Annotation Efficient Learning with Affinity Graphs

Zehua Cheng
Wolfson College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Trinity 2024

This thesis is dedicated to my parents for their indispensable, altruistic support.

Acknowledgements

In the heart of my Ph.D. journey, there lies a profound sense of gratitude that I wish to convey, particularly to those who have illuminated my path with their wisdom and unwavering support.

At the pinnacle of my acknowledgments rests Prof. Thomas Lukasiewicz, my esteemed supervisor, whose guidance has been a beacon throughout my research odyssey. His unwavering support and insightful mentorship have been the bedrock upon which my academic pursuits have flourished. Prof. Lukasiewicz possesses a rare gift for nurturing curiosity, fostering an environment where the pursuit of novel ideas is not just encouraged but celebrated. Each step of my research journey, marked by experimentation and the courage to venture into uncharted territories, was met with his steadfast encouragement. His belief in the value of exploring diverse avenues of thought, coupled with his willingness to embrace my every endeavor. His door was always open, ready to welcome ideas, however nascent or unconventional, and transform them into refined research paths.

Prof. Zhenghua Xu stands as a beacon in the early chapters of my research career. His insightful guidance, akin to a lighthouse guiding a ship through stormy seas, has been instrumental in shaping my scholarly compass. I am eternally thankful for the moments where his keen eye for detail and deep understanding of the discipline transformed vague notions into concrete theories. Our collaboration on my maiden research paper, alongside Prof. Thomas Lukasiewicz, was a defining experience. Their combined support, a symphony of encouragement and rigorous mentorship, not only facilitated the completion of that pivotal work but also instilled in me a resilience that has carried me through the entire thesis.

To Dr. Lianlong Wu, I owe a debt of gratitude that transcends words. Our shared enthusiasm for graph learning created a vibrant tapestry of ideas, woven during countless hours of animated discussions. Dr. Wu's presence was like a spark igniting the embers of creativity; his insights, sharp and illuminating, shed light on intricate pathways within the complex landscape of our research. Our collaborative explorations

felt like adventurous expeditions, where every challenge was an opportunity to delve deeper into the mysteries of our field. His unwavering support formed a pillar of strength, making our intellectual odyssey both rewarding and unforgettable.

Dr. Nanqing Dong holds a special place in this narrative of gratitude. His dedication to our joint endeavors, evident in the countless hours we spent refining papers, was nothing short of inspiring. His meticulous feedback, each comment a precious gem, polished my writing into a form that could effectively communicate the nuances of our research. Dr. Dong's mentorship extended beyond technicalities; he taught me the art of storytelling with data, revealing how a well-crafted narrative can breathe life into statistical analyses. His patient guidance not only honed my skills as a researcher but also nurtured in me a deeper appreciation for the beauty in scientific communication.

Over seven years, Dr. Wei Dai's exceptional engineering prowess illuminated my path. A virtuoso in his field, his unique talent and skillset remain unmatched in my experience. His intuitive and precise problem-solving approach has been my consistent guide and inspiration. Our collaboration transcended conventional bounds, fostering a synergy that saw us bravely dive into complex challenges, strengthening not just our technical skills but also forging an unbreakable bond. Countless brainstorming nights and coding marathons, leading to triumphs in deep learning competitions, are etched in my fondest memories. These victories were no accident; they signify Dr. Dai's relentless pursuit of excellence, his talent for simplifying complexity, and his steadfast trust in our team's potential—a testament to the power of partnership and shared vision.

My heartfelt gratitude also extends to my peer collaborators, Miss Di Yuan, Dr. Guosheng Hu, Dr. Emanuel Sallinger, Prof. Georg Gottlob, and Dr. Yuan Li whose timely and invaluable advice significantly shaped my research ideas and experiments. Their seasoned guidance proved indispensable, enriching my academic journey with a depth of perspective that was both timely and immensely beneficial.

I am equally grateful to my friends from the Department of Computer Science and Wolfson College, whose camaraderie turned moments into memories. Shared laughter during tennis matches, the elegance of formal dinners, the tranquility of punting on lazy afternoons, the excitement of balls, and casual drinks that turned into profound conversations. These experiences have colored my academic journey with joy and camaraderie, creating a vibrant backdrop against which my research flourished.

Lastly, my deepest gratitude is reserved for my family, especially my parents, whose unwavering support and boundless love have been the bedrock of my journey.

Their constant encouragement and belief in my dreams have been a beacon, guiding me through every challenge and triumph. Their love is the eternal fuel that sustained me, and for that, I am eternally thankful.

Abstract

Annotation-efficient learning has emerged as a critical area of research due to the scarcity of labeled samples posing a substantial barrier to developing robust fully-supervised deep neural networks. When the availability of labeled samples is limited, developing an effective learning system becomes a formidable challenge. Conversely, unlabeled data is often plentiful and can be obtained at a relatively low cost. Consequently, the concept of leveraging a substantial volume of unlabeled data to train deep models, despite the paucity of labeled samples, emerges as a compelling proposition.

This thesis explores novel approaches to annotation-efficient learning in machine learning, with a focus on leveraging implicit relationships within data to improve model performance in scenarios with limited labeled data. The research addresses three key challenges: effectively utilizing implicit structures in input data, integrating these structures into the learning process, and determining optimal representations for extracting and utilizing implicit relationships. The methodology centers on the development and application of affinity graph constraints across three domains: self-supervised learning for whole-slide image analysis, semi-supervised learning for medical image segmentation, and multi-modal learning for single-cell data integration.

The methodology centers on the development and application of affinity graph constraints across three domains: self-supervised learning for whole-slide image analysis, semi-supervised learning for medical image segmentation, and multi-modal learning for single-cell data integration. Specifically, we propose:

1. An affinity graph constraint (AGC) for self-supervised learning on whole-slide images, which captures fine-grained features and improve existing self-supervised methods.

2. An affinity-graph-guided contrastive learning framework for semi-supervised medical image segmentation, incorporating patch-wise class-centric sampling and hard-negative reweighting.

3. A single-cell Affinity Graph transFormer (scAGFormer) for multi-modal single-cell analysis, which employs an affinity graph prior to improve modality transformation.

Key findings demonstrate significant improvements over state-of-the-art methods across all three domains. The proposed affinity graph-based approaches consistently enhance model performance, particularly in scenarios with limited labeled data. The frameworks show remarkable generalizability across different datasets and tasks, highlighting their potential for broad application in medical image analysis and computational biology.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Deep learning has changed many domains like medical imaging by its extraordinary performance in various tasks such as diagnosis and treatment planning. However, the success of deep learning algorithms is highly dependent on large amounts of high-quality annotated data. Interestingly, despite the significant cost of annotating effort from highly-trained medical professionals, there is still no universally accepted gold standard for diagnosing many targets in the field of medical imaging [80, 190].

To fully utilize the limited annotation within medical imaging, annotation-efficient learning which focus on improving the annotation utilization via novel pipeline has been considered as an important option when deploying deep learning application in medical imaging area. Annotation-efficient learning is grounded in robust representation and models, particularly within the specialized context of medical imaging. This approach involves adapting learning pipelines to tackle specific challenges inherent from different modalities of input. Annotation-efficient learning typically involves representation schemes that are able to capture implicit interactions within data, a foundation model to extract reliable features, and the learning pipeline to utilize the novel representation and reliable features on scale. From a broader perspective, foundation models are trained in accordance with similar paradigms, such as GPT's unsupervised training on a massive corpus of text, CLIP [209] construct contrastive training on a massive collection of image-text pairs, and DINO [35] perform self-supervised learning through self-distillation on a massive dataset of images. Annotation-efficient learning typically leverages various foundation models, such as [157], which utilizes pre-trained Visual Transformer [74] and ResNet [107] to establish a large-scale patch contrastive learning for its self-supervised learning on whole-slide images and achieved significant improvement over downstream tasks. The advancement of representation

can broadly benefit annotation-efficient learning and the acquisition of foundation models. Given the exponential growth of models and data, reliable annotations become increasingly difficult to obtain. Better representations often lead to significant impact. Therefore, this thesis is **built upon successful foundation models and aim to find a reliable representation to improve the annotation utilization within different learning pipeline**.

Applying annotation-efficient learning to medical imaging presents unique challenges. The most commonly utilized modalities in medical imaging include Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and Whole-Slide Imaging (WSI), each of which has distinct issues. For instance, whole-slide images (WSIs) are characterized by their immense file sizes. A comprehensive pairwise comparison study often requires over 1 TB of storage due to the highresolution nature of WSIs, which can range from $50,000 \times 50,000$ pixels to over $100,000 \times 100,000$ pixels, depending on tissue dimensions and scanning resolution. Consequently, processing an entire WSI directly is computationally and memory-intensive, rendering such an approach infeasible. To address this, WSis are divided into smaller tiles, or patches, with typical dimensions of 256×256 or 512×512. Therefore, directly processing the full WSI is infeasible due to computational and memory constraints. Instead, WSIs are divided into smaller tiles (patches), with typical sizes like 256×256 and 512×512. This tiling approach introduces specific challenges for deep learning applications, **scale discrepancy** [24], **tiling effect** [216]. These issues arise due to the loss of contextual information and potential inconsistencies introduced by processing localized regions rather than the full slide.

Current approaches to the analysis of CT/MRI have achieved considerable success with the help of deep learning techniques. For instance, CNNs have been applied to the extraction of spatial features from CT/MRI slices; more recently, there is also the possibility of volumetric analysis enabled by advances in 3D CNNs. Other approaches include patch-based processing and region-based segmentation, which increased computational efficiency and improved the accuracy of the techniques. Semi-supervised methods using pseudo-labels and data augmentation have also been tried to overcome the scarcity of labeled data. Despite these advances, several challenges remain. Most of the methods fail to capture the global context because of inherent limitations due to processing individual slices or patches. Besides, pseudo-labeling methods may introduce error propagation and degrade model performance. The anatomical variability across patients further complicates generalization, and the requirement of consistent annotation across slices creates significant bottlenecks toward training robust models.

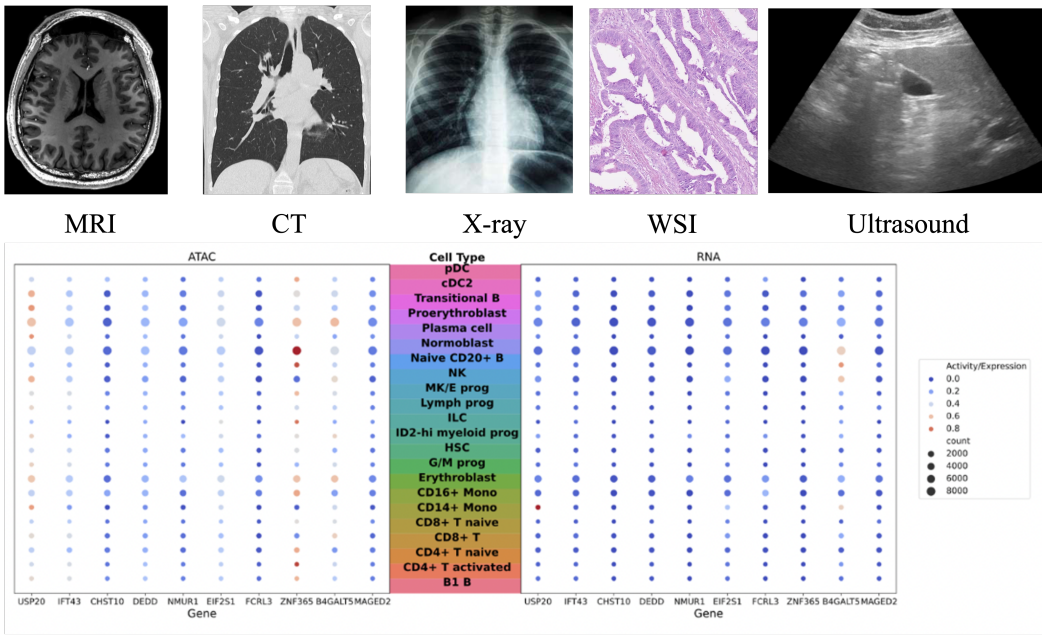

Figure 1.1: Representative data samples from diverse modalities used in biomedical and clinical research. The top row features imaging modalities including Magnetic Resonance Imaging (MRI), Computed Tomography (CT), X-ray, Whole Slide Imaging (WSI), and Ultrasound. The bottom panel highlights multi-modal single-cell data integrating ATAC and RNA sequencing modalities, showcasing gene activity/expression levels across various cell types.

Therefore, both WSIs and CT/MRI modalities are becoming very crucial in finding methods that are annotation-efficient. These methods have reduced the dependency on exhaustive manual annotation by leveraging the power of unlabeled data to capture meaningful patterns in terms of their implicit relationships, hence allowing efficient analysis of volumetric and high-resolution medical imaging data in a scalable way.

For example, the fine-tuning of the EfficientNet [243] model on the CytolmageNet [116] dataset achieved only 11.32% classification accuracy on the validation set, underlining the high difficulty of the dataset due to label imbalance and intricate categories. Therefore there is a huge domain gap between general and medical datasets (ImageNet v.s. CytoImageNet). Thus, direct application of foundation models to medical imaging may not be appropriate. One possible way to solve this could be combining general-purpose image signals with domain-specific medical image signals in a hybrid model for effective handling of downstream tasks. In conclusion, there are several challenges associated with medical imaging, and creative solutions need to be offered. Efficient annotation methods coupled with novel representations such

as the affinity graph constraints have proved promising for such limitations. These frameworks implicitly model the spatial and semantic relationships in data and thus provide efficient learning with limited annotations for scalability and robustness across a wide range of medical imaging domains.

Researchers have developed various learning pipelines to improve annotation utilization. For example, semi-supervised learning [52, 234, 162, 154, 195], and self-supervised learning [249, 95, 51, 303] are the most popular learning pipeline in annotation-efficient learning.

Semi-supervised learning [315] is a paradigm within machine learning that harnesses both labeled and unlabeled data to enhance model performance. Semi-supervised learning bridges the gap between supervised learning, which relies exclusively on labeled data, and unsupervised learning, which operates solely on unlabeled data. However, similar to active learning, semi-supervised learning also suffered from data quality and label noise [296, 298]. Errors in the initial labeled data can propagate through the learning process, especially in techniques like self-training and co-training, where mislabeled data can mislead the model during iterative training stages. This error propagation can degrade the overall performance of the model, making it less reliable. Thankfully, leveraging implicit relationships within the input data has the potential to mitigate the stated difficulties [148, 30]. For example [308] exploit the inherent relationships between data points within a batch, potentially improving the learning process by capturing implicit connections that might not be apparent in traditional supervised or unsupervised methods. focuses on explicit relationships, it provides a foundation for understanding how relationship information can be incorporated into semi-supervised learning frameworks. [312] using contrastive learning techniques, the model can better capture the implicit relationships within the data, potentially improving its robustness to label noise in semi-supervised settings. OpenRE [113] improve the generalization of semi-supervised learning by classifying both explicit and implicit relationships in known and novel classes from unlabeled data. Therefore, exploration of implicit relationships in semi-supervised learning would made semi-supervised learning to be an option for the annotation-efficient learning system.

Self-supervised learning [11] is a paradigm where the model generates its own labels from the input data, eliminating the need for manual annotation. This approach leverages the inherent structure within the data to create supervised signals, which can then be used to train the model [130]. Although self-supervised learning does not require human involvement, in certain instances, the dataset demands for self-supervised learning can surpass those of supervised learning since the model

Aspect	Sup. Learning	Semi-sup. Learning	Self-sup. Learning	Unsup. Learning
Data Utilization	$\mathcal{L}$	$\mathcal{L}$ and $\mathcal{U}$	$\mathbf{X}_u + \mathbf{X}_l$	$\mathbf{X}_u + \mathbf{X}_l$
Label Req,	High	Moderate	None	None
Human Involvement	High	Moderate	None	None
Complexity	Low	High	Moderate	Moderate
Scalability	Low	Moderate	High	High
Transferability	Low	Moderate	High	High

Table 1.1: High-level comparison of different learning paradigms across label requirements (Label Req.), data utilization, level of human involvement, complexity of the overall paradigm, the scalability and the transferability to new domain (Transferability).

must derive knowledge from the inherent patterns in the data without the guidance of explicit labels. For example, [92] found that increasing the amount of pretraining data from 1 million to 1 billion images led to substantial improvements in downstream task performance. Researchers are presently redirecting their focus towards the foundational premise of self-supervised learning, which entails initiating the learning process from the inherent, latent structure inherent within the data [221, 97, 254]. PCDNet [254] propose to decomposes an image into object components represented as transformed versions of learned object prototypes to extract the implicit representation with explicit reconstruction representation. [69] proposes a novel approach that utilizes the semantic structure of images by incorporating semantic relationships between different parts of an image. By exploiting the implicit relationships, self-supervised learning would able to more efficiently extract the inherent, latent structure inherent within the data and improve the overall performance.

I present the difference of the learning pipeline in Table 1.1. I got dataset $\mathcal{D} = \mathcal{L} \cup \mathcal{U}$, labeled $\mathcal{L} = \{(x_i, y_i)\}_{i=1}^{N_l}$ and unlabeled $\mathcal{U} = \{x_j\}_{j=1}^{N_u}$ datasets with N_l number of labeled data and N_u number of unlabeled data. $f(\cdot;\theta)$ is parametric function model parameterised by θ. The $\mathbf{X}_u = \{x_i\}_{i=1}^{N}$ and $\mathbf{X}_l = \{x_j\}_{j=1}^{N}$ are labeled and unlabeled inputs. λ is regularization parameter balancing labeled and unlabeled loss. Q is set of queried points in active learning. I would further discuss the details of self-supervised learning and semi-supervised learning in Chapter 2.

In summary, within the domain of annotation-efficient learning, both semi-supervised and self-supervised learning paradigms can significantly benefit from enhanced utilization of implicit relationships within the data. This approach addresses the fundamental challenge of limited labeled data by leveraging the inherent structure and associations present in both labeled and unlabeled samples. Furthermore, recent advancements in graph neural networks and attention mechanisms provide powerful

tools for modeling complex relationships in both semi-supervised and self-supervised settings [263, 231]. These architectures can capture long-range dependencies and multi-modal interactions, enabling more sophisticated exploitation of implicit data structures. Based on such insight, I explore the graph-based and attention-based module in self-supervised learning and semi-supervised learning area.

1.2 Research Challenges and Objectives

To improve the annotation utilization, I articulate it through the following research questions (RQ):

RQ1 What computational frameworks or mathematical representations can be utilized to efficiently elucidate and encode the implicit constraints within input data structures?

RQ2 How this representation (in RQ1) applied into the learning process?

RQ3 Can this representations (in RQ1) extract reliable implicit constraints from multiple source?

To answer these research questions, I research diverse domains in annotation efficient learning and figured out the effectiveness of affinity graphs. I believe affinity graph is an effective option to extract the implicit representation which answer the **RQ1**. I applied affinity graphs in self-supervised learning (in Chapter 3) and semi-supervised learning (in Chapter 4) to answer the **RQ2**. I then applied affinity graph into multi-modal learning to learn interactions between different modalities in bioinfomatics data (in Chapter 5) therefore answer the **RQ3**. I further discuss the details of how these research questions being addressed in Chapter 6.

1.3 Contributions

The research presented in this dissertation can be categorized into these key components:

- I propose a novel affinity graph constraint (AGC) for self-supervised learning on whole slide images (WSIs). This constraint is designed to capture fine-grained relationships between image patches and is compatible with existing self-supervised learning methods. I demonstrate that AGC significantly improves the performance of state-of-the-art self-supervised learning approaches

across multiple WSI datasets. Notably, the proposed method scales effectively to extremely large input images, as evidenced by successful application to a WSI dataset exceeding 1 TB in size. This scalability addresses a key challenge in applying self-supervised learning to pathology data. To scale up the application of the proposed method, I introduce a random query selection method based on multi-scale features to effectively utilize both global and local information in WSIs. This approach allows the proposed method to handle the immense size of WSIs while preserving important contextual information. By combining this selection method with our affinity graph constraint, the method achieved a balance between computational efficiency and feature richness, enabling effective learning from large-scale pathology data.

- I develop an affinity-graph-guided semi-supervised contrastive learning framework (Semi-AGCL) for medical image segmentation. This framework integrates the strengths of contrastive learning and semi-supervised learning, utilizing affinity graphs to enhance feature representation and improve segmentation accuracy with minimal annotations. Semi-AGCL demonstrates significant performance improvements over existing methods, particularly in scenarios with extremely limited labeled data. I propose a patch-wise class-centric sampling method guided by an entropy-based metric for my semi-supervised learning framework. This approach mitigates class collision issues in contrastive learning and provides a more informative supervision signal. By focusing on informative and diverse patches, the proposed method makes more efficient use of limited labeled data in medical image segmentation tasks.

- I propose a cell-agnostic framework called single-cell Affinity Graph transFormer (scAGFormer) for multimodal single-cell analysis. This method integrates statistical feature learning and representation-based learning through the application of correlation loss, enabling effective modality prediction without relying on cell embeddings. scAGFormer demonstrates superior performance in tasks such as predicting protein expression from gene expression data, outperforming existing cell-dependent methods.

