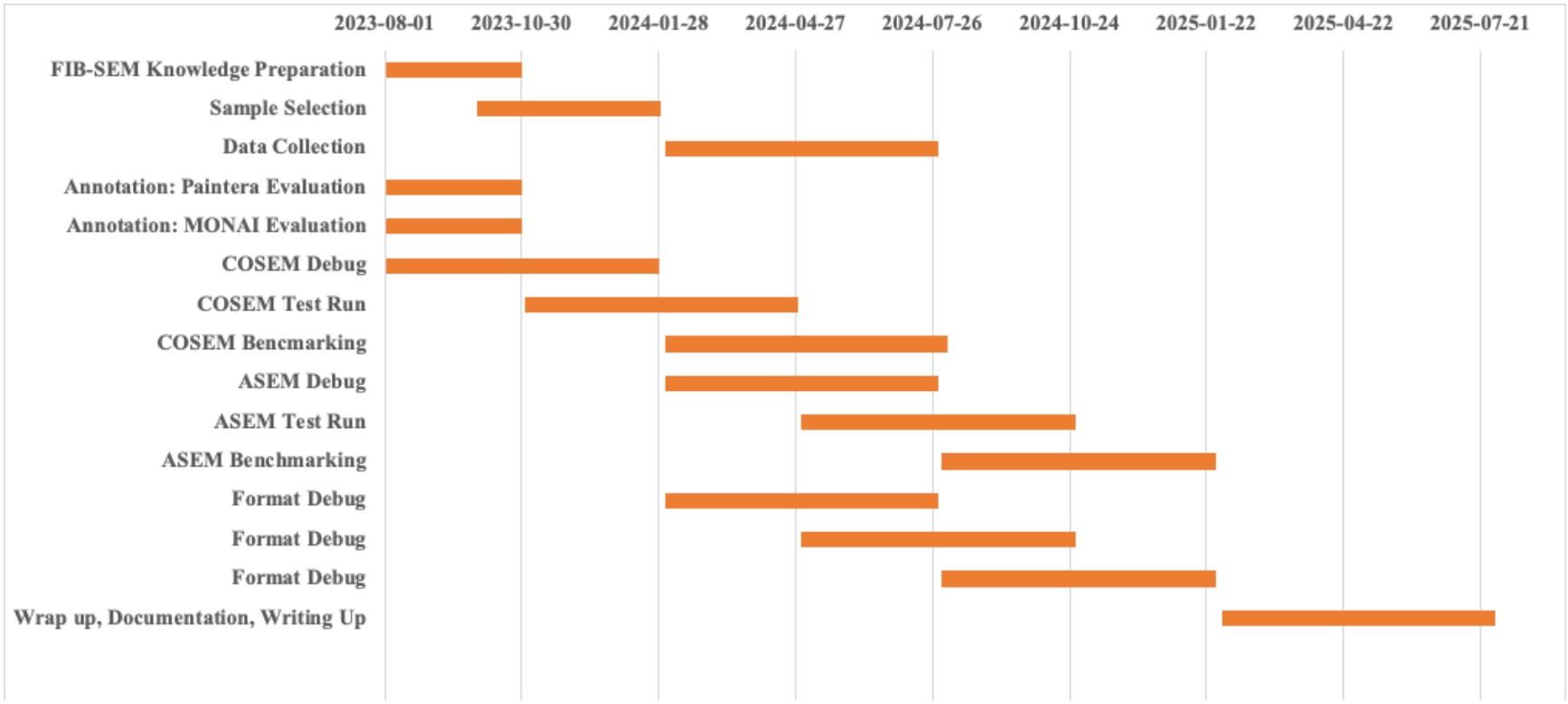
We plan to share our findings widely by publishing in academic journals and presenting at top conferences in library science, information retrieval, artificial intelligence, and biomed image processing. Potential venues include but are not limited to: JCDL, CNI, and/or IEEE Big Data. We will share all our documentation and software code on github as well as VT's institutional data repository (https://data.lib.vt.edu/).

3. Project Results

Curating large datasets is essential in the medical field for advancing research, developing treatments, and improving patient outcomes. This project is at the forefront of the biomedical field's digital transformation by prototyping a pioneering pipeline for curator-in-the-loop deep learning to curate large biomedical datasets. Through this project, we aim to equip library and archive professionals with practical skills in data-intensive science, image processing, machine learning, and applied Artificial Intelligence, enabling them to form deeper alliances with researchers and accelerate the transformation of libraries and archives from knowledge service provider to collaborative research partner. The lessons learned from this project have broad applications for handling vast datasets in other fields, such as engineering, physics, and environmental science. This project's efforts will establish a foundation for future advancements in data curation and analytics across all scientific research and innovation areas, promoting collaboration between data-intensive stakeholders and advancing scientific research and innovation.