- I applied affinity graph on diverse domains and diverse annotation efficient learning pipelines. I conduct extensive experiments on various datasets across different domains, including whole slide images (CAMELYON 16, NCT-CRC-HE-100K, TCGA Lung Cancer), medical image segmentation (LA, ACDC,

CARG), and single-cell multimodal data (NeurIPS 2021 and 2022 Multimodal Single-Cell Integration Challenge datasets). The experimental results demonstrate the effectiveness and generalizability of the affinity graph representation in self-supervised learning, semi-supervised learning, and multi-modality learning. I also provide detailed ablation studies and analyses of the proposed methods, offering insights into the workings of affinity graph-based approaches in annotation-efficient learning scenarios. These studies include investigations into the impact of different components of the proposed frameworks, the effectiveness of various loss functions, and the robustness of our methods under different experimental conditions.

These contributions advance the field of annotation-efficient learning by leveraging affinity graphs to improve performance in self-supervised, semi-supervised, and multimodal learning tasks. Our work demonstrates the potential of affinity graph-based methods to address key challenges in learning from limited labeled data across various biomedical applications, including whole slide image analysis, medical image segmentation, and single-cell multiomics data integration.

Throughout my D.Phil studies, there are relevant publications to support this thesis. I explicitly indicate my contributions according to CRediT taxonomy [25] to each included paper using numerical markers are outlined below:

1. Conceptualization: Developing the overarching ideas and framework for the research.
2. Methodology: Designing the methods and procedures used in the research.
3. Software: Writing and managing the code and software programs required for the study.
4. Validation: Verifying and assessing the reliability and accuracy of the findings or models.
5. Formal Analysis: Conducting detailed, structured data analysis and interpretation.
6. Investigation: Executing experiments and collecting data or resources.
7. Resources: Providing materials, instruments, or other essential support for the research.
8. Data Curation: Managing, organizing, and preserving the research data.
9. Writing - Original Draft: Composing the initial version of the manuscript or report.

10. Writing - Review & Editing: Refining and revising the manuscript or report for clarity and accuracy.
11. Visualization: Creating visual representations, such as graphs or diagrams, to support the findings.

For each paper, I specify the type(s) of contribution made by assigning the corresponding numbers in RED.

1. **Zehua Cheng**, and Lianlong Wu. *Hybrid Learning System for Large-scale Medical Image Analysis*. In IJCAI-ECAI 2022. 1,2,3,5,6,9,10

2. **Zehua Cheng**, Nanqing Dong, Di Yuan, Lianlong Wu, Xianhe Chen, and Thomas Lukasiewicz. Self-supervised Learning with Affinity Graph Constraint on Whole-Slide Images. under review of Neurocomputing. 1,2,3,6,9,10,11

3. **Zehua Cheng**, Nanqing Dong, Lianlong Wu, Xianhe Chen, Di Yuan, Wei Dai, Simiao Zhao, and Thomas Lukasiewicz. *Cell-agnostic Modality Transformation with Affinity Graph Prior in Single-Cell Multi-modal Prediction*. under review of IEEE/ACM Transactions on Computational Biology and Bioinformatics. 1,2,3,6,9,10

4. **Zehua Cheng**, Di Yuan, and Thomas Lukasiewicz. *Affinity-Graph-Guided Contractive Learning for Pretext-Free Medical Image Segmentation with Minimal Annotation*. In BIBM 2024. 1,2,3,5,6,9,10

The following papers, which I co-authored, are not directly relevant to the scope of this thesis:

5. Jiafeng Liu, Yuanliang Dong, **Zehua Cheng**, Xinran Zhang, Xiaobing Li, Feng Yu, Maosong Sun, *Symphony Generation with Permutation Invariant Language Model*. In ISMIR 2022. 2,3,6,9

6. **Zehua Cheng**,Wei Dai, and Jiahao Sun. *Visual Language Model for Preclinical Toxicologic Liver Histopathology Assessment*. In ACM Multimedia VLM4Bio Workshop 2024. 1,2,3,5,6,9,10

7. **Zehua Cheng**, Manying Zhang, and Jiahao Sun. *On Weaponization-Resistant Large Language Models with Prospect Theoretic Alignment*. In COLING 2025. 1,2,3,5,6,9,10

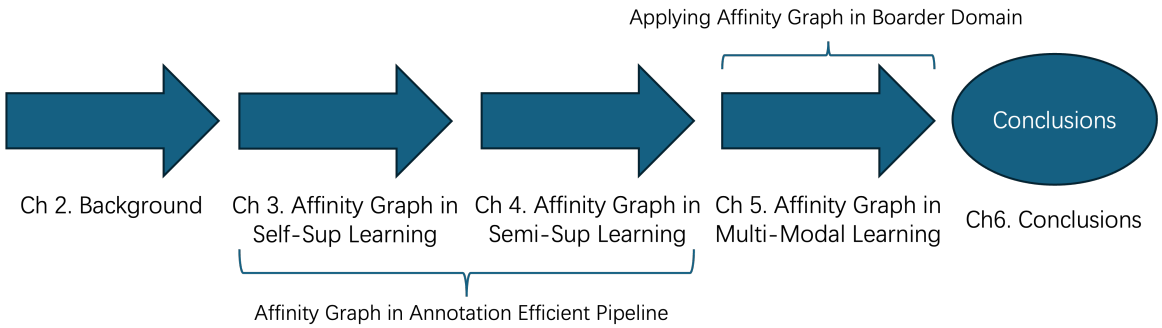

Figure 1.2: Flowchart for the thesis structure for each chapter. Chapter 3 and Chapter 4 applied affinity graph in self-supervised and semi-supervised learning.

8. **Zehua Cheng**, Di Yuan, Wenhu Zhang, and Thomas Lukasiewicz. *Effective and Efficient Medical Image Segmentation with Hierarchical Context Interaction.* In WACV 2025. 1,2,3,5,6,9,10

1.4 Thesis Structure

We present the flowchart for the thesis structure in Figure 1.2. The main content and structure of each chapter in this paper are summarized below:

- **Chapter 2** present the foundational concepts and methods relevant to the research. It begins with a review of the role of deep learning in medical imaging and then transitions to an in-depth exploration of annotation-efficient learning techniques, including self-supervised and semi-supervised learning. The theoretical underpinnings of these approaches, such as the cluster and manifold assumptions, are discussed. The chapter also introduces the concept of implicit constraints and affinity graphs, emphasizing their potential to enhance learning by capturing intrinsic relationships within data. Recent breakthrough with foundation models and large-language models also discussed and present the limitation of applying foundation models and large-language models in medical imaging and bioinformatics area.

- **Chapter 3** introduce **Affinity Graph in Self-Supervised Learning**, introduces the application of affinity graphs to self-supervised learning, focusing on whole-slide image analysis. It presents the development of an Affinity Graph Constraint (AGC) methodology, which captures fine-grained features and improves existing self-supervised methods. The chapter provides a detailed explanation of the methodology, followed by experimental results and

evaluations across various datasets, demonstrating significant improvements in self-supervised learning performance.

- **Chapter 4** introduce **Affinity Graph Constraint on Semi-Supervised Learning**, extends the use of affinity graphs to semi-supervised learning in the context of medical image segmentation. This chapter introduces a semi-supervised framework that integrates affinity-graph-guided contrastive learning, innovative sampling strategies, and loss functions. It addresses challenges such as reliance on pretext tasks and insufficient supervision signals. Through rigorous experimentation, the chapter showcases the framework's effectiveness in improving segmentation accuracy, particularly with minimal annotations.

- **Chapter 5** introduce **Affinity Graph in Multi-Modal Learning**, explores the use of affinity graphs in multi-modal single-cell data analysis. It proposes the scAGFormer model, which combines statistical and representational feature learning to predict relationships between different biological modalities, such as gene expression and protein levels. The model's cell-agnostic approach enables it to handle diverse and complex biological data. This chapter includes extensive experimental evaluations, demonstrating that the proposed method outperforms existing state-of-the-art approaches in multi-modal predictions.

- The last **Chapter 6** synthesizes the findings and contributions of the thesis. It revisits the research questions posed in the introduction, providing comprehensive answers that emphasize the utility and versatility of affinity graphs in various learning paradigms. The chapter also discusses the broader implications of the work and outlines potential future directions, including extensions beyond Transformers, further theoretical exploration of affinity graphs, and their applications in multi-modal large-language models.

Chapter 2

Background

2.1 Annotation Efficient Learning

Annotation efficient learning refers to methods that yield learning reliable representation without relying on massive carefully labeled training datasets [242]. It has attracted significant attention from the medical imaging research community due to the difficulty and expense of collecting large, representative, and accurately annotated datasets. Annotation efficient learning has covered a broad spectrum machine learning pipelines, including the following concepts:

- **Semi-supervised learning**: Approaches that leverage both a small amount of labeled data and a large amount of unlabeled data during training. The key idea is to use the labeled examples to provide initial supervision, and then exploit the structure and patterns in the unlabeled data to infer labels for those examples and iteratively improve the model. Techniques include consistency regularization, pseudo-labeling, and generative models.

- **Self-supervised learning**: Methods that learn useful representations from unlabeled data by solving pretext tasks that do not require manual annotations. The learned representations can then be fine-tuned on a small labeled dataset for the target task. Contrastive learning and pixel restoration are examples of self-supervised pretext tasks used in computer vision.

- **Few-shot learning**: Techniques that enable models to learn new concepts from just a handful of labeled examples per class. This is achieved through approaches like meta-learning, which trains models on a variety of tasks so they can quickly adapt to new tasks.

- **Incremental learning**: Approaches that allow models to continuously learn from new data over time, without forgetting previously acquired knowledge. The key challenge is avoiding catastrophic forgetting, where the model overwrites past information when learning from new data. Incremental learning is important for adapting to evolving data distributions and incorporating new classes over time.

However, although these methods have achieved great success in annotation efficient learning, semi-supervised learning and self-supervised learning have emerged as two of the most promising approaches. Few-shot learning and incremental learning are not as effective for annotation efficient learning in the medical domain because they either still require substantial labeled data for pretraining (few-shot learning) or do not reduce the need for continuous expert annotations (incremental learning), while also introducing complexities that make them less practical for critical and sensitive medical applications. Semi-supervised learning and self-supervised learning enable models to harness the vast amounts of complex, unlabeled medical data to learn intricate patterns and representations without relying heavily on scarce and expensive expert annotations, thus improving diagnostic accuracy and efficiency.

A significant challenge in annotation efficient learning is scaling the approach to handle more complex real-world tasks without sacrificing performance [247, 263]. Both incremental learning and few-shot learning techniques rely on dynamically updating the model as it is being applied, to adapt to new data with limited annotations [244, 247]. This would much more similar to an ongoing service instead of an ultimate solution for the annotation efficient learning. To tackle the scalability issue, we have to face up to core challenge of the annotation efficient learning where using reliable model to learn patterns from the unlabeled data. Semi-supervised learning methods like consistency regularization [279, 186, 185] and pseudo-labeling [154] can effectively leverage large amounts of unlabeled data alongside a small labeled dataset . This allows them to exploit the structure and patterns in the unlabeled examples to improve model performance. Self-supervised learning aims to learn useful representations from unlabeled data by solving pretext tasks that don't require manual annotations. The learned representations can then be fine-tuned on a small labeled set for the target task, greatly reducing the need for expensive labeled examples.

Therefore, we would focus on the solution of self-supervised learning and semi-supervised learning to improve the annotation utilization. We would further discuss the self-supervised learning and semi-supervised learning in Section 2.2 and Section 2.3.

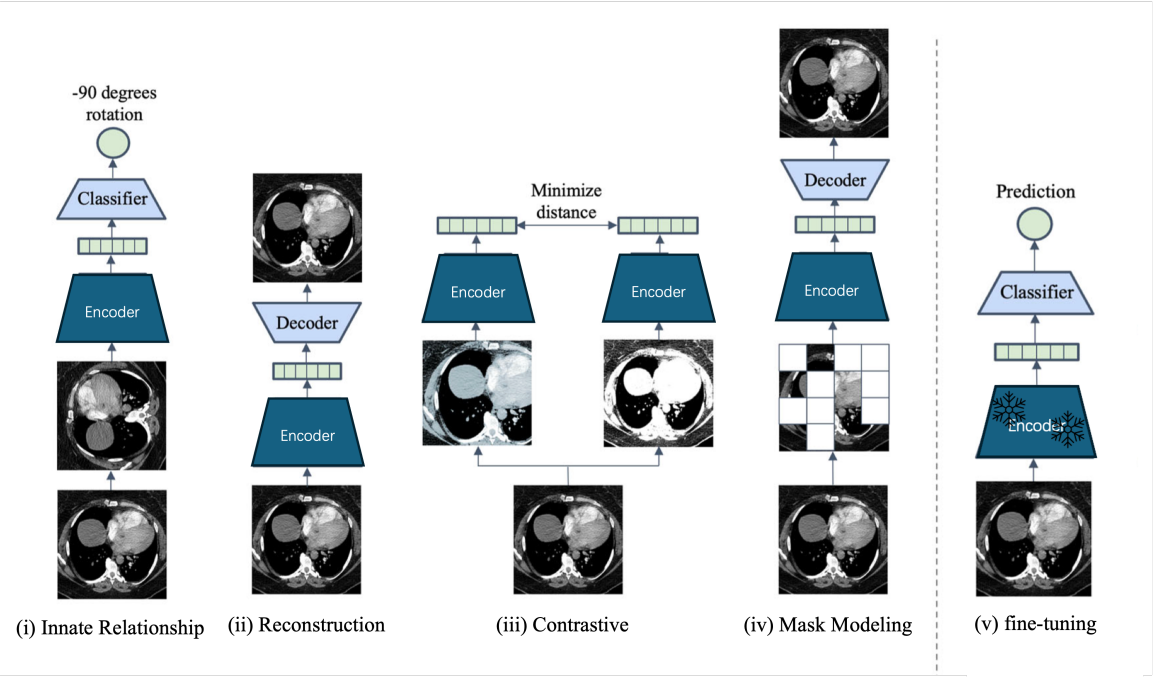

Figure 2.1: Illustration of different self-supervised learning pipelines. (i) Innate Rel. training the self-supervised learning model by utilize pre-crafted tasks, focusing internal relationship of the data; (ii) Reconstruction training the self-supervised learning model by learning the distribution and trying to reconstruct the original input signal; (iii) Contrastive self-supervised learning form the positive pairs between different views of augmentation of the same input and minimizes representational distances of positive samples. A more theoretical framework illustration is presented in Figure 2.2; (iv) Mask Modeling entails initially obscuring selected portions of an image and subsequently reconstructing it based on intact areas. The model then applying to downstream task with (v) fine-tuning.

2.2 Self-supervised Learning

Self-supervised learning (SSL) has emerged as a crucial method in machine learning, particularly within deep learning, where it addresses the challenge of limited labeled data by leveraging large amounts of unlabeled data. SSL generates labels from the data itself through various techniques, thus facilitating model training without the need for extensive labeled datasets.

Self-supervised learning has been widely applied in diverse domains like computer vision [36, 48, 104], natural language processing [70, 87, 26], speech [115, 171, 10], robotics [222], and healthcare [151]. This thesis mainly explore applying SSL over computer vision tasks. The main strategies for SSL to construct universal and robust features from the data could be summarized in Figure 2.1.

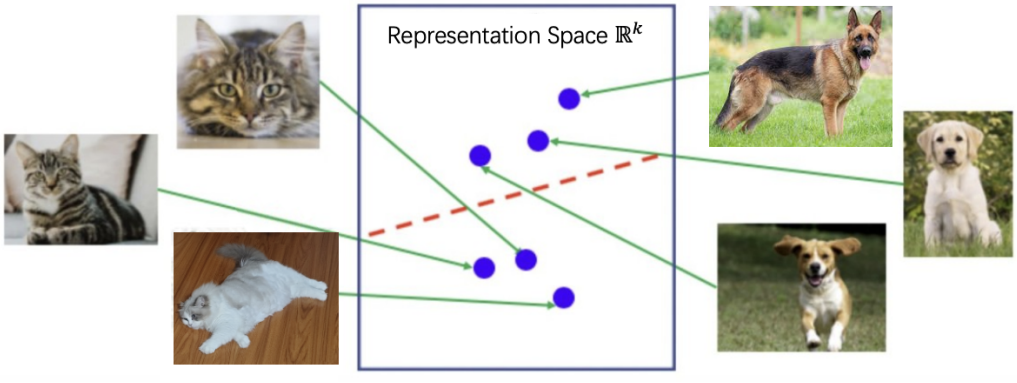

Figure 2.2: Contrastive learning: images are embedded in a representation space $\mathbb{R}^k$, where similar samples (e.g., cats) are pulled closer together, and dissimilar samples (e.g., cats vs. dogs) are pushed apart to enhance feature separation.

Contrasive learning has been considered as one of the most successful pipeline in self-supervised learning. The success of contrastive learning is built upon capturing invariant features while discarding irrelevant features through contrasting positive and negative pairs by pushing similar examples together and dissimilar ones apart [48]. We present the illustration of concept in Figure 2.2 with cats and dogs. A dashed red line suggests a potential classification boundary between cats and dogs. The figure shows that the cat and dog embeddings (blue points) are clustered in different regions. In Table 2.2, there is around two thirds of the reviewed articles—specifically, forty-four out of seventy-nine—employed a form of unsupervised pretraining known as contrastive learning. This approach is crucial in preparing these studies' datasets to be analyzed effectively by subsequent machine learning models. SimCLR [47], MoCo [105] and BYOL [95] were the three most used frameworks. Some papers leveraged medical domain priors to develop tailored methodologies for generating positive pairs. Specifically, in pathology slices, [155] utilized a strategy that areas surrounding particular lesions tend to exhibit homogeneity. This pre-clustering techniques designed to identify incongruent patches within the dataset.

In recent research, contrastive learning has enabled self-supervised visual representation learning that rivals or even surpasses supervised pretraining on tasks such as image classification, object detection, and segmentation [147]. While contrastive learning approaches have demonstrated impressive results in various domains, they are not without certain constraints. Such as, require careful design of data augmentations and negative sampling strategies, which can significantly impact performance [146]. Moreover, these methods often necessitate large batch sizes [252, 38] and a substantial

Dataset	Organ	Modality	#Data	Year	Task	Performance	Solution/Ref
MIMIC-III [133]	–	EHR	40,000	2016	Length-of-Stay Prediction	RMSE ≈ 2.65 days	[219]
	_			2016	Mortality Prediction	AUC ≈ 0.87	[219]
BRATS [191]	Brain	MRI	220	2016	Segmentation	DSC = 86%	[217]
		fMRI	285	2017	Segmentation	DSC = 90.9%	[207]
		MM-MRI	285	2018	Segmentation	DSC = 86.9%	[259]
		MM-MRI	335	2019	Segmentation	DSC ≈ 92%	[241]
		MM-MRI	369	2020	Segmentation	DSC = 88.5%	[223]
		MM-MRI	1,251	2021	Segmentation	DSC = 92.2%	[304]
MIMIC-CXR [132]	Chest	X-ray	371,920	2019	Multi-label Classification	AUC = 83.4%	[225]
CheXpert [120]	Chest	X-ray	224,316	2019	Classification	AUC = 93.3	[137]
ChestX-ray14[267]	Chest	X-ray	112,120	2017	Pneumonia Detection	AUROC = 84.4%	[292]
CAMELYON16 [15]	Lung	WSIs	270	2016	Segmentation	AUC = 94.6 %	[305]
GlaS [232]	CRC [1]	WSIs	165	2016	Segmentation	DSC= 93.0%	[205]
ACDC [18]	Heart	MRI	150	2017	Segmentation	DSC = 94.26%	[141]
MSD [230]	Multi-organs	–	2,633	2019	Segmentation	DSC = 90%	[121]
Kvasir [124]	GI polyp	Endoscope	8,000	2019	Segmentation	DSC = 92.6%	[211]

Table 2.1: Dataset and modality. MM-MRI refers to multi-modal MRI. EHR refers to electronic health record which is mainly natural language processing tasks.

number of negative examples to function effectively [57], thereby increasing computational cost. Additionally, the learned representations may be biased and may not capture all relevant features [57, 93], potentially leading to class collapse or feature suppression [289]. Addressing these limitations is essential for realizing the full potential of contrastive learning approaches and their practical application in various domains.

To address these challenges, researchers have developed several strategies. One approach is to constrain representation norms, which prevents dimensional collapse and encourages uniformity on the hypersphere [282]. Contrastive learning aims to map similar examples close together and dissimilar ones far apart in the embedding space. Without constraints on the representation norms, the embeddings can collapse into a lower-dimensional subspace instead of spanning the full embedding space. Constraining the representations to have L2 norm projects them onto a hypersphere, forcing the embeddings to be uniformly distributed on the hypersphere surface and preventing dimensional collapse [266]. Another effective strategy is to combine instance and class discrimination losses, providing complementary constraints for self-supervised learning (SSL) to produce more informative, robust, and adaptable representations [266]. Therefore, constructing external constraints is considered a helpful tool to improve contrastive learning and learn diverse representations. By incorporating these strategies, researchers can enhance the performance and generalizability of SSL models.

Masked image modeling (MIM) is a self-supervised learning approach that trains models to reconstruct missing parts of an input image [284]. By training on large-scale datasets without manual labels, MIM models learn robust and transferable features

[1] Colorectal adenocarcinoma

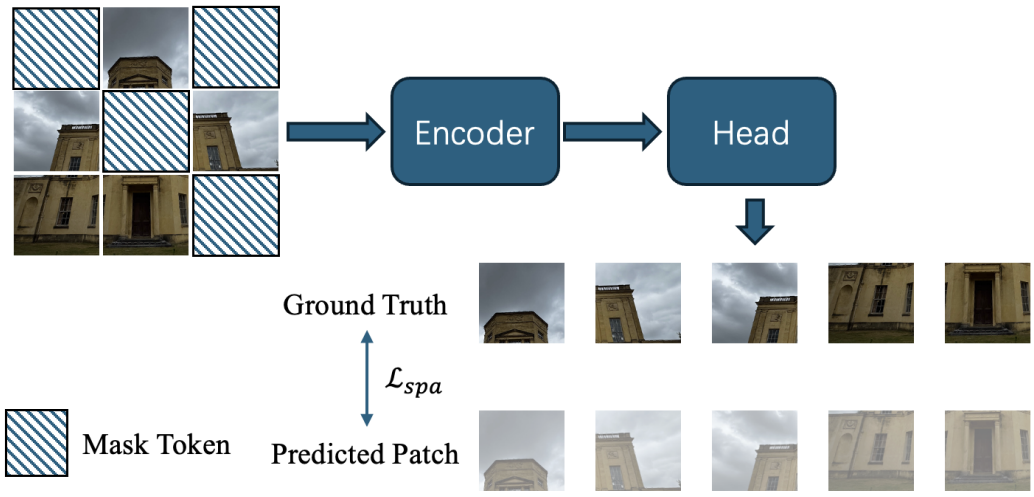

Figure 2.3: Illustration of Mask Image Modeling [284]

that achieve state-of-the-art results on a variety of computer vision tasks. The ability to focus on local details and learn the underlying structure enables MIM to generalize well to new domains. The illustraction of MIM is presented in Figure 2.3. By masking a portion of the image, MIM allows training on large unlabeled datasets, reducing the need for costly human annotations. The goal is to learn useful visual representations without relying on manual labels. State-of-the-art MIM models such as MAE [104], BEiT [12], and SimMIM [284] have achieved impressive performance on downstream tasks including image classification, object detection, and semantic segmentation, often surpassing supervised pre-training [283, 206, 43]. The masking process in MIM introduces a locality inductive bias, encouraging the model to attend to local details and maintain diversity in attention heads across all layers. By reconstructing the masked patches, the model is forced to learn meaningful representations that capture the underlying structure and semantics of the data [206]. This allows MIM to effectively leverage the inherent patterns in unlabeled images.

However, masked image modeling still faces efficiency and design challenges. Training MIM models can be computationally expensive due to the high masking ratio and need for a deep decoder to reconstruct the masked regions [178, 104]. Furthermore, the choice of masking strategy and reconstruction target significantly impacts performance and requires careful tuning [178]. For example, MIM models may struggle to generate high-frequency details and produce semantically meaningful reconstructions of the masked patches. To improve efficiency, some recent works like MixMIM [176] propose replacing the masked tokens of one image with visible tokens from another

image, and reconstructing both original images from this mixed input. To encourage learning of higher-level features, some methods predict hand-crafted descriptors of the masked patches, such as HOG (Histogram of Oriented Gradients)[258, 288] and clustering-based tokenization [284] as prediction targets. These provide more semantic, discriminative targets compared to raw pixels.

For both contrasive learning and MIM based methods, introducing external constraints has emerged as a robust and efficient strategy to enhance the effectiveness of self-supervised learning methods. This approach leverages additional information or structure from external sources to guide the learning process, enabling the model to learn more informative representations from unlabeled data. By incorporating external constraints, self-supervised learning models can achieve high accuracy with reduced reliance on large amounts of unlabeled data [307]. Furthermore, external constraints can improve computational efficiency by focusing the learning on the most informative aspects of the data [291] and enhance performance for specific tasks such as medical imaging [157]. Therefore, introducing external constraints to self-supervised learning methods has been considered as a competitive strategy to improve the self-supervised learning.

We summarized recent breakthrough of self-supervised learning in medical imaging with different modalities and the corresponding self-supervised learning performance and supervised performance in Table 2.2. From Table 2.2, it is evident that self-supervised learning has become both widespread and increasingly effective in medical imaging, encompassing diverse modalities such as X-ray, CT, MRI, endoscopy, and pathology slides in multiple clinical domains, including radiology, gastroenterology, and ophthalmology. In the context of chest X-ray interpretation—an area often limited by the scarcity of detailed radiologist annotations—contrastive learning approaches such as MoCo [105] or SimCLR [47] have been employed to derive robust latent representations that, after minimal fine-tuning, can rival or surpass purely supervised models in tasks like pneumonia detection [237]. Similarly, researchers investigating MRI data for tumor segmentation or lesion classification have adopted patch-based inpainting techniques, where models learn to reconstruct missing regions of images as a means of pretraining. These methods have shown impressive adaptability, especially when faced with the high variability of MRI contrast and the heterogeneity of tumor appearances. Researchers have proposed a range of pretext tasks to extract meaningful representations from unlabeled data. Despite this variety of techniques, many of the published results indicate that the performance of self-supervised models, once fine-tuned with limited labeled data, frequently approaches

or even matches purely supervised benchmarks. Although there is currently no single method that consistently outperforms all others, these findings underscore the versatility of self-supervised learning in addressing a wide array of medical imaging tasks. By leveraging the abundance of unlabeled scans available in clinical practice, self-supervised models can narrow the gap to fully supervised methods, reducing reliance on large-scale human annotation.

Strategy	Year	Imaging modality	Clinical domain	framework	Metrics	Self. perf	Sup. perf	Ref.
Generative, Contrastive	2020	Chest X-ray	Radiology	Autoencoder, MoCo (modified), other	AUROC	0.917	0.861	[23]
Generative, Contrastive	2020	MRI	Radiology	Autoencoder, Multimodal contrastive	Accuracy	0.594	-	[114]
Contrastive, Innate Rel.	2020	Ultrasound	Obstetrics & Gynecology	Constrastive learning, other	F1	0.726	0.725	[128]
Contrastive, Innate Rel.	2021	Colonoscopy	Gastroenterology	Contrastive Learning, Augmentation Prediction, Patch Position Prediction	AUROC	0.972	-	[248]
Generative, Innate Rel., Self-prediction	2021	CT	Radiology	Autoencoder, patch pseudo label prediction, perturbed image restoration	AUROC	0.985	0.943	[100]
Self-prediction, Generative	2021	Endoscopy	Gastroenterology	GAN, Other	Accuracy	0.838	0.792	[63]
Contrastive, Innate Rel.	2021	Fundus Image	Ophthalmology	Multi-view contrastive learning, rotation prediction	AUROC	0.991	0.98	[166]
Generative, Contrastive	2021	MRI	Radiology	Autoencoder, SimCLR	AUROC	-	-	[81]
Contrastive, Generative	2021	MRI	Radiology	Longitudinal Neighborhood Embedding, Autoencoder	Accuracy	0.836	0.794	[204]
Generative, Contrastive	2021	Whole Slide Image	Pathology	CycleGAN, Contrastive Learning, Clustering	Accuracy	0.91	-	[144]
Generative, Contrastive	2021	Whole Slide Image	Pathology	Contrastive learning, other	Accuracy	0.914	0.844	[294]
Contrastive	2020	Chest X-ray	Radiology	MoCo	AUROC	0.953	0.949	[237]
Contrastive	2020	Chest X-ray	Radiology	Other	AUROC	0.893	0.879	[314]

Table 2.2: Current Self-supervised learning pipeline and the corresponding performance in supervised (Sup.) and self-supervised (Self.) learning

Strategy	Year	Imaging modality	Clinical domain	framework	Metrics	Self. perf	Sup. perf	Ref.
Contrastive	2020	Fundus Image	Ophthalmology	Multi-modal contrastive learning	AUROC	0.986	0.98	[167]
Contrastive	2020	MRI	Radiology	Mutual Information Maximization	AUROC	0.841	0.88	[82]
Contrastive	2020	Whole Slide Image	Ophthalmology	SimCLR	Accuracy	0.923	0.904	[196]
Generative, Contrastive	2020	Chest X-ray	Radiology	Autoencoder, MoCo (modified), other	AUROC	0.917	0.861	[23]
Generative, Contrastive	2020	MRI	Radiology	Autoencoder, Multi-modal contrastive	Accuracy	0.594	-	[196]
Contrastive, Innate Rel.	2020	Ultrasound	Obstetrics & Gynecology	Constrastive learning, other	F1	0.726	0.725	[128]
Contrastive, Innate Rel.	2021	Colonoscopy	Gastroenterology	Contrastive Learning, Augmentation Prediction, Patch Position Prediction	AUROC	0.972	-	[248]
Generative, Innate Rel., Self-prediction	2021	CT	Radiology	Autoencoder, patch pseudo label prediction, perturbed image restoration	AUROC	0.985	0.943	[100]
Self-prediction, Generative	2021	Endoscopy	Gastroenterology	GAN, Other	Accuracy	0.838	0.792	[63]
Contrastive, Innate Rel.	2021	Fundus Image	Ophthalmology	Multi-view contrastive learning, rotation prediction	AUROC	0.991	0.98	[166]
Generative, Contrastive	2021	MRI	Radiology	Autoencoder, SimCLR	AUROC	-	-	[81]
Contrastive, Generative	2021	MRI	Radiology	Longitudinal Neighborhood Embedding, Autoencoder	Accuracy	0.836	0.794	[204]

Table 2.3: Current Self-supervised learning pipeline and the corresponding performance in supervised (Sup.) and self-supervised (Self.) learning (CONTINUE)

Strategy	Year	Imaging modality	Clinical domain	framework	Metrics	Self. perf	Sup. perf	
Generative, Contrastive	2021	Whole Slide Image	Pathology	CycleGAN, Contrastive Learning, Clustering	Accuracy	0.91	-	[144]
Generative, Contrastive	2021	Whole Slide Image	Pathology	Contrastive learning, other	Accuracy	0.914	0.844	[294]
Contrastive	2020	Chest X-ray	Radiology	MoCo	AUROC	0.953	0.949	[237]
Contrastive	2020	CT	Radiology	Contrastive learning	Accuracy	0.963	0.775	[314]
Contrastive	2020	Fundus Image	Ophthalmology	Deep InfoMax	AUROC	0.835	0.833	[29]
Contrastive	2020	Whole Slide Image	Pathology	SimCLR	AUROC	0.963	0.726	[155]
Contrastive	2020	Whole Slide Image	Pathology	MoCo v2	AUROC	0.987	0.829	[67]
Contrastive	2020	Whole Slide Image	Pathology	SimCLR	F1	0.914	0.801	[58]
Contrastive	2021	Chest X-ray	Radiology	MoCo v2	-	-	-	[215]
Contrastive	2021	Chest X-ray	Radiology	Contrastive learning	AUROC	0.825	-	[172]
Contrastive	2021	Chest X-ray	Radiology	SimCLR	Sensitivity	0.936	0.907	[103]
Contrastive	2021	Chest X-ray	Radiology	SimCLR	AUROC	0.9	0.915	[158]
Contrastive	2021	Chest X-ray	Radiology	Contrastive Learning	AUROC	0.977	-	[88]
Contrastive	2021	Chest X-ray	Radiology	MoCo	Accuracy	0.916	0.796	[72]
Contrastive	2021	Chest X-ray	Radiology	BYOL	AUROC	0.988	0.95	[200]
Contrastive	2021	Chest X-ray	Radiology	SimCLR (modified)	AUROC	0.773	0.763	[7]
Contrastive	2021	Chest X-ray	Radiology	SimCLR, SwAV, DINO	AUROC	0.984	0.94	[250]
Contrastive	2021	Chest X-ray	Radiology	SimCLR (modified)	AUROC	0.889	0.84	[311]
Contrastive	2021	Chest X-ray	Radiology	MoCo (modified)	AUROC	0.906	0.858	[255]
Contrastive	2021	Chest X-ray	Radiology	Multimodal Contrastive, Text to Region Alignment	AUROC	0.932	0.91	[125]
Contrastive	2021	CT	Radiology	Contrastive learning (modified)	Accuracy	0.854	0.836	[71]
Contrastive	2021	CT	Radiology	SeLa-v2	AUROC	0.957	0.947	[122]
Contrastive	2021	CT	Radiology	SimSiam	AUROC	0.975	0.95	[13]
Contrastive	2021	Endoscopy	Gastroenterology	SimCLR	F1	0.9	-	[126]
Contrastive	2021	Fundus Image	Ophthalmology	MoCo, MSE	AUROC	0.966	0.941	[135]

Table 2.4: Current Self-supervised learning pipeline and the corresponding performance in supervised (Sup.) and self-supervised (Self.) learning (CONTINUE)

Modality	Year	2D/3D	Dataset	Label Strategy	Ref
MRI	2022	3D	BraTS 2020 [191]	Online	PLRS [246]
CT	2022	2D	COVID-19-CT-Seg [188]	Online	SSA-Net [268]
CT, MRI	2021	2D/3D	Pancreas CT [59], MR Endocardium [3], ACDC [18]	Online	CoraNet [227]
Microscope	2022	2D	CRAG [94]	Online	ECLR [309]
CT	2021	2D	UESTC-COVID-19 [257], COVID-19-CT-Seg [188]	Online	SECT [156]
CT	2022	2D	LiTS [20]	Label prop.	LoL-SSL [102]
X-ray, Dermoscopic	2021	2D	ISIC Skin [60], Chexpert [120]	Label prop.	NM-SSL [262]
X-ray	2023	2D	JSRT [228]	Label prop.	RPG [224]

Table 2.5: Current state-of-the-art performance on semi-supervised medical image segmentation based on pseudo labels. Label prop. refers to label propagation.

2.3 Semi-supervised Learning

Semi-supervised learning (Semi-SL) bridges the gap between supervised and unsupervised learning by leveraging both labeled and unlabeled data to train models. This approach is particularly valuable in domains where acquiring labeled data is expensive or labor-intensive, such as medical image segmentation.

Semi-SL occupies a unique middle ground, harmoniously integrating principles from both supervised and unsupervised learning paradigms. This integration is facilitated by a set of key assumptions that serve as theoretical underpinnings, guiding the design and implementation of Semi-SL algorithms. These fundamental assumptions—namely, the smoothness, cluster, manifold, and low density separation assumptions—play pivotal roles in establishing connections between the realms of supervised, unsupervised, and self-supervised learning.

The **smoothness assumption** posits that data points in close proximity are likely to share the same class label, suggesting that decision boundaries should traverse areas of low data density. In supervised learning, methods akin to k-nearest neighbors or kernel-based classifiers inherently exploit this principle by assigning similar labels to neighboring points. Similarly, in unsupervised settings, clustering algorithms like k-means cluster together proximate points under the implicit belief in local homogeneity. Self-supervised strategies, including contrastive learning or autoencoder architectures, further underscore this assumption by encouraging similar inputs to map to similar representations.

The **cluster assumption** asserts that data belonging to the same class tend to aggregate into distinct clusters. Within supervised learning, this assumption is har-

nessed when algorithms such as decision trees or support vector machines detect and leverage natural data clusters corresponding to classes. Conversely, in unsupervised learning, clustering techniques directly apply this premise to group unlabeled data, banking on the notion that clusters represent meaningful categories. In self-supervised learning, mechanisms that generate positive and negative sample pairs for contrastive learning mirror this cluster-based reasoning.

The **manifold assumption** suggests that high-dimensional datasets can be adequately represented on a lower-dimensional manifold. In supervised contexts, techniques like principal component analysis (PCA) [208] or linear discriminant analysis (LDA) [83] reduce dimensions under the premise of manifold existence. Supervised neural networks also learn these manifolds implicitly. Unsupervised learning algorithms, including t-distributed stochastic neighbor embedding (t-SNE) [111] and Isomap [245], are designed to uncover these low-dimensional structures. Self-supervised methods, exemplified by autoencoders [150], contribute to manifold discovery, thereby enriching feature representations before potential supervised fine-tuning.

Lastly, the **low density separation principle** advocates for decision boundaries in sparse data regions, separating classes by areas of minimal data presence. Supervised learning capitalizes on this through algorithms like support vector machines, which optimize margins between classes. Regularization techniques further promote simplicity in models, indirectly supporting this assumption. In unsupervised learning, density-based clustering algorithms like DBSCAN [79] operate on the principle of isolating clusters in low-density spaces. Self-supervised learning, by maintaining consistent representations under variations, fosters class separability in sparsely populated regions, indirectly echoing this principle.

Here we summarized the four different assumptions as below:

- **Smoothness assumption**: Data points that are close to each other are more likely to share the same label. This assumes that the decision boundary should lie in low-density regions.

- **Cluster Assumption**: Data points in the same cluster are likely to be of the same class. This assumes that the classes form well-separated clusters.

- **Manifold assumption**: The high-dimensional data lies on a lower-dimensional manifold. This assumes that the structure of the manifold can be learned from both labeled and unlabeled data.

Modality	Year	2D/3D	Dataset	Augmentation Strategy	Reference
X-ray	2019	2D	JSRT [228]	Elastic deformations	SemiTC [22]
US	2021	3D	UCLA [236]	Shadow augmentation	SCO-SSL [286]
Microscopy	2022	3D	Kasthuri15 [139], CREMI[2]	RandAug	SSN-RCL [118]
TOF-MRA	2022	3D	MIDAS [28]	RandAug	GCS [39]
CT, MRI	2020	3D	LA dataset [285], KiTS [109]	Random noise, dropout	DUW-SSL [270]
CT	2021	3D	BraTS 2019 [191], Pancreas CT [59]	Randomly cropped patches, multi-level pyramid predictions	URPC [187]
MRI	2021	3D	Longitudinal Multiple Sclerosis Dataset [34], ISLES 2015 [189], BraTS 2018 [191]	Multi-scale features	Mtans [41]
CT, MRI	2022	3D	BraTS 2019 [191], KiTS [109]	Different input images	CPCL [287]
CT, MRI	2021	3D	LGE-CMR datasets [138, 161]	Different domain inputs	AHDC [42]
MRI	2021	3D	LA dataset [285]	Random flipping, random rotating	UA-MT [302]
MRI	2021	3D	LA dataset [285]	Task-level consistency	SASSNet [163]
CT, MRI	2020	3D	LA dataset [285], Pancreas CT [59]	Task-level consistency	DTC [185]
MRI	2023	2D	ACDC dataset [18], PROMISE [170]	Task-level consistency	T-UncA [261]

Table 2.6: Semi-supervised medical image segmentation methods with consistency learning. RandomAug refers to perform different level of data augmentation for different views.

- **Low density separation**: The decision boundary should lie in low-density regions of the input space. This assumes that different classes are separated by low-density regions.

To develop an efficacious semi-supervised learning model, it is imperative to employ a judicious combination of underlying assumptions. particularly in the realm of medical image segmentation, where extracting global-local semantic features is of paramount importance [165]. The challenge lies in constructing such a model based on a well-informed integration of these assumptions.

Fusing the principles of the smoothness assumption and the cluster assumption constitutes an optimal strategy for addressing the intricacies of medical image segmentation. These two cornerstones of semi-supervised learning methodologies, while

[2] `https://cremi.org`

distinct, harmoniously integrate to facilitate enhanced data interpretation and utilization. The smoothness assumption encourages decision boundaries to pass through regions where data points are sparse, thereby minimizing the likelihood of misclassification for nearby points with unknown labels. This assumption is more focused on the idea of local similarity and the gradual change of class probabilities across the data landscape. However, cluster assumption deals with the global structure of the data, asserting that points belonging to the same class tend to form distinct, separable clusters in the feature space. It implies that classes occupy separate volumes in the data distribution, and there is a clear division between them. Unlike the smoothness assumption, which operates on a point-to-point local scale, the cluster assumption considers the aggregation of points into coherent groups, regardless of their immediate proximity to points from other classes. Thus, whereas the smoothness assumption operates at a granular, neighborhood-centric level, the cluster assumption undertakes a holistic perspective, aggregating data points into cohesive classifications based on their collective positioning in the higher-dimensional feature space – a synergy that is especially pertinent in the nuanced realm of medical image segmentation.

2.4 Limitations of Current Annotation-efficient Learning

Although annotation-efficient learning pipelines, such as semi-supervised learning (SSL) and selfsupervised learning (S4L), have made significant advancements in the field of medical imaging, several critical limitations hinder their effective application in real-world clinical settings. These limitations primarily revolve around challenges in representation learning, which is pivotal for the success of these methods. Theses limitations are summarized as below:

- **Data Quality & Heterogeneity**. Medical images often come from different hospitals, scanner vendors, and acquisition protocols, leading to domain shifts that can degrade model performance.
- **Noisy Labels & Class Imbalance**. Real-world clinical datasets can be highly imbalanced (e.g., few positive cases) and include noisy labels (e.g., ambiguous diagnoses, disagreements among radiologists).
- **Artifact-prone**. Imaging artifacts (e.g., motion, metal streaks) can distort subtle features, which misleads self-supervised or semi-supervised models trying to learn robust patterns.