The project's emphasis on practical skills and cutting-edge technologies will have far-reaching implications, enabling researchers in various fields to curate and process large research datasets more efficiently. The project's deliverables, including a curated FIB-SEM image dataset, pipeline prototypes, a comprehensive white paper, a publicly available documentation and software code repository on Github, and technical support and consultation, will establish a foundation for future advancements in data curation, analytics, and preservation.

Schedule of Completion



Curating Very Large Biomedical Image Datasets For Librarian-In-The-Loop Deep Learning
Digital Products Plan

Type

This planning project will generate multiple datasets containing annotated 3D medical images created by human experts and machines. Moreover, prototype software tools will be developed to aid in creating and analyzing these images. Publications, presentations, and white papers will be prepared at various project stages to share progress and preliminary results.

Availability

The results of this project will be made widely available through multiple channels to ensure availability. These channels include depositing datasets, whitepapers, publications, and presentations in the Virginia Tech (VT) Scholarly Repository (https://vtechworks.lib.vt.edu/) and Virginia Tech Data Repository (https://data.lib.vt.edu/), hosted by the VT Libraries (https://lib.vt.edu/). The project's publications and datasets will be displayed on a dedicated website created and maintained by the Center for Digital Research & Scholarship at VT Libraries.

This project website will serve as a comprehensive information source about the research, offering links to datasets, software tools, and publications generated throughout the study. A GitHub repository will be used to share software, scripts, and documentation produced during the project's planning phase. This repository will facilitate interaction between the project team and other parties interested in applying the software, scripts, and workflows in their own research. The VT Libraries has a history of effectively providing access to software tools and applications through GitHub (https://github.com/vtul).

Access

All publications and presentations will be made available through the VT Scholarly Repository. Datasets will be deposited in the Virginia Tech Data Repository. Furthermore, all software tools will be accessible via the project's GitHub page, which will feature repositories created for software, scripts, and documentation during the planning project.

Rights will be assigned to this planning project output with a flexible reuse license based on the resource type. Software, scripts, and documentation will be a GNU open-source General Public License (GPL). Publications and presentations, when feasible, will be made accessible through a Creative Commons License, such as CC-BY. Datasets will be made available under a CC0 Public Domain Dedication license.

Sustainability

These resources become a permanent part of the VT Libraries' digital assets by depositing datasets, publications, and presentations into the VT Scholarly Repository and Virginia Tech Data Repository. Although long-term access to these materials is expected to be infinite, the project team commits to providing access to all products and data from this project for at least five years after the grant period ends. Software, scripts, and documentation will be available on GitHub as long as it remains a viable and free platform for accessing software and code. If GitHub no longer exists, the team will transfer the repositories to another available platform or archive the final version in the VT Libraries.

As a planning project, the software, scripts, and algorithms shared via GitHub will be supported for at least two years after project completion. However, these are not expected to be maintained through changes in language versions or the need to migrate to new tools.

Organizational Profile

Mission Statement
Inspired by our land-grant identity and guided by our motto, Ut Prosim (That I May Serve), Virginia Tech is an inclusive community of knowledge, discovery, and creativity dedicated to improving the quality of life and the human condition within the Commonwealth of Virginia and throughout the world. (from https://vt.edu/about/facts-about-virginia-tech.html)

Governance Structure
Virginia Tech is Virginia's most comprehensive university and a leading research institution. It has more than 37,000 undergraduate, graduate, and professional students. It manages a research portfolio of $556 million and has over 2,000 instructional faculty members. It had a 1.89 billion operating budget in 2022-23. (from https://vt.edu/about/facts-about-virginia-tech.html)

Service Area
The Commonwealth of Virginia, the nation, and the world community.

Brief History
Virginia Polytechnic Institute and State University, popularly known as Virginia Tech, officially opened on Oct. 1, 1872, as Virginia's white land-grant institution. During its existence, the university has operated under three other legal names: Virginia Agricultural and Mechanical College since 1872, Virginia Agricultural and Mechanical College and Polytechnic Institute since 1896, and Virginia Polytechnic Institute since 1944. The state legislature sanctioned university status and bestowed upon it the present legal name effective June 26, 1970.

Established in 1872 with 500 volumes, Virginia Tech Libraries now includes holdings of more than 2 million volumes physically located in Newman Library and four branches. The library is a selective depository for federal documents and is an invited member of the Association of Research Libraries (ARL). The library's mission is to invent the future of libraries at Virginia Tech. We honor tradition as we excel in our core mission to provide access to information. We acknowledge change as we adapt to address the new information needs and Open Web's reframing of the academic and research enterprises in higher education. We embrace a diversity of thought and culture as we find solutions to information challenges when meeting user needs. Over the next decade, we anticipate seismic shifts in the nature of libraries across the globe. The form, function, and overall identity of the library as an institution will evolve. At Virginia Tech, we envision the library of the future emerging as a platform for student success and faculty innovation in a global context, a hub for strategic partnerships, and a regenerating entity that adapts to changing user needs and expectations.