- **Pretext vs. Downstream Task Mismatch**. Generic pretext tasks (e.g., rotation prediction, inpainting) might not capture the subtle anatomic or pathologic features needed for complex clinical tasks like lesion detection. For example, a model trained to predict image rotations on lung CT scans fails to learn the distinguishing features for ground-glass opacities, which are critical for early lung pathology detection.

- **High Intra-class Similarity & Subtle Inter-class Differences**. Subtle disease indicators or slight morphological changes can be critical for diagnosis but are not always emphasized in self-supervised objectives. For example, In retinal OCT scans, diabetic retinopathy may manifest as tiny microaneurysms. A general SSL approach (e.g., colorization) might overlook such small changes necessary for diagnosis.

The identified limitations of current annotation-efficient learning pipelines-encompassing semi-supervised learning (SSL) and self-supervised learning have prompted extensive research endeavors aimed at mitigating these challenges. The community has devoted considerable effort to addressing each specific issue, resulting in a multifaceted and systematic engineering approach to improving model performance in medical imaging. However, despite these advancements, the complexity of these solutions often leads to intricate systems that may still fall short of fully resolving the inherent limitations. In this context, developing powerful and reliable representations emerges as a promising strategy to comprehensively address multiple shortcomings simultaneously.

To address the challenge of **data heterogeneity and domain shifts**, researchers have employed various domain adaptation techniques. Methods such as Domain-Adversarial Neural Networks (DANN) [85] and CycleGAN-based approaches aim to align feature distributions across different domains, enabling models to generalize better to unseen data sources. For example, a study by [8] demonstrated that adversarial domain adaptation significantly improved the performance of models on cross-domain medical imaging tasks by learning domain-invariant features.

Addressing **noisy labels and class imbalance** has led to the development of robust training algorithms and specialized loss functions. Techniques like co-training, label smoothing, and focal loss have been proposed to mitigate the impact of mislabeled data and to ensure that minority classes are adequately represented during training. For instance, [169] introduced focal loss to address class imbalance in object detection, which has since been adapted for medical imaging applications with promising results [299].

To alleviate the effects of **imaging artifacts**, researchers have integrated preprocessing steps and artifact-aware learning strategies. Approaches such as image denoising using CNNs and incorporating artifact detection modules within the learning pipeline have shown efficacy in enhancing model robustness. An example is the work by [220], which utilized CNN-based denoising to improve the quality of MRI scans, thereby facilitating more accurate downstream diagnostic tasks.

Recognizing the mismatch between generic pretext tasks and clinical needs, researchers have started designing task-specific pretext objectives that better capture clinically relevant features. For example, instead of using rotation prediction, some studies have employed lesion segmentation or anomaly detection as pretext tasks to ensure that the learned representations are more aligned with diagnostic requirements. This alignment has been shown to enhance the performance of downstream tasks, as evidenced by the work of [251], who tailored pretext tasks for skin lesion classification, resulting in improved diagnostic accuracy.

To address high intra-class similarity and subtle inter-class differences, advanced feature extraction techniques and attention mechanisms have been incorporated into models. Attention-based models, such as the Attention U-Net, focus on diagnostically significant regions within images, thereby enhancing the model's ability to detect minute pathological changes. For instance, [202] demonstrated that incorporating attention gates into U-Net architectures significantly improved the segmentation of small and subtle lesions in medical images.

While these targeted strategies have yielded improvements, they often require the integration of multiple specialized components, leading to complex and cumbersome systems. For instance, a typical robust medical imaging pipeline may involve:

Given the complexity and partial effectiveness of current systematic approaches, focusing on developing powerful and reliable representations offers a compelling alternative. Enhanced representation learning aims to create unified feature spaces that inherently address multiple limitations, thereby simplifying the overall learning pipeline and improving model robustness and generalizability.

A well-designed representation can encapsulate domain invariance, noise robustness, and feature sensitivity within a single framework. The model will learn from comprehensive and discriminative features, generalize across domains, be resilient against artifacts, and effectively highlight subtle pathological differences with no need to have multiple specialized modules. Moreover, focusing on representation learning simplifies the learning pipeline by reducing the reliance on numerous auxiliary components. Consequently, this streamlines model development by expediting training

and deployment processes while ensuring robustness across varied datasets. This efficiency meets the needs of practical clinical applications where data variability is frequent.

2.5 Implicit Constraints

Building upon the identified limitations of current annotation-efficient learning pipelines and the advocacy for robust representation learning, this section explores a novel approach to integrate learned representations as implicit constraints within semi-supervised and self-supervised learning frameworks. This integration aims to address multiple shortcomings simultaneously by embedding meaningful structural and semantic information directly into the learning process, thereby enhancing model robustness, generalizability, and diagnostic accuracy.

Implicit constraints are conditions or properties that are not explicitly stated or enforced in the problem formulation or learning process, but are inherently desired or assumed to hold true for the solution or learned model [61]. Unlike explicit constraints that are directly incorporated into the optimization objective or model architecture, implicit constraints are indirectly imposed through the design of the learning algorithm, loss functions, or regularization terms.

- **Smoothness assumptions**: Encouraging the learned function or decision boundary to be smooth and continuous, without explicitly specifying the degree of smoothness.
- **Invariance or equivariance properties**: Assuming that the learned representations or model outputs should be invariant or equivariant to certain transformations of the input data, such as translations, rotations, or scaling.
- **Consistency regularization**: Enforcing consistent predictions or representations across different views, augmentations, or perturbations of the input data, without explicitly aligning the representations.
- **Physical or domain-specific laws**: Incorporating knowledge about the underlying physical system or domain, such as conservation laws or symmetries, into the learning process through carefully designed architectures or loss functions.

The main advantage of implicit constraints is that they allow for more flexible and expressive learning algorithms that can capture the desired properties or structures without the need for explicit annotation or supervision. By guiding the learning process towards solutions that satisfy these implicit constraints, the resulting models can be more robust, generalizable, and interpretable.

In the context of semi-supervised learning and self-supervised learning, implicit constraints play a crucial role in enabling the effective utilization of unlabeled data. They provide additional training signals and regularization that help the model extract meaningful representations and generate reliable pseudo-labels, even in the absence of explicit supervision. This leads to improved performance and generalization, especially when labeled data is scarce or expensive to obtain. For example, smoothness assumptions is one of the fundamental assumptions in semi-supervised learning, and the consistency regularization [234] is also an widely applied technique to improve the effectiveness of semi-supervised learning. Invariance or equivariance properties is the cornerstone of contrastive learning [47]. It is obvious that implicit constraints has been considered as an important component for annotation efficient learning.

Recent literature has introduced several forms of representation constraints in different learning paradigms. Locally Constrained Representations (LCR) [134] impose a soft constraint on state representations in reinforcement learning, assuming linear predictability of neighboring states [198]. In semi-supervised learning, pairwise constraints, such as must-link and cannot-link constraints, are employed to specify instance relationships and augment labeled data during classifier training. Self-supervised learning models utilize orthogonality constraints on neural network weights to prevent dimension collapse and learning of trivial scalar representations [78]. Knowledge graph embedding models incorporate constraints based on entity domain and range to ensure the semantic validity of corrupted triples used in training [152].

The successful of the Transformer [253] and its applications in visual [73, 104, 179] and natural language processing [26, 70, 210]. Transformer utilize an simple but effective structure, the self-attention to construct alignment among queries, keys and values where the keys and values are constructed with an affinity graph. Such design is highly scalable and effective for information extraction. We consider affinity graph would be an ideal option to construct effective and reliable implicit constraint representations in annotation efficient learning.

Beyond the framework of the Transformer, affinity graph is a graph structure where nodes represent data points (e.g. pixels, features, objects) and edges represent

the similarity or affinity between those data points. The edge weights capture the strength of the relationship. Affinity graphs are used to model the intrinsic structure and constraints present in the data. The application of affinity graphs have been widely applied in capturing the reliable representation between the implicit structure of the inputs. Learning an adaptive low-rank affinity graph allows capturing the subspace structure of data and improves clustering performance compared to using a fixed graph [281, 131, 300]. [235] proposed to use multiple affinity graph interaction network to fuse features from RGB and thermal images improves detection of salient objects, especially under poor illumination conditions. By capturing intrinsic structures and relationships, even in the absence of explicit supervision, affinity graph techniques enable more flexible and expressive learning that remains consistent with underlying domain knowledge and requirements.

In summary, affinity graphs serve as a powerful tool to encode structural constraints and guide representation learning. By introducing the affinity graph to extract the implicit constraints, we are able to construct reliable annotation efficient learning pipelines.

2.6 Preliminaries

As discussed in Section 2.5, affinity graph has been discussed as a reliable constraints to guide representation learning. In this section, we introduce the details of affinity graph. The affinity graph would be applied to support all the findings and further be discussed its application in semi-supervised learning, self-supervised learning. We would further setup a similar form of affinity graph but bridge different modalities to construct a multi-modal learning pipeline.

We follow the notation defined in the Vision Transformer (ViT) [74], where a given image $\mathbf{X} \in \mathcal{R}^{H \times W \times 3}$ is sliced into N patches. A special token that represents the specific class of this image is $\mathbf{X}_{[CLS]}$. The input is as follows:

$$z = [\mathbf{X}_{[CLS]}; \mathbf{X}_1; \ldots; \mathbf{X}_N] + \mathbf{E}_{pos}\,, \tag{2.1}$$

where $\mathbf{E}_{pos}$ is the positional embedding. The input z is then applied to Transformer [253] to learn the visual representation. The transformer layer is made up of multi-head self-attention (MSA) blocks. An attention function can be described as mapping a query ($\mathbf{Q}$) and a set of key($\mathbf{K}$)-value($\mathbf{V}$) pairs to an output, where the ($\mathbf{Q}$), ($\mathbf{K}$), ($\mathbf{V}$), and output are all vectors. The self-attention has been formulated in Eq. 2.2 where the $\frac{1}{\sqrt{d_k}}$ is the scaling factor. d_k is the key of dimension. MSA is made

up of a concatenation of self-attention, where each head h_i is in the Eq. 2.2. For an MSA with k number of heads is formulated in Eq. 2.3.

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Softmax(\frac{\mathbf{Q} \cdot \mathbf{K}^T}{\sqrt{d_k}})\mathbf{V} \tag{2.2}$$

$$MultiHead(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Concat(h_1, \cdots h_k) \tag{2.3}$$

The affinity graph is a mathematical representation of the relationships between input and output elements in a sequence. Explicitly, in multi-head attention, the affinity graph is represented as $\mathbf{Q} \cdot \mathbf{K}^T$, where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ denote the query, key, and value, respectively. The affinity between two elements is calculated using the dot product between their vector representations, which are learned representations extracted by the model of interest. The affinity values are then normalized via a softmax function, resulting in a probability distribution over the input elements. This probability distribution is used to weight the input elements when generating the output.

Given n queries $q_1, q_2, ..., q_n$, an adjacency matrix W of $n \times n$ can be used to represent the affinity matrix as follows:

$$W = \begin{bmatrix} W_{11} & W_{12} & \cdots & W_{1n} \\ W_{21} & W_{22} & \cdots & W_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ W_{n1} & W_{n2} & \cdots & W_{nn} \end{bmatrix}, \tag{2.4}$$

where $W_{ij} > 0 \in \mathbb{R}$, indicating the similarity between the query q_i and q_j. A higher value of W_{ij} suggests a closer relationship between q_i and q_j. Being an adjacency matrix of a graph, W is also called an "affinity graph".

Affinity graphs is aim to extract and leverage implicit constraints within data, providing a foundation for robust and scalable annotation-efficient learning. The mathematical formulation of affinity graphs allows for the effective modeling of relationships between data elements, making them highly suitable for a variety of tasks in machine learning. The unique strength of affinity graphs lies in their ability to establish meaningful alignments between two forms of data, whether within a single modality or across multiple modalities. This alignment capability is particularly beneficial for self-supervised and semi-supervised learning, where the lack of explicit supervision necessitates the discovery of latent structures and relationships. Affinity graphs can provide a structured representation of these relationships, enhancing the learning process by embedding domain-relevant constraints directly into the model.

To evaluate the utility and flexibility of affinity graphs, this thesis proposes their application in three distinct learning paradigms:

Semi-Supervised Learning. By embedding affinity graphs within semi-supervised frameworks, we aim to enhance the propagation of information from labeled to unlabeled data. The alignment between labeled and unlabeled examples facilitated by the affinity graph has the potential to mitigate issues such as noise in pseudo-labels and the oversimplification of decision boundaries.

Self-Supervised Learning. In self-supervised learning, affinity graphs can be used to impose structural constraints on representations learned through pretext tasks. By aligning features from different augmented views of the same data, affinity graphs can help the model learn more robust and transferable representations, overcoming limitations related to task-specific augmentation strategies.

Multi-Modal Learning. Multi-modal learning represents an extreme setup, where the alignment between disparate forms of data, such as images and text or gene and protein expression profiles, is crucial for success. The applicability of affinity graphs in this context serves as a litmus test for their reliability as a representation framework. If affinity graphs can effectively capture and model the complex relationships inherent in multi-modal datasets, their role as a versatile and powerful tool in annotation-efficient learning would be firmly established.

The overarching idea behind the affinity graph is its ability to align two forms—whether they represent augmented views in self-supervised learning, labeled and unlabeled samples in semi-supervised learning, or different modalities in multi-modal learning. This capability positions affinity graphs as a unifying representation that not only facilitates improved learning outcomes across diverse domains but also provides a consistent framework for addressing the inherent challenges of each paradigm.

Future chapters will delve into these applications, rigorously evaluating the performance and generalizability of affinity graphs in each learning paradigm. By systematically exploring their potential in increasingly complex setups, this work aims to establish affinity graphs as a cornerstone for robust, scalable, and reliable annotation-efficient learning.

2.7 Conclusions

In this Chapter, we present the transformative potential of annotation-efficient learning in leveraging vast amounts of unlabeled data for robust representation learning.

The exploration of semi-supervised learning revealed its reliance on implicit constraints, such as the smoothness and manifold assumptions, which enable the efficient propagation of limited labeled data across a larger dataset. The integration of methods like pseudo-labeling and consistency regularization was shown to bridge the gap between fully supervised and unsupervised approaches. However, issues such as noise in pseudo-labels and the difficulty in tuning hyperparameters for consistency were noted as significant challenges. Addressing these limitations requires further refinement in both theoretical frameworks and computational implementations.

In the context of self-supervised learning, the versatility of pretext tasks, such as contrastive learning and masked image modeling, was emphasized. These tasks harness the inherent structure of data to learn transferable representations. Despite the success of these methods, their dependency on large-scale datasets and sophisticated augmentation techniques presents computational and scalability hurdles. The use of external constraints, such as affinity graphs, emerged as a promising approach to address these limitations by embedding structural information directly into the learning process.

A critical insight from this chapter is the applicability of affinity graphs as a unifying representation across learning paradigms. By modeling relationships between data points, affinity graphs can capture spatial, semantic, and multi-modal interactions, thus addressing the inherent complexities of real-world data. This potential paves the way for innovative applications in medical imaging and beyond. By addressing these challenges and building on the strengths of current methodologies, annotation-efficient learning can continue to evolve, ultimately enabling the development of robust, scalable, and generalizable models for diverse applications.

Future chapters will delve into these applications, rigorously evaluating the performance and generalizability of affinity graphs in each learning paradigm. By systematically exploring their potential in increasingly complex setups, this work aims to establish affinity graphs as a cornerstone for robust, scalable, and reliable annotation-efficient learning.

Chapter 3

Affinity Graph in Self-supervised Learning

3.1 Introduction

Digital pathology has become an important form of supporting evidence in clinical trials [233]. Whole-slide images (WSIs) are high-resolution representations of microscopy slides that contain a tissue sample in a single digital copy. Therefore, WSIs can be extremely large in file size and contain billions of pixels, and thus exhibiting significantly greater detail than conventional microscopy images. So far, the collections of WSIs have significantly grown, which makes them a non-trivial asset in diagnostic pathology research. Several pioneering studies have already achieved a diagnostic performance equivalent to that of human specialists using supervised deep learning models on pathological pictures. However, the demand for exhaustively annotated datasets severely limits the scale and applicability of these investigations, since finely detailed expert annotations are difficult to obtain [199, 201].

A potential solution to the label scarcity of WSIs is self-supervised learning (SSL). SSL has recently achieved a consistent improvement in general computer vision tasks [9].

WSI presents unique challenges due to the extremely large input size of the images, as well as the high similarity between local details and the complexity of their texture features. To leverage these features, SSLP [157, 168] applies intra-invariance and inter-invariance constraints to utilize fine-grained features of WSI and significantly enhance the performance of SSL in WSI analysis tasks. Inspired by SSLP, we introduce *affinity graph constraint* (AGC) as an external constraint to use the fine-grained features of WSIs.

As one of successful SSL approaches on natural images, contrastive learning investigates the relationship of different views from the same patch of an image [105, 47]. Harnessing the power of Vision Transformers (ViTs) [74], DINO [35] and LOST [229] have shown success in SSL tasks, *i.e.* learning transferable representations. For example, without any segmentation mask supervision, the representations learned by DINO can be directly used for semantic segmentation tasks and demonstrate competitive performance with supervised methods. Motivated by the empirical findings discovered in DINO and LOST, we further explore the potential of Transformer-based architecture in SSL. Formally, the affinity graph is a mathematical representation that bridges an element-wise interaction between input and output elements in Transformer-based models [253]. Concretely, we propose a similarity constraint between two affinity graphs. Given a WSI image input, two augmented views are generated. An encoder and a following Transformer-layer will output two affinity graphs for two views, respectively. We aim to maximize the similarity between two affinity graphs and use this to regularze the SSL process to learn better reprsentations for WSIs. It is worth mentioning that the proposed method can be easily intergrated with common SSL frameworks (*e.g.* MoCo v2 [49] and SSLP [157])to pre-train common encoders (*e.g.* ResNet [107] and ViT [74]).

We evaluate the efficiency and robustness of AGC on three public benchmark WSI datasets and we find that incorporating the affinity graph constraint in the SSL training can efficiently boost the performance of SSL. There are three important findings. First, AGC significantly outperforms the state-of-the-art SSL methods in all datasets and these SSL methods significantly benefit from adding AGC to the training. Second, we use NPID [280] and MoCo v2 [105] as baselines to perform two sets of ablation studies. We find that AGC not only outperforms the existing constraints, but also complements the existing constraints by further improving the model performance, which establishes the new state of the art. Third, we show that AGC is robust under various convolutional or Transformer-based backbones.

The main contributions of this section are briefly summarized as follows:

- We introduce the constraints at the affinity graph level to perform SSL on WSI. This affinity graph constraint is compatible with existing SSL constraints.

- We introduce the random query selection based on multi-scale features to effectively utilize both global and local information. The scalable approach has been evaluated on the large-scale WSI dataset exceeding 1 TB in size.

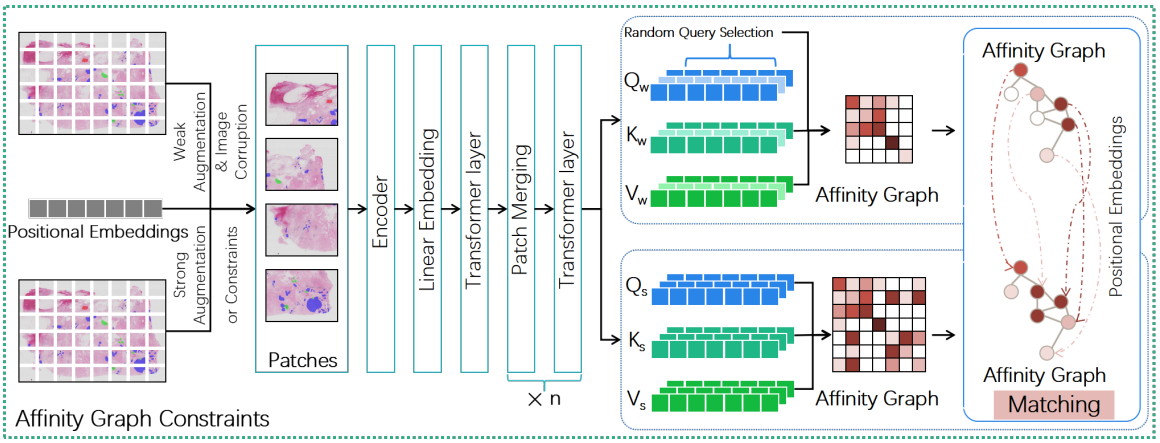

Figure 3.1: Overall structure of the affinity graph constraint (AGC) for SSL on WSIs.

- The experimental results on multiple WSI benchmarks show that the affinity graph constraint can further improve the performance of existing SSL. Our results demonstrate a significant improvement in accuracy at TCGA Lung Cancer Dataset, exceeding 10% compared to a baseline approach.

3.2 Methodology

We propose an affinity graph constraint based on self-supervised learning (SSL) to extract more detailed fine-grained information from whole slide images (WSIs). The structure diagram is shown in Fig. 3.1. Specifically, the approach first generates two different views of an input WSI through distinct data augmentation methods and divides them into non-overlapping patches. Subsequently, these patches, along with positional embeddings, are fed into an encoder (e.g., ResNet50) and Transformer layers. The process involves a combination of linear embedding, Transformer layers, and a series of patch mergings with the Transformer layer.

To scale up our proposed method, we do not directly perform further operations on the affinity graphs of both views. At the weak augmentation view, we randomly sample from a query $\mathbf{Q}_w = [q_{w,1}, q_{w,2}, \ldots, q_{w,n}]$, where n is the total number of queries. We compute the affinity graph $\mathbf{G_w}$ based on the sampled query $\mathbf{Q}_w$ and the corresponding key $\mathbf{K}_w$. We compute the affinity graph $\mathbf{G_s}$ for the strong augmentation view.

We then match two affinity graphs G_w and G_s based on the selected position embedding. As G_w is obtained from a subset of query-key setup, we cannot directly

match G_w and G_s. We map the G_w with the same position of query in G_s based on the corresponding positional embedding.

Note that our proposed method is compatible with prevailing self-supervised learning constraints, and does not suffer from any particular pre-processing perspective or stringent data augmentation requirements. The Transformer used here can be replaced with any other Transformer.

3.2.1 Affinity Graph Constraint

To extract fine-grained features (*c.f.* the features extracted by standard feature extractor, *e.g.* ResNet [107]), we introduce affinity graph constraints; the overall workflow is shown in Fig. 3.1. We argue that the affinity graph constraint can be applied to ViT [74], its variants, and the convolution structure. We integrate the visual encoder (*i.e.*, ResNet50) and the Transformer layer in an encoder for the self-supervised learning pipeline. Such design initially came to light in ViT, where the concluding feature map of ResNet50 serves as the ViT input.

The vast input size of WSI coupled with significant variations in pathology samples necessitates the consideration of scalability as a pivotal factor in the successful implementation of self-supervised learning in WSI. Formally, the affinity graph of the Transformer can be formulated as $\mathbf{Q} \cdot \mathbf{K}^T$. In this case, for an input length of n, the dimension of the affinity graph is an $n \times n$ matrix. Effectively optimizing the similarity between entities within such a high-dimensional structure is a formidable undertaking. Recent research has revealed that the inherent mechanism of the Transformer exhibits a proclivity towards a considerably sparse affinity graph in the context of self-supervised learning, where the Transformer operates as an encoder [55, 123]. A widely used approach to scale the self-supervised Transformer on a large-scale dataset is to apply sparse attention to the Transformer [123]. Nonetheless, it should be noted that WSI diverges significantly from the language modeling task, wherein global-local features occupy a prominent position in the pipeline. However, WSI deviates significantly from the language modeling task where the local and global features have taken a great place in the pipeline. Recent advancements in self-supervised learning for WSI have demonstrated the efficacy of exploiting spatial similaritites between local and global features [40, 157]. These methodologies strengthen spatial constraints between local and global features by maximizing their similarity. Typically, in such designs, the global features undergo substantial downsampling to enable the encoder to handle large inputs. Building upon these principles, our study introduces a strategy for direct query sampling, facilitating scalable self-supervised learning. Therefore,

to adapt the Transformer for extensive WSI datasets and to incorporate both local and global features effectively, our methodology aligns with these prior findings and implements random selection in the input queries.

The sampled query from the weak augmentation view can be formulated as $\mathbf{Q}_w = [q_{w,1}, q_{w,2}, \ldots, q_{w,n}]$, where n is the total number of queries. We compute the affinity graph $\mathbf{G_w}$ based on the sampled query $\mathbf{Q}_w$ and the corresponding key $\mathbf{K}_w$. It is worth mentioning that query selection has been proven to be effective in Deformable DETR [316].

To leverage both local and global features, we implement an extension of the affinity graph constraint within an alternative layer of the Transformer. Numerous self-supervised learning (SSL) and weakly-supervised learning pipelines for whole slide images (WSIs) have incorporated a pyramid structure to represent visual features [294, 177]. Typically, a WSI representation is obtained from patch-level embeddings through a global pooling operator at the end. The work by [45] further scales visual feature learning in self-supervised visual transformers with Giga-level pixel input scale by introducing a design similar to that in [177]. In subsequent experiments, we find that simply adopting this combination yields better performance in the WSI analysis task compared to others who construct complex densely connected constraints.

To further enhance the learning efficiency of the affinity graph constraint, we establish an affinity graph constraint between different Transformer layers, which is as follows:

$$\mathcal{L}_{\text{AGC}} = -\sum_{i=1}^{l} \frac{\mathbf{G}_{w,i} \cdot \mathbf{G}_{s,i}}{||\mathbf{G}_{w,i}||||\mathbf{G}_{s,i}||}, \tag{3.1}$$

where l is the number of layers, and $\mathbf{G}_{w,i}$ and $\mathbf{G}_{s,i}$ are the i_{th} layers of the two views of the self-supervised learning pipeline. L_{AGC} is used to minimize the negative cosine similarity loss [95, 50] between $\mathbf{G}_{w,i}$ and $\mathbf{G}_{s,i}$. The layer-wise computation of AGC is illustrated in Fig. 3.2 (a) and the visualization of each layer is shown in Fig 3.2 (b).

3.2.2 AGC with Existing Constraints

We can add AGC as a flexible module to improve the existing SSL pipelines. Since AGC is formed between two views, we can optimize $\mathcal{L}_{\text{AGC}}$ jointly with any contrastive loss or similarity loss without changing the model structure.

$$\mathcal{L}_{\text{Total}} = \lambda\mathcal{L}_{\text{AGC}} + (1-\lambda)\mathcal{L}_{\text{SSL}}, \tag{3.2}$$

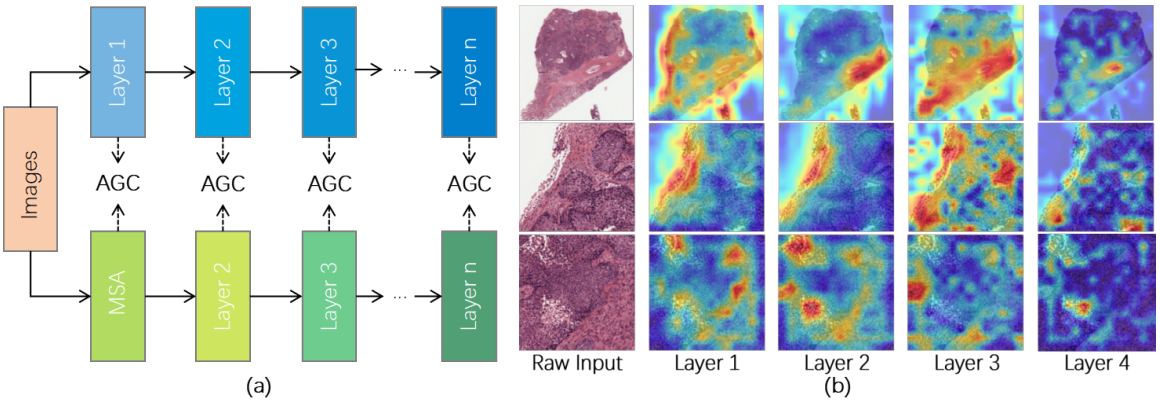

Figure 3.2: (a) Layer-wise computation of AGC. It has two streams. The left stream (the upper branch in Fig. 3.1) only uses the affinity graph $\mathbf{G_w}$ where queries are randomly selected. The right stream (the lower branch in Fig. 3.1) uses the affinity graph $\mathbf{G_s}$ with complete queries, where the first layer is represented by MSA. We can add AGC between each layer of the two streams, where the feature similarity is maximized. Each layer denotes multiple Transformer layers in ViT [74]. (b) AGC visualization across layers: Vertical sample size gradation with quadruple magnification and horizontal transition from raw input to heat maps of different layers, illustrating information acquisition for downstream tasks - a case from the TCGA Lung Cancer dataset dataset [273].

where $\lambda \in (0, 1)$ is a weighting hyperparameter.

3.3 Experiments

3.3.1 Datasets

Our experiments encompass three datasets of varying sizes. Specifically, CAMELYON 16 and NCT-CRC-HE-100K datasets are two widely used datasets for SSL on WSI. CAMELYON 16 and NCT-CRC-HE-100K datasets are relatively small in comparison with the TCGA Lung Cancer dataset. To our knowledge, this is the first instance of successful SSL on the TCGA Lung Cancer dataset due to its immense volume. It is noteworthy that datasets containing hundreds of WSI have the potential to provide abundant data samples for SSL because of the exceptionally large size of WSI images.

- **CAMELYON 16** [15] is a compilation of digitized WSIs focusing on lymph node tissue for the purpose of diagnosing breast cancer. It encompasses a total of 400 WSIs, with 270 designated for training and 130 for testing. Each image measures $100,000 \times 100,000$ pixels at a resolution of 0.25 microns per pixel. The training set comprises 159 normal slides and 111 malignant slides. While

CAMELYON 16 provides both pixel-level annotation and slide labels, only the slide labels are utilized for our training and testing purposes. A significant challenge presented by this dataset is that the majority of positive slides contain only small sections of malignancies dispersed throughout large tissue areas. The complete dataset demands over 700 GB of storage.

- **NCT-CRC-HE-100K** [140] is a collection of $100,000$ non-overlapping image patches from histological images. All images are stored in 224×224 pixels at a resolution of 0.5 microns per pixel. The dataset encompasses diverse tissue samples, including CRC original tumor slides and tumor tissue from CRC liver metastases; to improve the variety, normal tissue classes are supplemented with non-tumorous areas from gastrectomy specimens. The entire dataset requires over 10 GB of storage.

- **TCGA Lung Cancer dataset** [273] contains 534 LUAD and 512 LUSC slides, representing two subtypes of malignancy. Most of the slide is in $100,000 \times 100,000$ pixels. The dataset is generated as part of the TCGA project, a comprehensive effort to characterize the genomic changes associated with cancer. For this dataset, only slide-level labels are provided. Notably, tumour areas within this dataset are much larger than those in CAMELYON 16. The entire dataset requires over 1 TB of storage.

3.3.2 Experimental Setup

All SSL algorithms are trained without any supervised labels. For all convolutional structures, we use SGD with a momentum of 0.9, weight decay of $1e^{-4}$, and batch size of 32 per GPU on 8 NVIDIA A100 GPUs. For all Transformer structures, we use the AdamW [182] optimizer with an initial learning rate $5e^{-4}$. Unless otherwise stated, we use the same data augmentation methodologies as in [105] for all experiments in this paper. The backbone of all the experiments is ResNet50 [107]. For supervised learning, we use SGD with a base learning rate of 0.05 for 100 epochs. Without loss of generality, we use equal weights in Eq. (3.2), *i.e.* $\lambda = 0.5$ when adding AGC to existing constraints.

3.3.3 Evaluation Methods

The area under the curve (AUC) is the major performance metric across all trials since it is more comprehensive and less susceptible to class imbalance. Moreover, the

Method	CAMELYON		NCT-CRC		TCGA Lung Cancer	
	Acc	AUC	Acc	AUC	Acc	AUC
Supervised	79.6	84.2	95.9	93.7	83.3	**90.1**
NPID	73.1	78.7	93.4	93.1	70.0	72.4
LocalAgg	76.7	82.3	93.8	93.2	71.2	74.0
SSLP	82.9	85.7	95.2	93.6	75.6	82.4
MoCo v2	80.1	84.9	93.7	93.7	70.4	72.6
DenseCL	80.5	85.0	93.6	93.1	70.4	73.8
VICRegL	82.0	86.1	94.4	94.0	72.1	74.9
AGC (Ours)	81.5	85.5	95.5	94.1	81.2	83.3
NPID w/ AGC	77.5	80.4	94.6	94.0	78.4	83.2
SSLP w/ AGC	85.4	90.4	**97.5**	95.0	**85.6**	**90.1**
MoCo v2 w/AGC	82.5	86.1	95.7	95.1	81.9	88.5
DenseCL w/ AGC	82.5	86.1	95.5	95.1	80.4	87.2
VICRegL w/ AGC	**87.5**	**91.8**	96.8	**95.2**	84.5	89.4

Table 3.1: Experimental results (%) of AGC and the state-of-the-art baselines on the CAMELYON 16, NCT-CRC, and TCGA Lung Cancer datasets. **Bold** denotes the best result.

slide-level accuracy (Acc) is also considered in our evaluation. For the CAMELYON 16 experiments, we follow the same setup[1] as in [305], where the official training set is randomly divided by the 9 : 1 ratio. Each experiment is repeated five times, and the mean values of performance metrics are presented on the official test set, as well as the accompanying 95% confidence interval (CI-95). We evaluated the test set with the same test mask as in [305]. For the NCT-CRC-HE-100K experiments, we follow the setup in [157] and use a patch-level accuracy (Acc) as the evaluation metric. For the TCGA lung cancer experiments, a random partition is performed, allocating 70% of the data for training and 30% for validation.

3.3.4 Experimental Results

To study the effectiveness of the proposed affinity graph constraint (AGC), we conduct experiments on three publicly available datasets and compare AGC against the state-of-the-art self-supervised methods. In addition to a fully-supervised baseline, we have six self-supervised baselines:

[1] `https://github.com/hrzhang1123/DTFD-MIL`

- **NPID** [280] calculates direct distances between instances from the features in a non-parametric way.
- **LocalAgg** [317] trains an embedding function to maximize a metric of local aggregation.
- **SSLP** [157] introduces additional constraints to NPID to improve the model's performance.
- **MoCo v2** [49] modifies MoCo using an MLP projection head and more data augmentation of SimCLR. As MoCo v2 has been viewed as an improvement method based on SimCLR, we do not perform experiments on SimCLR.
- **DenseCL** [269] is a variant of MoCo v2 which emphasizes the optimization of a contrastive loss function for measuring the level of (dis)similarity between different image views at the pixel level.
- **VICRegL** [14] is another variant of MoCo v2 by two distinctively distorted renditions of a single visual input through symmetrical branches to learn global and local features simultaneously.

Quantitative experimental results are presented in Table 3.1.

Generally, as shown in Table 3.1, the self-supervised model outperforms all baselines after including the proposed AGC, demonstrating the usefulness of our AGC within the limits of picture information. While SSLP performs better than MoCo v2 and its variants on the current benchmark, it falls short. However, because SSLP [157] is tailored for WSI, the resulting performance enhancement is greater than that of the upgraded MoCo v2 solution, which is not at all tailored for WSI. This also demonstrates the validity and importance of the constraints designed for self-supervised learning in the WSI analyses.

In particular, we show that on all measures in the three datasets, AGC (8_{th} row) beats all SSL approaches. Second, we note that the SSL techniques significantly outperform the original SSL methods when AGC is included (i.e., from 9_{th} to 13_{th} row). This is due to the fact that AGC's features are more fine-grained than those acquired by other feature extractors at the level of the affinity map.

After that, we find that the results in the last row in Table 3.1 are significantly better than MoCo v2 with AGC, proving that combining AGC and any SSL method will significantly improve the effect of SSL methods. Besides, we observe that adding AGC significantly outperforms the best self-supervised methods on the TCGA Lung

Constraints				CAMELYON		NCT-CRC		TCGA Lung Cancer	
$\mathcal{L}_{Spatial}$	$\mathcal{L}_{Cluster}$	$\mathcal{N}_{semi}$	AGC	Acc	AUC	Acc	AUC	Acc	AUC
				73.1	78.7	93.4	93.1	70.0	72.4
			✓	77.5 (+4.1)	80.4 (+1.7)	94.6 (+1.2)	94.0 (+0.9)	78.4 (+ 8.4)	83.2 (+10.8)
✓				80.0	84.4	94.0	93.1	74.5	79.3
✓			✓	80.5 (+0.5)	85.3 (+0.9)	94.1 (+0.1)	93.1 (+0.0)	81.2 (+ 6.7)	88.1 (+ 8.8)
✓	✓			80.6	85.5	94.4	94.0	74.8	80.1
✓	✓		✓	82.3 (+1.7)	85.5 (+0.0)	97.4 (+3.0)	94.3 (+0.3)	83.8 (+ 9.0)	89.0 (+ 8.9)
✓	✓	✓		82.9	85.7	95.2	94.0	75.6	82.4
✓	✓	✓	✓	**85.4** (+2.5)	**90.4** (+4.7)	**97.5** (+2.3)	94.4 (+0.4)	**85.6** (+10.0)	**90.1** (+ 7.7)

Table 3.2: Ablation study results (%) of the proposed AGC and different constraints based on NPID [280] on three datasets. $\mathcal{L}_{Spatial}$ is the spatial neighbourhood invariance constraint in SSLP [157], $\mathcal{L}_{Cluster}$ is the clustering neighbourhood invariance constraint, and $\mathcal{N}_{semi}$ is the semi-hard negative mining constraint. **Bold** denotes the best result, and () indicates the improvement after adding AGC.

Cancer dataset. For example, the Acc metric of VICRegL with AGC on the TCGA dataset is 14.1% higher than that of MoCo v2, and 7.4% and 3.8% higher on the CAMELYON and the NCT-CAC dataset, respectively. This is because the TCGA Lung Cancer dataset is not only large in size but also has a relatively large lesion area. Finally, we observe that in the NCT-CRC and TCGA Lung Cancer datasets, SSLP with AGC achieves the best performance. We speculate that this is because the latter has better constraints on datasets with larger tumour regions.

Therefore, all the above phenomena not only prove that AGC is superior to the fully supervised baseline and state-of-the-art SSL methods, but also prove that AGC can be combined with any single SSL method or a combination of these SSL methods, and can significantly improve performances of these SSL methods.

3.3.5 Effect of AGC on Existing Methods

The proposed affinity graph constraint (AGC) directly exploits the properties of the Transformer and is therefore theoretically compatible with existing SSL methods that introduce different constraints on the pipeline. To further demonstrate that the proposed method can bring additional improvements to existing SSL methods, we perform a series of ablation studies based on two SSL methods (*i.e.* NPID [280] and MoCo v2 [49]). Based on these two methods, we study the effect of AGC and the other baseline constraints on NPID and MoCo v2.

3.3.6 Effect of AGC based on NPID

SSLP [157] introduces 3 effective constraints to NPID to improve the model performance on WSI analysis. In SSLP, the authors propose to introduce spatial neigh-

Constraints				CAMELYON		NCT-CRC		TCGA Lung Cancer	
$\mathcal{L}_q$	$\mathcal{L}_s$	$\mathcal{L}_d$	AGC	Acc	AUC	Acc	AUC	Acc	AUC
				80.1	84.9	93.7	93.7	70.4	72.6
			✓	84.5 (+4.4)	89.4 (+4.5)	95.7 (+2.0)	95.1 (+1.4)	81.9 (+11.5)	88.5 (+15.9)
✓				80.5	85.0	93.6	93.2	70.4	73.8
✓			✓	82.5 (+2.0)	86.1 (+1.1)	95.5 (+1.9)	95.1 (+1.9)	80.4 (+10.0)	87.2 (+14.4)
✓	✓			82.0	86.1	94.3	93.7	71.6	74.7
✓	✓		✓	87.3 (+5.3)	91.6 (+5.5)	95.9 (+1.6)	94.7 (+1.0)	80.2 (+ 8.6)	86.8 (+12.1)
✓	✓	✓		82.0	86.1	94.4	93.7	72.1	74.9
✓	✓	✓	✓	**87.5** (+5.5)	**91.7** (+5.6)	**96.8** (+2.4)	95.2 (+1.5)	**84.5** (+12.4)	**89.4** (+14.5)

Table 3.3: Ablation study results (%) of the proposed AGC and different constraints based on MoCo v2 [49] on three datasets. $\mathcal{L}_q$ are the dense constraints in DenseCL [269]. $\mathcal{L}_s$ and $\mathcal{L}_d$ are the locational-based constraints and feature-based in VICRegL [14]. **Bold** denotes the best result, and () indicates the improvement after adding AGC.

bourhood invariance constraints ($\mathcal{L}_{Spatial}$), clustering neighbourhood invariance constraints ($\mathcal{L}_{Cluster}$), and semi-hard negative mining constraints ($\mathcal{N}_{semi}$). We use the same notation as in [157], and the experimental results are in Table 3.2. We have the following findings. (i) The performance of the backbone improves with the addition of any type of constraint, indicating the effectiveness of these constraints. For example, compared to the Baseline (1_{st} row) with Baseline+AGC in the three datasets. Acc increases by 4.1%, 1.2%, and 8.4%, respectively. (ii) $\mathcal{N}_{semi}$ has the most significant impact, ensuring even sampling based on relative distances, correcting biases, and guaranteeing every data point's sampling opportunity [275]. A comparison between experimental findings in the 1_{st}, 2_{nd}, 3_{rd}, 5_{th}, 7_{th}, and 8_{th} rows that all metrics across all datasets improve with each additional constraint applied. In addition, by comparing the experimental results with and without AGC, we observe that all metrics are significantly improved after adding AGC. (iii) By adding AGC, the performance has been consistently improved, suggesting AGC's complementarity with other constraints.

3.3.7 Ablation Studies - Effect of AGC based on MoCo v2

MoCo v2 [49] is a renowned SSL baseline, excelling in WSI analysis. DenseCL [269] extends MoCo v2 by introducing dense constraints ($\mathcal{L}_q$), akin to our proposed dense matching. VICRegL [14] adds feature-based ($\mathcal{L}_d$) and locational-based ($\mathcal{L}_s$) constraints to MoCo v2. Our work integrates these constraints, and the experimental results are shown in Table 3.3. We have the following findings. (i) All constraints boost the performance of the baseline model. (ii) The improvements caused by $\mathcal{L}_s$ and $\mathcal{L}_d$ are comparable, while $\mathcal{L}_s$ has the minimal impact on global feature quality [14].

Backbone	CAMELYON		NCT-CRC	TCGA Lung Cancer	
	Acc	AUC	Acc	Acc	AUC
ResNet50 [107]	81.5	85.5	95.5	81.2	83.3
ConvNeXt-Tiny [180]	81.6	85.7	95.5	82.1	84.8
ViT-Small/16 [74]	74.5	79.9	92.8	69.5	70.4
SwinT-Tiny [179]	81.5	85.5	94.5	81.4	83.7

Table 3.4: Performance under different backbones.

Layer	1	2	3	4	Baseline (w/o AGC)
Acc	74.10	74.22	75.51	75.78	73.10
AUC	81.00	80.50	82.30	82.33	78.70

Table 3.5: Ablation study on the layer-wise effect of AGC on CAMELYON 16.

(iii) AGC complements all constraints, but its addition after $\mathcal{L}_s$ on the TCGA Lung Cancer dataset shows a dip, possibly due to parameter settings in [14] are intended to balance global and local feature learning capabilities.

The experiments also demonstrate that AGC can capture more fine-grained constraint information than existing constraints, and can be combined with existing constraints to further improve the ability of capturing information. Comparing results from rows 2_{nd}, 3_{rd}, 5_{th}, and 7_{th}, we find that adding AGC outperforms any existing constraint or their combinations, improves the performance of SSL methods. All the above analyses prove that adding AGC can further improve the complementarity between different constraints because AGC can capture more fine-grained constraint information than other constraints to improve the model's performance.

3.3.8 Ablation Studies - Robustness under Different Backbones

To prove that AGC can be used with different backbones, we conduct experiments on ConvNeXt-Tiny [180], ViT-Small/ 16 [74], and SwinT-Tiny [179]. The results are shown in Table 3.4, in conjunction with the experimental result of ResNet50. It is worth mentioning that, for ViT-based methods, we do not apply additional Transformer layers after the backbone ("Encoder" in Fig. 3.1) and the original layer-wise AGC is added at different blocks of ViT. We choose the small or tiny models because they have similar numbers of parameters. From Table 3.4, we can see that AGC can consistently achieve promising results with popular backbones.

3.3.9 Ablation Studies - Layer-wise Signicance of AGCs

We study the layer-wise effect of AGC on CAMELYON 16, where the SSL method is NPID, and the network backbone is ResNet50. The numbers are reported in Table 3.5. We see that AGC leads to a larger performance boost on a deeper layer.

3.4 Summary

In this Chapter, we proposed a new constraint on the affinity graph, called affinity graph constraint (AGC). Our research has extensively investigated the efficacy of the proposed AGC within the realm of self-supervised learning on Whole-slide images. The experiments show that including AGC in SSL is an effective way to capture fine-grained features and to improve the performance of the existing SSL methods in WSI analysis. Moreover, AGC can be combined with any single SSL method or combination of them. The robustness of AGC was also verified across various model architectures. It emphasized the adaptability and reliability of AGC in different settings. In-depth analyses of the layer-wise significance of AGC illuminated their pronounced impact on deeper layers, indicating the ability of AGC to capture intricate patterns in the data. The consistent improvements observed across datasets, backbones, and SSL methods underscore the broad applicability and effectiveness of AGC in advancing self-supervised learning for Whole-slide image analysis.

3.5 Limitations

The affinity graph constraint (AGC) establishes connections between different Transformer layers to improve the performance of the self-supervised learning pipeline. This inter-layer constraint is tightly integrated with the Transformer's multi-layer architecture, allowing for better information flow and feature refinement across layers. We found constructing affinity graph constraint is highly compatible with existing constraints in self-supervised learning and achieved great success in WSI analysis.

For **scalability** concern, the affinity graph constraint, while effective in improving performance, is currently tightly integrated with the Transformer architecture. This coupling may limit its applicability to other model structures. Exploring AGC beyond Transformers could potentially unlock new possibilities for improving model performance across different architectures. By generalizing AGC to work with other neural network structures, researchers could potentially address scalability issues in a broader range of models and applications. The tight coupling between AGC and

the Transformer structure may also make it challenging to modify the model for specific use cases. Researchers or practitioners looking to adapt the model for particular applications may find it difficult to disentangle the effects of AGC from the core Transformer functionality. This could limit the model's flexibility and adaptability to new domains or tasks.

Chapter 4

Affinity Graph Constraint on Semi-supervised Learning

4.1 Introduction

The precise delineation of medical imagery furnishes pivotal and discerning data for medical practitioners for suitable diagnostic evaluations, monitoring disease evolution, and formulating effective treatment strategies. Supervised methods based on deep learning have achieved a remarkable performance in medical image segmentation [44]. However, these methods largely benefit from extensive annotation datasets [272], and acquiring pixel-level annotations on a broad scale frequently demands a significant time investment and specialized knowledge, and entails substantial expenses. To alleviate the dependence on a large amount of annotated data, semi-supervised learning (SemiSL) and contrastive learning (CL) complement each other and are widely used in medical image segmentation. In detail, the pseudo-labels generated by SemiSL enhance the discriminative ability of CL by providing supplementary guidance for the metric learning method [5], while the crucial class discriminative feature learning of CL enhances the multi-class segmentation efficacy of SemiSL, allowing SemiSL to produce more ideal pseudo-labels [106, 37].

However, these methods have two obvious shortcomings: (i) **Relying on pretext tasks leads to a poor generalization ability**. First, these methods suffer from sampling biases and exacerbated class collision [57] that undermine the model's performance. Then, these methods do not account for substantial domain differences, resulting in a poor performance across datasets for models trained well in pretext tasks [98], which is particularly prevalent in medical images.(ii) **The lack of supervision signal leads to an overfitting problem**. Most of these methods use

regression, pixel-wise cross entropy or mean square error loss terms, and their variants to evaluate and generate ideal pseudo labels to assist the model in generating relatively accurate segmentation results. However, the loss functions have notable limitations, i.e., they cannot enforce intra-class compactness and inter-class separability [278, 164], thus limiting their full learning potential. Besides, there is a domain shift problem, i.e., these methods employ self-integration strategies and are designed for a singular dataset [277], which brings challenges to generalization across different domains.

The essence of solving the above problems lies in further utilizing the feature consistency in the manifold space with respect to the regional feature interconnections. Essentially, the effectiveness of SemiSL is based on the manifold assumption and the cluster assumption [295], which conceptualizes data points as components of low-dimensional manifolds embedded within a larger, high-dimensional space. Data points situated in the same feature space possess identical labels. However, when dealing with a limited amount of labeled data, the demarcation of cluster boundaries becomes ambiguous, which impedes the accurate delineation of the manifold's shape, leading to difficulties in correctly assigning labels to unlabeled data and thus hurting the quality of the learned feature representation. To mitigate these issues, the construction of a semantic graph presents a viable solution. By representing data points as nodes and their interconnections as edges based on feature similarities, a semantic graph encapsulates the intricate relationships within the data with explicitly representation. This structure not only enhances the understanding of manifold geometry but also provides a more nuanced view of cluster boundaries, even in scenarios with minimal labeled data. The integration of a semantic graph into SemiSL can be instrumental in exploiting the manifold's inherent structure, thereby facilitating more accurate label propagation in sparsely annotated environments.

Therefore, in this work, we propose an affinity-graph-guided semi-supervised patch-based CL framework that avoids the displayed pretext task. Specifically, the framework first uses an average-patch-entropy-driven new inter-patch semantic disparity mapping to select the positive and negative patches, and to improve sampling within our CL method, thus providing a powerful initial feature space by avoiding class conflicts. Then, the framework introduces affinity-graph supervision as an external constraint on the pseudo-label generated by the student and teacher networks, to enrich the supervision signal and enhance the discriminative ability required for accurate segmentation by exploiting the inherent structure of the data. Finally, the framework

proposes a new hard negative sampling method, making hard negative samples similar to positive samples but with different labels, and designs a loss function based on the affinity graph between positive and negative samples, to combine the advantages of SemiSL and CL, which can improve the quality of the learning representation and the model's generalization ability.

The main contributions of our proposed method are listed below:

- To alleviate the problem of the reliance on pretext and the overfitting problem caused by the lack of supervision signals, we propose an affinity-graph-guided semi-supervised contrastive learning framework (Semi-AGCL) to achieve high-precision medical image segmentation with extremely few annotations.

- We use an affinity graph as an external constraint on the generated pseudo labels with our affinity mass loss to minify class-discriminative features without any explicit training on pretext tasks, thereby demonstrating generalizability across multiple domains. Furthermore, we utilize a patch-based CL framework, wherein the selection of positive and negative patches is steered by an entropy-based metric, informed by the pseudo-labels garnered in the SemiSL setting. This approach averts class collision, i.e., the forceful and unguided contrasting of semantically akin instances within the CL framework.

- Following evaluation across three datasets from diverse domains, our method has demonstrated effectiveness, showcasing its generalizability and robustness even with a limited number of annotated samples.

4.2 Methodology

The overall structure of our proposed framework is presented in Figure 4.1. Given a labeled image set alongside its respective label set D_L and an unlabeled image set D_U, comprising N_L and N_U images (where $N_L << N_U$), respectively, we propose a patch-wise contrastive learning strategy with a teacher-student model to target the assimilate information from both D_L and D_U directed by pseudo-labels. In our framework, we first delineate the patch generation process, steered by the effective employment of (true or pseudo) labels; we then devise a new contrastive loss function incorporating an affinity graph between pseudo labels of student and teacher networks; subsequently, we construct an affinity graph between the positive and negative samples to guide the student network to learn from diverse distributions.

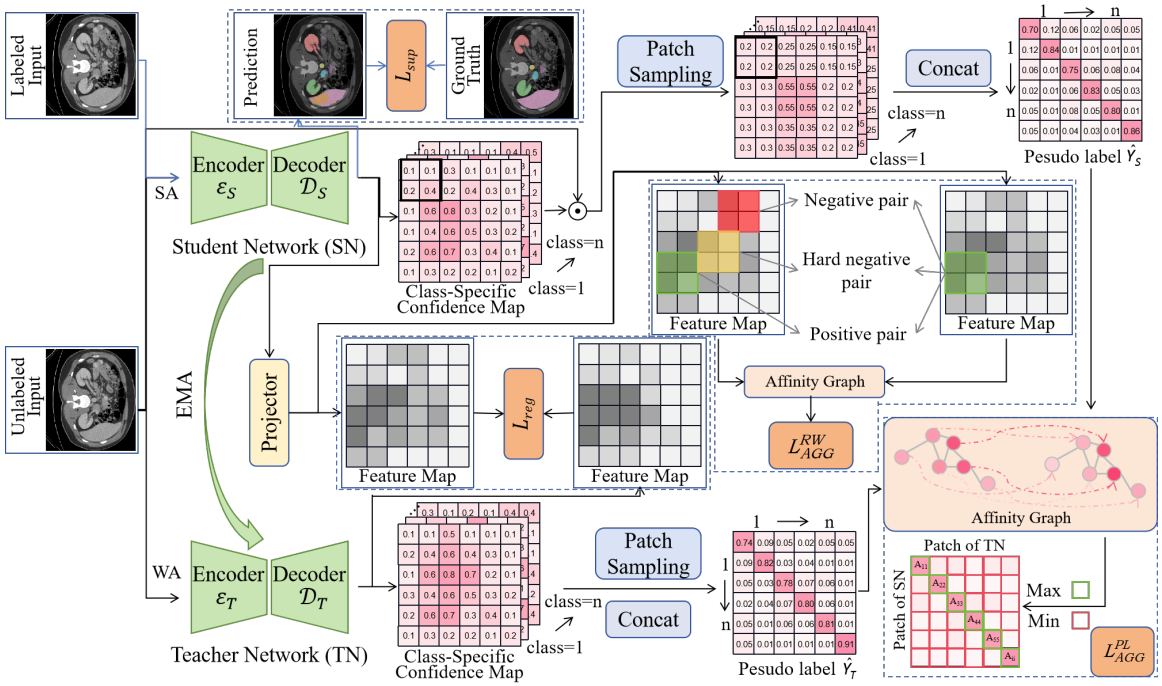

Figure 4.1: The proposed framework. For labeled data, we directly use the supervised loss $\mathcal{L}_{sup}$ to update the student network. For unlabeled data, we first slice the image into patches, then bridge an affinity graph loss $\mathcal{L}_{AGG}^{PL}$ between pseudo labels of student and teacher networks, and also design a new loss $\mathcal{L}_{AGG}^{RW}$ using the reweighting hard negative sample based on the edge of affinity graph. In the affinity-graph-based losses, we use low A_{ii} to construct a negative hard sample and try to pull positive pairs closer (increase A_{ii}) and push negative pairs away. Besides, the blue arrows use labeled data, and the rest (black arrows) are unlabeled data; we use a mixture of labeled and unlabeled data, so it is a semisupervised task rather than a self-supervised task. SA: Strong Augmentation, WA: Weak Augmentation.

4.2.1 Patch-wise Class-centric Sampling

Let $X_i \in \mathbb{R}^M$ denote the i^{th} image in a mini-batch, containing M pixels. The value of the m^{th} pixel in image X_i is denoted by $X_i(m)$. The key idea behind our patch-wise class-centric sampling is to select positive and negative patches for contrastive learning in an informed manner using the pseudo-labels. This prevents forceful contrasting of semantically similar patches, i.e., class collision. To achieve this, we first generate a class-specific confidence map C_i^k for each class $k \in \{1, 2, ..., K\}$, where $K(\geq 1)$ indicates the total number of classes. This confidence map reflects the likelihood of each pixel belonging to class k. By performing an element-wise product between C_i^k and the image X_i, which accentuates the regions relevant to class k, i.e., $X_i'^k = X_i \odot C_i^k$. Although the confidence map C_i^k is much more informative comparing to

the segmentation mask. $X_i'^k$ retains the image intensity/texture information along with masking information from the groundtruth and provides a richer representation for entropy calculation.

To sample informative patches, we compute an average patch entropy $Ent_{i,j}^k$ for each patch $P_{i,j}^k$ based on the pixel intensity values in the attended image $X_i'^k$. This entropy reflects three key types of information: confidence of belonging to class k, uncertainty regarding other classes, and intensity appearance from the original image X_i. A high entropy value indicates the patch likely contains the class k object but also has some confusion with other classes. The entropy thereby provides a richer metric for sampling, instead of simple random selection. This guided sampling focuses the learning on informative patches. Therefore, the average patch entropy allows robust, semantically meaningful sampling of positives and negatives to serve as a highly informative supervision signal to the self-supervised learning model. $Ent_{i,j}^k$ is formulated as follows:

$$\begin{aligned} Ent_{i,j}^k = -\frac{1}{|P_{i,j}^k|} \sum_{m \in P_{i,j}^k} & X_i'^k(m) \log(X_i'^k(m)) \\ & + (1 - X_i'^k(m)) \log(1 - X_i'^k(m)), \end{aligned} \tag{4.1}$$

where $X_i'^k(m)$ is the intensity value of pixel m in patch $P_{i,j}^k$. For an anchor patch of class k, patches with Top-n Ent_i^k values are **positives**, and the rest are **negatives**. This entropy-based sampling allows sampling positives that have high confidence for class k while also sampling challenging negatives from other classes. The patch appearance information also helps avoid ambiguity.

4.2.2 Affinity-Graph-Guided Contrastive Loss between Pseudo Labels

Unlike traditional semi-supervised learning frameworks, a patch-wise approach necessitates the incorporation of regional information to maximize the utility of the data. Consequently, it is our contention that not all patches should be regarded as equally significant. Therefore, we introduce an affinity graph to regularize patch importance by constructing fine-grained alignment in the outputs of student network ($\hat{\mathbf{Y}}_S$) and teacher network ($\hat{\mathbf{Y}}_T$). By directly encoding prediction vector similarities as edge weights between graph nodes, the discrete topology inherently captures the continuous semantic affinities that we intend to align. Concurrently, the graph Laplacian regularization enforces smoothness priors, forefending collapse into trivial solutions.

Maximizing the resultant diagonal trace impels convergence of the patch-wise pseudo-labels.

Specifically, we construct a patch-wise affinity graph $\mathbf{A} \in \mathbb{R}^{N \times N}$ between the pseudo-labels from the teacher network $\hat{\mathbf{y}}_t$ and student network $\hat{\mathbf{y}}_s$, where N is the number of patches. The edge weight A_{ij} is defined using a Gaussian kernel based on the L_2 distance between pseudo-label vectors $\hat{\mathbf{y}}_t^i$ and $\hat{\mathbf{y}}_s^j$:

$$A_{ij} = \exp\left(-\frac{|\hat{\mathbf{y}}_t^i - \hat{\mathbf{y}}_s^j|_2^2}{2\sigma^2}\right), \tag{4.2}$$

σ is an adaptive bandwidth parameter. We choose the Gaussian kernel, because it has strictly localized support, smooth variation, and efficient computability, which are theoretically well-founded for representing granular semantic relationships among samples. First, locality is imparted through the exponential term that precipitously decays affinity weight as distance in the embedding hyperspace grows. This realizes the expectation that closer samples exhibit greater semantic similarity and relatedness. It also guarantees gradual weight transitions with distance alterations, regulated by σ, thereby preventing abrupt changes. The modulating impact of σ further provides control over the rate of falloff and spatial scope of similar neighborhoods. Additionally, the Gaussian kernel satisfies constraints of radial symmetry and positive semi-definiteness suitable for modeling sample-wise relationships rather than discrete differences. Efficient computability facilitates constructing weighted graphs over large corpora encompassing tens of thousands of nodes.

The affinity graph construct provides a prudent approach here, as directly encoding prediction vector similarities as edge weights inherently captures the desired semantic affinities to align. Concurrently, the discrete topology is regulated through smoothness priors. In other words, we harness the diagonal entries A_{ii}, measuring self-similarities between teacher and student pseudo-labels on patch i. By maximizing the trace $\mathbb{E}[\text{tr}(\mathbf{A})]$, we promote convergence of the patch-wise student and teacher predictions. Simultaneously, we minimize the nuclear norm $\|\mathbf{A}\|_*$ via convex relaxation.

The nuclear norm serves as a convex lower bound on the intractable matrix rank function. We first perform this convex relaxation through singular value thresholding [31]. This decomposes $\mathbf{A}$ into $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T$ via singular value decomposition, where Σ contains the singular values σ_i. We then soft-threshold these singular values by an amount proportional to the subgradient of the nuclear norm, iteratively zeroing out unimportant dimensions. Thereby, minimizing $||\mathbf{A}||*$ serves as an efficient, tractable proxy for minimizing rank.

In summary, by excavating this salient low-rank alignment pattern between student and teacher outputs amidst noisy inconsequential variations, our approach safeguards the model from overfitting during contrastive learning. In effect, the convex relaxation extracts the most essential signals while filtering out extraneous dimensions. Explicitly, we get the affinity graph guided contrastive loss $\mathcal{L}_{\text{PL}}$ between the pseudo labels of $\hat{Y}_S$ and $\hat{Y}_T$ with its affinity graphs $\mathbf{A}$ as:

$$\mathcal{L}_{\text{AGG}}^{PL} = \sum_{i=1}^{N} \exp\left(-\frac{|\hat{\mathbf{y}}_t^i - \hat{\mathbf{y}}_s^i|_2^2}{2\sigma^2}\right) + \gamma||\mathbf{A}||*, \tag{4.3}$$

where γ balances the relative importance. By directly encoding alignment similarities and extracting low-rank structure, this loss function applied alongside standard supervised objectives aligns student and teacher representations robustly even with few labels. The affinity graph topology provides an interpretable, flexible mechanism for semi-supervised contrastive learning. We note that the quantities needed for the affinity graph in Equation 4.2 implementation are easily computed in parallel across examples within a batch using deep learning. We show a code snippet in Listing **??**.

4.2.3 Affinity-Graph-Guided Hard-Negative Reweighting

[218, 127] argue that the performance of contrastive learning could be improved by the incorporation of hard negative samples (i.e., samples y_i that are difficult to distinguish from an anchor x_i). In this context, rather than considering two arbitrary data points as negative pairs, these methods construct a negative pair from two random data points that are not too far from each other. The affinity graph constructed by Equation 4.2 also can meassure the hard negative samples under L_2 distance between $\hat{\mathbf{y}}_t^i$ and $\hat{\mathbf{y}}_s^i$. Therefore, utilizing the hard negative samples selected from Equation 4.2 could further improve the performance of the current framework.

In contrastive learning, we get the query $\mathbf{q}$ and the corresponding key $\mathbf{k}$ embeddings from the positive pair. We construct the query-key pairs $(\mathbf{q}, \mathbf{k})$ with the encoder-projection routine for the student networks, where we get $\mathbf{q} = \mathcal{H}(\mathcal{E}_S(P_i^k))$, $\mathcal{E}_S$ is the encoder of student networks, and $\mathcal{H}$ is the projection head. The key vector set $\mathcal{K}$ is formulated by amalgamating both positive and negative keys, represented as $\mathcal{K} = \mathcal{K}^+ \cup \mathcal{K}^-$, $\mathcal{K}^+$ consisting of positive keys $\mathbf{k}_i^+$ with the same distribution as $\mathbf{q}_i$, $\mathcal{K}^-$ consisting of negative samples $\mathbf{k}_i^-$. A widely recognized and effective loss function

utilized in contrastive learning is delineated as follows:

$$\mathcal{L}_{\mathbf{q},\mathbf{k}^+,Q} = -\log \frac{\exp(\frac{\mathbf{q}^T \cdot \mathbf{k}^+}{\tau})}{\exp(\frac{\mathbf{q}^T \cdot \mathbf{k}^+}{\tau}) + \sum_{\mathbf{n} \in Q} \exp(\frac{\mathbf{q}^T \cdot \mathbf{n}}{\tau})}, \tag{4.4}$$

where τ is a temperature parameter. The positive pair $(\mathbf{q}, \mathbf{k}^+)$ is contrasted with every feature n in the bank of negatives Q [106] with a fixed size K.

The log-likelihood function of Equation 4.4 is delineated based on the probability distribution, which emerges from the application of the softmax function to each $\mathbf{q}$. Let $p_{\mathbf{z}_i}$ as the correspondence probability between the query and feature $\mathbf{z}_i \in Z = Q \cup \mathbf{k}^+$. In this case, we get $p_{\mathbf{z}_i} = \frac{\exp \mathbf{q}^T \cdot \mathbf{z}_i / \tau}{\sum_{j=Z} \exp \mathbf{q}^T \cdot \mathbf{z}_j / \tau}$. The derivative of the loss function relative to the $\mathbf{q}$ is presented as:

$$\frac{\partial \mathcal{L}_{\mathbf{q},\mathbf{k}^+,Q}}{\partial \mathbf{q}} = -\frac{1}{\tau}((1 - p_k) \cdot \mathbf{k}^+ - \sum_{\mathbf{n} \in Q} p_n \cdot \mathbf{n}), \tag{4.5}$$

where p_k and p_n represent the probabilities of matching the key and negative feature, respectively. This formulation encapsulates the likelihoods of the feature vector $\mathbf{z}_i$ being aligned with either the key or the negative feature in the given context. It is trivial that the impact of both positive and negative logits on the loss function mirrors that observed in a $(K+1)$-way cross-entropy loss. In this scenario, the logit corresponding to the key is indicative of the latent class of the query. Additionally, all gradients in this framework are uniformly scaled by a factor of $\frac{1}{\tau}$. Therefore, sampling effective hard-negative samples from the memory bank Q has become the most effective way to improve the learning Equation 4.4.

We sample the hard negative samples based on two principles. First, the sampled hard negative instances must possess labels that are distinct from those of the anchor instances. This criterion ensures the maintenance of a fundamental dissimilarity at the label level. In our framework, for the same patch P_i^k with different views of augmentation, a hard sample/patch would expect to have the lowest similarity score between the two views. Then, the most advantageous hard negative samples are those which, according to the current state of the embedding, appear to be similar to the anchor. This perceived similarity, albeit misleading, renders these samples particularly challenging and, therefore, intrinsically valuable for the training process. Such samples, by virtue of their difficulty, provide a robust mechanism for enhancing the discriminative capability of the learned representations.

Based on the abovementioned principles, we set a dynamic threshold $\theta \in (0, 1)$ to identify hard negative samples with low edge weight A_{ii} based on the diagonal element

of **A** constructed by Equation 4.2. Motivated by [136] which proved the effectiveness of a mixture of the query and hard negative sample with a joint projection function. For the negative features $\mathbf{n} \in Q$ from a memory bank Q of size K, we get:

$$\mathbf{h}_k = \frac{A_{ii}\mathbf{n}_i + (1 - A_{ii})\mathbf{n}_j}{||A_{ii}\mathbf{n}_i + (1 - A_{ii})\mathbf{n}_j||_2}, \tag{4.6}$$

where $||\cdot||_2$ is the l_2-norm. $\mathbf{n}_i, \mathbf{n}_j \in Q$ are randomly chosen negative features from the set Q of the closest N negatives. H is the hard negative samples set where $\mathbf{h}_k \in H$. The A_{ii} represents the diagonal element of the affinity graph, and the suffix j represents other negative features in Q. We now get our affinity-graph-guided hard-negative reweighting on Equation 4.4 as:

$$\mathcal{L}_{\text{AGG}}^{RW} = -\log \frac{\exp(\frac{\mathbf{q}^T \cdot \mathbf{k}^+}{\tau})}{\exp(\frac{\mathbf{q}^T \cdot \mathbf{k}^+}{\tau}) + \sum_{\mathbf{h_k} \in H} \exp(\frac{\mathbf{q}^T \cdot \mathbf{h_k}}{\tau})}. \tag{4.7}$$

Since Equation 4.4 calculates the $L2$ distance between the positive sample and negative samples, so only the diagonal element of the affinity graph A_{ii} is used here.

Therefore, the overall loss function is as follows:

$$\mathcal{L}_{all} = \mathcal{L}_{sup} + \mathcal{L}_{reg} + \mathcal{L}_{AGG}^{PL} + \mathcal{L}_{AGG}^{RW}, \tag{4.8}$$

where $\mathcal{L}_{reg}$ is the cross entropy loss between the outputs of student and teacher networks.

4.3 Experiments

4.3.1 Datasets

The **LA dataset** [285] is an Atrial Segmentation Challenge dataset including 100 3D gadolinium-enhanced magnetic resonance image scans with labels.

The **ACDC dataset** [18] is a public segmentation dataset with four classes, i.e., background, right ventricle, left ventricle, and myocardium, containing 100 patients' scans.

The **CRAG dataset** is a Colorectal Adenocarcinoma Gland dataset [94] containning 213 H&E WS histopathological images taken with an OmnyxVL120 scanner. It has images with 20× objective magnification with a resolution of $0.55\mu m/pixel$ with each tile possessing full instance-level annotation.

We employ four metrics to evaluate the framework performance on all datasets, namely, Dice Similarity Score (DSC), Jaccard Index (Jaccard), Hausdorff Distance

95 (HD95), and Average Symmetric Distance (ASD) [37]. Given the outputs and the ground truths, DSC and Jaccard mainly evaluate the overlap value between them, HD95 measures the closest point distance between them, and ASD computes the average distance between their boundaries.

We follow the official setup for the training and testing split of all the datasets: both LA and ACDC are 80% and 20% for training and validation; and CARG is 80%, 10%, and 10% for training, validation, and testing.

4.3.2 Implementation Details

We conduct all experiments on a DGX A100 server with fixed random seeds. The model's convergence is achieved through the utilization of an ADAM optimizer, with the specifications of a batch size set at 16 and a learning rate designated at $1e-4$. The parameters τ and λ in Equation 4.5 are assigned values of 0.2 and 4, respectively, as guided by the precedent set in [32]. In Equation 4.3, we set the γ as -1 in all experiments. Within the scope of Section 4.2.1, which discusses n-nearest entropy-based sampling, the parameter n is determined through validation to hold the values of 0.999, 0.25, 0.2, and 20, respectively. For all the baselines, we follow the hyper-parameters defined in the original paper, and use Optuna [2] to tune the learning rate.

4.3.3 Main Results

The proposed framework is compared with the state-of-the-art CL- and SemiSL-based segmentation methods at different markers (i.e., 5% and 10%), that is, UA-MT [**?**], Double-UA [271], SASSNet [163], DTC [185], URPC [186], MC-Net [279], and SS-Net [278].

LA dataset. Results from other competitors are reported in the identical experimental setting in SS-Net [278] for fair comparisons. As shown in Table 4.3.3, our framework achieves the best performance on all four evaluation metrics, significantly outperforming other competitors. For settings with a 5% labeled ratio, we achieve significant improvements over Dice, Jaccard, HD95, and ASD (i.e., 3.11%, 2.90%, 2.19, and 0.20 over the second one, respectively). This is because of the loss function designed based on the affinity graph, which can directly utilize the inherent structure of the data and encapsulate the geometric and topological relationships between data, helping to enhance intra-class compactness and inter-class separability

Method	Scans used		Metrics			
	Labeled	Unlabeled	DSC↑	Jaccard↑	HD95↓	ASD↓
V-Net	5%	0	52.55	39.60	47.05	9.87
	10%	0	82.74	71.70	13.33	3.26
	100%	0	91.44	84.55	5.48	1.53
UA-MT	5%	95%	82.26	70.98	13.71	3.82
Double-UA			82.73	71.73	12.53	3.80
SASSNet			81.60	69.63	16.16	3.58
DTC			81.25	69.33	14.90	3.99
URPC			82.48	71.35	14.65	3.65
MC-Net			83.59	72.36	14.07	2.70
SS-Net			86.33	76.15	9.97	2.31
Semi-AGCL			**89.44**	**79.05**	**7.78**	**2.11**
UA-MT	10%	90%	87.79	78.39	8.68	2.12
Double-UA			88.53	78.83	8.42	2.10
SASSNet			87.54	78.05	9.84	2.59
DTC			87.51	78.17	8.23	2.36
URPC			86.92	77.03	11.13	2.28
MC-Net			87.66	78.25	10.03	1.82
SS-Net			88.55	79.62	7.49	1.90
Semi-AGCL			**90.33**	**82.53**	**6.68**	**1.78**

Table 4.1: Comparisons with state-of-the-art semi-supervised learning on LA dataset.

to improve the framework's ability to learn effective features, thereby improving the segmentation performance of medical images.

CARG dataset. We follow [226] to split the data into $80-10-10$ training, test, and validation ratios. Table 4.3.3 shows that our method achieves great improvement in the CARG dataset, and even at 10% labeled ratio, our method performs better segmentation than U-Net with 100% labeled ratio. This is mainly because (i) 2D slices can generate more combinations than combinations of 3D data. Therefore, knowledge from labeled data can be more fully transferred to unlabeled data, especially when the amount of labeled data is extremely small. This may be the reason why such a significant improvement is achieved when the labeled ratio is 5% compared to 10% (the ACDC dataset also has this advantage). And (ii) the CARG dataset has WS histopathological images, which contain additional texture features that can enrich the edge information for the affinity graph. This may be the reason why the improvement performance in the CARG dataset is more obvious than that of other datasets. Besides, as can be seen from Figure 4.2, our method can accurately seg-

Method	Scans used		Metrics			
	Labeled	Unlabeled	DSC↑	Jaccard↑	HD95↓	ASD↓
U-Net	5%	0	40.77	33.57	30.11	11.66
	10%	0	75.42	70.05	8.22	2.82
	100%	0	91.10	83.28	1.19	1.98
UA-MT	5%	95%	47.75	38.81	18.44	6.36
Double-UA			50.42	44.45	15.87	7.05
SASSNet			48.87	40.63	18.87	6.77
DTC			50.50	45.60	15.92	6.51
URPC			58.85	48.89	13.99	5.95
MC-Net			58.88	50.50	9.50	5.25
SS-Net			58.95	48.88	10.75	4.95
Semi-AGCL			**84.42**	**70.49**	**1.48**	**2.88**
UA-MT	10%	90%	81.46	71.42	1.48	2.23
Double-UA			87.01	77.58	1.50	2.63
SASSNet			86.43	76.98	1.67	2.66
DTC			84.13	75.24	1.83	2.73
URPC			83.36	71.79	1.61	2.33
MC-Net			83.30	72.11	1.61	2.13
SS-Net			83.40	70.25	1.88	2.58
Semi-AGCL			**91.93**	**83.37**	**1.08**	**1.76**

Table 4.2: Comparisons with state-of-the-art semi-supervised learning on the CARG dataset.

ment all objects, and the segmentation details are closer to ground truths than other baselines (see the red boxes).

ACDC dataset. Following SS-Net [278], we use 2D U-Net as the backbone, set the input patch size as 256×256 and the size of the zero-value region of mask $\mathcal{M}$ as 170×170. The batch size, pre-training iterations, and the self-training training iterations are set as 24, 10k and 30k, respectively. Table 4.3.3 shows the averaged performance of four-class segmentation results on ACDC dataset with 5% and 10% labeled ratios. It can be seen that our method is clearly optimal, e.g., with 5% labeled ratio, we obtain a huge performance improvement of up to 23.09% in DSC. The HD95 and ASD in the 10% labeled ratio have decreased compared to that of 5%, which may be because the loss function based on the affinity graph requires an appropriate increase in training iterations for the complex details of the edges. However, overall, our method significantly outperforms the competition on all metrics for all labeled ratios.

Method	Scans used		Metrics			
	Labeled	Unlabeled	DSC↑	Jaccard↑	HD95↓	ASD↓
U-Net	5%	0	47.82	37.01	31.16	12.66
	10%	0	78.22	68.05	9.33	2.70
	100%	0	91.44	84.55	4.30	1.00
UA-MT	5%	95%	46.04	35.97	20.08	7.75
Double-UA			56.88	45.53	22.70	6.26
SASSNet			57.77	46.14	20.05	6.06
DTC			56.90	45.66	23.33	7.38
URPC			55.58	43.66	13.66	3.78
MC-Net			62.85	52.29	7.62	2.33
SS-Net			65.83	55.38	6.67	2.28
Semi-AGCL			**88.92**	**78.84**	**1.90**	**0.66**
UA-MT	10%	90%	81.66	70.56	6.88	2.00
Double-UA			84.48	73.97	5.52	1.90
SASSNet			84.50	74.34	5.42	1.88
DTC			84.29	73.72	12.81	4.00
URPC			83.11	72.41	4.84	1.55
MC-Net			86.47	77.13	5.50	1.83
SS-Net			86.78	77.44	6.00	1.40
Semi-AGCL			**89.98**	**80.96**	**3.66**	**1.16**

Table 4.3: Comparisons with state-of-the-art semi-supervised learning on the ACDC dataset.

4.3.4 Computational Efficiency

Since most of semi-supervised learning methods are constructed based on complicated pipeline setup, we present the quantitative comparison of network's parameters and training time are listed in Table 4.4 on LA dataset. For all the semi-supervisd learning pipeline, our proposed Semi-AGCL achieved the second best training time over all existing semi-supervised learning methods. However, the performance of Semi-AGCL is much more better then DTC in LA datasets. Although our proposed method involves extensive matrix manipulation, it is highly parallelizable, providing an advantage in computation time. Furthermore, the patch-wise class-centric sampling method does not require parameter tuning. Despite the complexity of this computation, it did not substantially increase our computation time. As shown in Table 4.4, the training time difference between Semi-AGCL with patch-wise class-centric sampling and Semi-AGCL with class confidence (cls conf) sampling is only 1.8 minutes.

Table 4.4: Quantitative comparison of computational time between our methods and other semi-supervised learning methods on Left Atrium MRI dataset. We also present the Semi-AGCL without patch-wise class centric sampling (see Semi-AGCL w/cls conf). The Params is refer to the number of trainable parameters using the same backbone.

Method	Scaned Used		Computational Cost	
	Labeled	Unlabeled	Params (M)	Training time (mins)
VNet	5%	0	9.44	36.5
	100%	0	9.44	37.8
UA-MT	5%	95%	9.44	67.5
SASSNet			20.46	73.6
DTC			9.44	47.1
MC-Net			15.25	88.9
SS-Net			9.44	70.8
ACTION			10.14	471.9
ARCO			10.14	421.1
Semi-AGCL(Ours)			9.44	48.6
Semi-AGCL w/cls conf			9.44	46.8

4.3.5 Ablation Studies - Effectiveness of each module

In Table 4.3.5, under different labeling ratios, we compare different loss functions between pseudo-labels and loss functions based on different positive and negative sample selection methods, proving the effectiveness and optimality of the proposed loss functions (i.e., $\mathcal{L}_{AGG}^{PL}$ and $\mathcal{L}_{AGG}^{RW}$). First, for $\mathcal{L}_{AGG}^{PL}$, we find that the experimental results of $GK + L*$ used to calculate A_{ij} are not much different, but adding $\lambda||A||*$ used to calculate A_{ii}, the segmentation results have improved (the *random* of $\mathcal{L}_{AGG}^{RW}$ does not meet this conclusion, which may be because the randomness of the selection of samples is too high). Then, for $\mathcal{L}_{AGG}^{RW}$, the designed sampling method (A_{ii}) is better than the other two methods in most cases, but it incorporating $\lambda||A||*$ will significantly improve all metrics. These phenomena prove the optimality and complementarity of our loss function and sampling method based on the affinity graph. Finally, we find that the improvement under the labeling ratio of 5% is more obvious than that of 10%, further proving the effectiveness of our method in extreme data.

4.3.6 Ablation Studies - Effectiveness of the patch-wise class-centric sampling.

The comparison is conducted among three patch sampling methods: Cosine Similarity, Class Confidence, and our proposed entropy-based technique. Cosine similarity is

$\mathcal{L}_{AGG}^{PL}$			$\mathcal{L}_{AGG}^{RW}$			Label=5%				Label=10%			
GK+L2	GK+L1	$\lambda\|\|\mathbf{A}\|\|*$	A_{ii}	1:1	random	DSC↑	Jacard↑	HD95↓	ASD↓	DSC↑	Jacard↑	HD95↓	ASD↓
-	-	-	-	-	-	68.82	68.77	8.67	2.40	86.78	77.63	6.68	2.00
✓		✓				66.43	68.78	8.58	2.23	87.13	77.53	6.21	1.90
			✓			**88.92**	**78.84**	**1.90**	**0.66**	89.98	**80.96**	**3.66**	**1.16**
				✓		86.06	75.83	7.88	1.93	87.93	78.83	8.44	2.00
					✓	85.56	73.26	8.53	2.33	88.10	79.82	7.78	1.83
						67.78	68.88	8.54	2.89	87.78	77.56	6.70	2.10
			✓			71.71	72.53	7.74	2.11	87.78	77.58	6.71	2.15
				✓		70.71	71.33	8.21	2.10	87.72	77.50	5.93	1.98
					✓	69.83	70.88	8.11	2.00	87.78	77.47	6.32	1.98
	✓	✓				66.52	69.00	7.93	2.07	85.55	74.32	7.01	2.11
			✓			87.93	78.63	2.33	1.39	**90.12**	**80.96**	3.68	1.19
				✓		87.78	75.53	3.53	1.66	86.59	74.44	7.43	2.31
					✓	79.35	71.33	7.01	1.99	85.58	75.00	6.66	2.13
						66.52	69.10	7.53	2.01	85.75	74.42	6.83	1.89
			✓			83.10	73.11	5.56	1.66	85.72	74.58	5.38	1.66
				✓		82.33	70.31	6.87	1.77	85.22	74.46	6.82	1.89
					✓	82.33	70.52	7.05	1.86	85.40	74.23	6.50	1.84

Table 4.5: Comparisons with different setting of affinity graph loss on the ACDC dataset. GK+L* $= \sum_{i=1}^{N} \exp\left(-\frac{L*}{2\sigma^2}\right)$ in Eq. 4.3.

a prevalent metric for gauging similarity between two vectorized patches. The class confidence for a given patch $P_{i,j}^{k}$ requires computing the average patch confidence and subsequently classifying patches with analogous confidence values as positive and the rest as negative. We observe from Table **??** that the cosine similarity-centric patch-sampling from $X_i^{'k}$ is not satisfactory. Class confidence only achieves limited improvement relative to cosine similarity. This is mainly because the classification of positive and negative samples is not perpetually exclusive, and the misclassification rate of the above methods may increase, resulting in suboptimal performance. Our method achieves huge improvements, this is because our method advocates using entropy in $X_i^{'k}$ to sample positive and negative patches according to the class confidence map, and considers it as a more efficient measure for disparity mapping amidst patches.

4.4 Summary

To alleviate the problem that methods combining semi-supervised learning and contrastive learning rely on pretext tasks and insufficient supervision signals, we propose an affinity-graph-guided semi-supervised contrastive learning framework (Semi-AGCL) to achieve medical image segmentation without pretext under extremely few annotations. Semi-AGCL first designs an average-patch-entropy-driven inter-patch sampling method, which can provide a powerful initial feature space without pretext

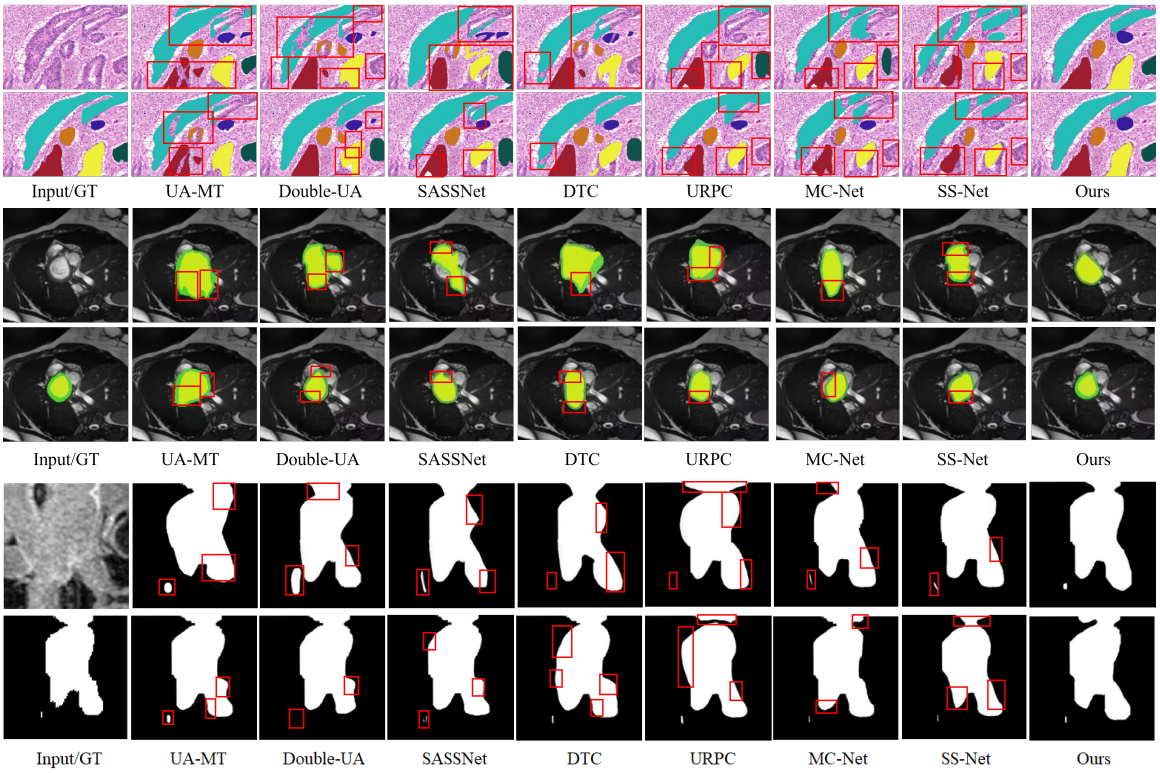

Figure 4.2: The visualization of the proposed framework and baselines on the CRAG, ACDC and LA dataset (from the top to down). The first and second rows are the segmentation results with labeled ratios of 5% and 10%, respectively. The red boxes indicate that our method outperforms other baselines. GT: Ground Truth.

tacks; and then it designs a new affinity-graph-based loss function between the student and teacher networks to improve the model's generalization ability. Evaluation on three medical segmentation datasets spanning multiple domains, our framework outperforms SOTA methods with minimal annotations, confirming its effectiveness and generalizability.

4.5 Discussion

In summary, while the Semi-AGCL method shows empirical effectiveness in leveraging unlabeled data for medical image segmentation, it has limitations in its reliance on pseudo-label quality and representativeness of the labeled and unlabeled data used for training. If the labeled data is not sufficiently diverse or representative of the full data distribution, it may limit the model's ability to generalize well to unseen cases. Careful curation of the training data is important for optimal performance.

Secondly, the representativeness of training data significantly influences performance. If the labeled data fails to capture the diversity of the full dataset, the learned representations may generalize poorly to unseen samples. For example, in a dataset with diverse tissue types, over-representing one type in the labeled set while under-representing others can lead to poor segmentation accuracy in under-represented categories.

While the entropy-based patch sampling addresses issues such as class collision, it risks introducing biases by overemphasizing high-entropy regions and under-sampling low-entropy but informative areas [313]. In retinal image analysis, for instance, high-entropy patches near the optic nerve may dominate, while low-entropy but diagnostically relevant regions, such as macular areas, could be underrepresented.

Scalability is another challenge. Despite the achievements that have been made in matrix operations and parallel computation, the construction and utilization of affinity graphs in high-dimensional datasets pose significant computational demands. For example, generating affinity graphs for 3D MRI scans of the brain, each comprising millions of voxels, can result in substantial computational overhead.

Recent advances suggest that affinity graph designs must adapt to dynamic, context-dependent relationships within data. The current implementation assumes static graphs, which may not fully capture temporal or evolving relationships in datasets such as time-series medical imaging or real-time diagnostic applications. For example, in fetal ultrasound imaging, dynamic graphs could better account for growth and positional changes over time.

To address these limitations, automated affinity graph initialization using meta-learning or self-supervised graph construction could reduce the reliance on manual tuning and domain-specific preprocessing. For instance, automatic graph construction for pathology images could dynamically adapt to varying staining techniques without requiring extensive preprocessing. Extending the affinity graph concept to cross-domain applications, such as integrating imaging data with electronic health records or genomics, could unlock new possibilities in multi-modal learning frameworks. For example, constructing graphs that connect imaging features with genomic markers could improve precision medicine approaches for cancer treatment.

Real-time scalability can be improved by employing sparse graph representations or dynamic graph pruning techniques to address computational bottlenecks. For instance, using sparse graphs in intraoperative imaging could enable real-time surgical guidance systems [117]. Developing mechanisms to visualize and interpret the learned graph structures, possibly through explainable AI approaches, could enhance the

usability and trustworthiness of affinity graph-based systems in sensitive domains like healthcare. For example, a visual explanation of graph edges and weights could help radiologists understand how specific features influence diagnostic outcomes.

By addressing these limitations and capitalizing on emerging trends, the potential of affinity graphs in semi-supervised learning and beyond can be fully realized.

Chapter 5

Affinity Graph in Multi-modal Learning

5.1 Introduction

The advent of single-cell technologies has revolutionized our understanding of cellular heterogeneity and function, providing unprecedented resolution to explore the complex biological systems at the individual cell level. Among these technologies, multimodal single-cell analysis, which simultaneously measures multiple types of biological information from the same individual cells, has emerged as a powerful tool for dissecting the intricate relationships between different biological modalities, such as gene expression [240], protein levels, and chromatin accessibility [33, 46]. However, the integration and interpretation of such high-dimensional and heterogeneous data pose significant computational challenges, necessitating the development of novel methodologies.

Modality prediction is one of the most critical tasks in multimodal single-cell analysis. The goal of modality prediction is to infer one type of biological information (*e.g.*, DNA methylation status) based on another type (*e.g.*, protein expression). This task is particularly important as it permits the imputation of missing or unmeasured modalities, thereby enriching the multimodal data and facilitating a more comprehensive understanding of cellular states and functions [181]. So far, deep learning has demonstrated remarkable success in the task of single-cell analysis [77]. As shown in Figure 5.1, in the context of modality prediction, most existing deep learning approaches construct joint representations of cells, genes, and proteins, and then leverage these representations to perform downstream tasks. Most contemporary techniques, including BABEL [276], CMAE [293], scMoGNN [274], scMM [194], and scGPT [65], utilize cell embeddings or representations as an intermediary step

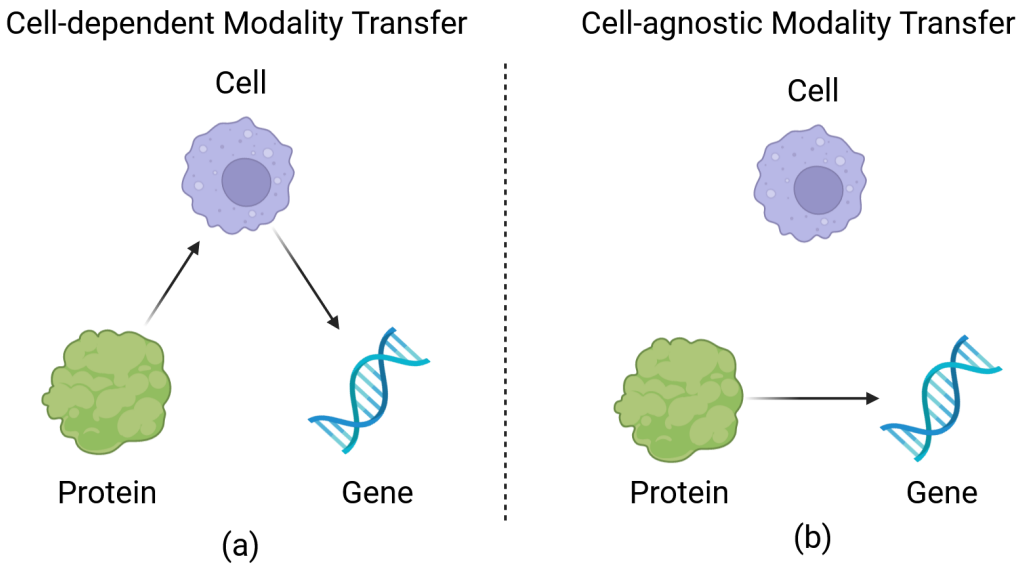

Figure 5.1: Dichotomy of modality transfer between protein and gene. (a) Cell-dependent methods: CMAE [293], scMM [194], scMoGNN [274], scFormer [64]. (b) Cell-agnostic methods: BABEL [276] and Gradient Boosted Decision Trees (GBDT) [274]. Our scAGFormer bridges between gene and protein expression without cell representations thus is a cell-agnostic method.

within their overarching workflow. These representation-based methods, while powerful, have their limitations. Specifically, they tend to underperform when the given cell type deviates from the dataset's extant cell base distribution, a scenario frequently encountered in biological research given the immense heterogeneity and fluidity of cellular systems.

In addressing this concern, there is an imperative requirement for a method capable of predicting modalities independent of cell embeddings. We present a cell-agnostic modality prediction approach. Classical statistical learning techniques like logistic regression might serve as potential candidates, given their ability to harness statistical features directly for prediction. However, these methods often struggle with out-of-distribution (OOD) instances, which are prevalent in the biological field due to the aforementioned cellular diversity and dynamics. Moreover, these statistical methods are incapable of capturing complex interactions between different representations [17], which are crucial to understanding correlations in single-cell analysis [110].

Statistical feature learning methods, such as Gradient Boosting Decision Trees (GBDT) [84], primarily focus on statistical manipulations of input features for predictive modeling. Representation learning methods like autoencoder [150] or Transformer [253] architectures generate high-level feature representations through neural

networks. Informed by these seminal constructs, we recognize an urgent requirement for a hybrid approach. Statistical feature learning methods excel in capturing the intrinsic features of the data through feature engineering and selection, thereby delivering models that are often interpretable and computationally efficient. Moreover, statistical feature learning methods are completely cell-embedding-free solutions. On the other hand, representation learning methods specialize in learning complex and hierarchical representations from the raw data, which are adept at capturing nonlinear relationships but may sacrifice interpretability. The integration of statistical feature learning and representation learning methods shall provide a feasible solution for the multimodal single-cell analysis task.

Therefore, we introduce an innovative cell-agnostic multimodality prediction model leveraging on statistical and representational features. We present a dual-module architecture with a joint optimization strategy. The proposed framework consists of a statistical feature learning module and a representational feature learning module. The statistical feature learning module takes the differential gene expression as the input [62] and is implemented through a Transformer-based model with regression tasks. Simultaneously, the representational feature learning module is devised to establish an affinity graph. The affinity graph plays a critical role in enabling the model to discern and learn from the interactions between the source and target modalities. This process involves the transformation of the gene data using gene2vec [75], alongside the application of protein2vec [86] to the protein data.

We further design a novel Interaction Extraction Module (IEM) that can extract the interactions between features in both statistical feature learning module and representational feature learning module.

We evaluate our model on four benchmark multimodal modality prediction tasks over two public datasets and achieve state-of-the-art performance. The proposed method outperforms other baseline methods by a substantial margin. This work not only advances the field of modality prediction in multimodal single-cell analysis but also provides a new perspective on how to effectively integrate and interpret multimodal single-cell data.

Our contributions can be summarized as follows:

- This Chapter introduces a cell-agnostic framework for multimodal single-cell analysis, called single-cell Affinity Graph transFormer (scAGFormer). This methodology can extract high-order feature interactions between modalities without the need for cell representations, marking a departure from traditional cell embedding techniques.

- The proposed framework consists of two novel modules, a statistical feature learning module can perform accurate prediction based on a multi-step feature selection, and a representational feature learning module that leverage an affinity graph prior to learn the interactions between the source and target modalities.
- The proposed framework achieves superior performance across different prediction tasks against all the existing GNN-based or GPT-based neural models as well as the classical statistical learning techniques.

5.2 Related Works

In the current research, multimodal single-cell analysis techniques can be categorized into matrix factorization, statistical approaches, autoencoder-derived methods, and GNN-based methodologies. Subsequent sections provide a concise overview of these techniques.

5.2.1 Multimodal Data Integration

The prevailing literature on multimodal data integration can be primarily categorized into two distinct categories: 1) matrix factorization or statistical methods and 2) autoencoder-based methodologies. Matrix factorization or statistical techniques have been extensively applied in research studies such as [76, 129, 238]. These methods provide a valuable approach to handling the complexity of multimodal data, but they often neglect the possible non-linear relationships between different modalities. Methods based on auto-encoders such as [90, 276] have gained momentum due to their inherent ability to capture nonlinearities in data. In particular, BABEL [276] employs an autoencoder architecture comprising dual encoders and decoders. This structure permits BABEL to handle a single modality at a time and infer the counterpart using a combination of reconstruction loss and cross-modality loss, providing a powerful and flexible framework for integrating multimodal data. Cobolt [90] obtains a joint embedding by applying a modified form of the multimodal variational autoencoder (MVAE [297]). By leveraging this innovative adaptation of the MVAE, Cobolt presents a unique approach to multimodal data integration, capitalizing on the power of autoencoders to capture complex non-linear relationships between different modalities. scMM [194] exploits a mixture-of-expert framework with GANs [91]

thereby achieving end-to-end learning by directly modeling the raw counts of each modality.

These integration techniques predominantly concentrate on deciphering the mapping relationship between the embeddings of the source and target modalities. They typically achieve this either through a direct connection between the source and target modalities or via intermediary representations like cell embedding, subsequently mapping directly to the target modality's embedding. In contrast, our approach utilizes an affinity graph to directly establish the relationship between the embeddings of the source and target modalities. This relationship is further assimilated through joint optimization with the statistical feature learning module. Notably, since this module operates directly on the source modality's statistical features, it eschews the need for intermediate representations.

5.2.2 GNNs in Single-Cell Analysis

In the quest to encapsulate the intricate biological interactions that occur at the molecular and cellular levels, a growing body of literature has begun to demonstrate the efficacy of Graph Neural Networks (GNNs) [260, 96] and Transformer frameworks [64, 65, 290] within the domain of single cell analysis. In general, GNN-based solutions focus on building the interaction between gene and cell representation, and transformer-based solutions learn from the implicate representation of gene information from a string format [290]. ScGNN [260] modeling the cell-cell interaction by incorporating GNN with multimodal autoencoders. Specifically, scGNN builds a cell graph by capturing cell-type-specific regulatory signals and utilizes a Left Truncated Mixture Gaussian model [256] for scRNA-Seq data analysis. The utilization of transformer architectures, in particular, has been heralded for their capacity to capture long-range dependencies across single-cell datasets, thus providing a more comprehensive global perspective on cellular interactions [64]. [65] propose to use a GPT [210] model to perform unsupervised learning on large-scale single cell data to obtain a foundation model [21] in single cell analysis area. Such a language model was built with embeddings on the genes and cells and trained by performing generative modeling on diverse single-cell data.

The efficacy of Graph Neural Network (GNN)-based multimodal single-cell prediction hinges on the premise that GNNs can achieve an optimal mapping relationship among cell modality, source modality, and target modality. This mapping is realized through the interactions between the learned features. While GNNs excel as tools for

representing interaction correlations, they are less adept at statistical feature learning. Given that single-cell data fundamentally comprises statistical features, this limitation of GNNs may hinder their ability to fully leverage these features, potentially diminishing the accuracy of the final predictions. Contrary to the conventional approach, we posit that statistical features hold significant importance in single-cell tasks. Consequently, we have developed a joint optimization framework for multimodal single-cell prediction. This framework is tailored to simultaneously harness the power of statistical feature learning and representational feature learning, ensuring the comprehensive utilization of both feature types.

5.3 Problem Formulation

We follow the setting of the NeurIPS multimodal single-cell integration competition of the year 2021 [184] and 2022 [66] and collect the joint measurements of gene expression and surface protein levels datasets from the competitions. Both datasets contain the raw counts, which represent the number of reads per gene per cell, as well as the normalized counts. Both tasks work on the same objective, where to perform modality prediction to discern the correlation between two individual cell profiles and subsequently offer a probability distribution pertaining to these conjectures. More formally, this task can be articulated as follows:

Given source modality $\mathbf{M} \in \mathbb{R}^{N \times k_1}$ *and goal modality* $\mathbf{M}' \in \mathbb{R}^{N \times k_2}$ *where* N *is the number of instances (for example, the number of cells), and* k_i *refers to the number of features given in the* i*th modality. The objective is to find a valid mapping function* f_θ *parameterized by* θ *where* $\mathbf{M}' = f_\theta(\mathbf{M})$.

The objective is to perform modality prediction among mRNA data (GEX), DNA data (ATAC) and protein data (ADT). For NeurIPS multimodal single-cell integration competition in 2021[1], the modality prediction task is built between GEX→ADT, ADT→GEX, GEX→ATAC and ATAC→GEX respectively. Subsequently, for the NeurIPS 2022 competition[2], the challenge was curated to involve data measurements obtained from two distinct single-cell assays, namely *CITEseq* and *Multiome*. Within this context, the *CITEseq* task aimed at the modality prediction between GEX and ADT, while the *Multiome* task focused on the modality prediction between ATAC and GEX. We discuss more details about the datasets in Sec. 5.5.

[1] `https://openproblems.bio/events/2021-09_neurips/`
[2] `https://openproblems.bio/events/2022-08_neurips/`

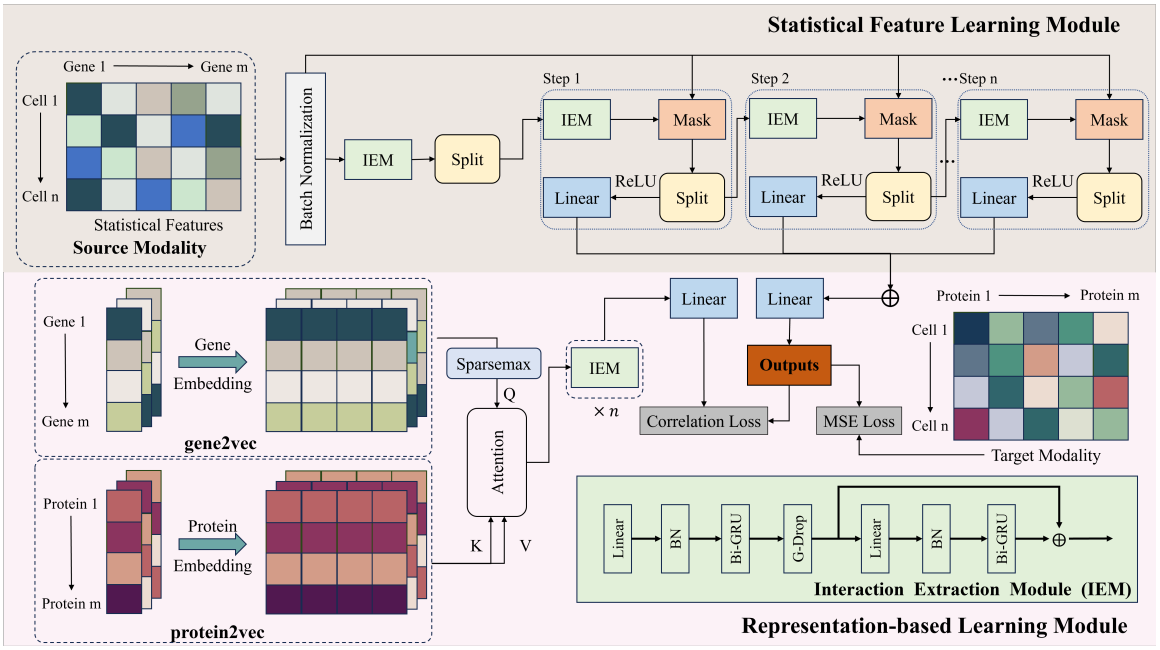

Figure 5.2: Overall structure of scAGFormer. The scAGFormer methodology employs a two-branch Transformer architecture. The upper branch initiates with batch-normalized single-cell data, advancing through a multi-step feature selection process guided by a learnable mask and the Interactions of Extracted Modalities (IEM) technique. This process, involving shared and step-specific layers, culminates in refined predictions of target modality expressions. Concurrently, the lower branch constructs a cross-attention between source and target modality embeddings, enhanced by sparsemax operation for sparse data, thereby capturing intricate relationships. This dual approach, integrating advanced attention mechanisms and loss functions, enables the scAGFormer to efficiently analyze and predict multimodal single-cell data. G-Drop refers to Gaussian Dropout [264]. Bi-GRU refers to Bidirectional Gated Recurrent Unit. Linear refers to linear transformation.

5.4 Methodology

5.4.1 Affinity Graph for Multi-modal Single Cell Prediction

Affinity graph is a graphical structure widely adopted in machine learning to represent the relationship between entities to extract finer-grained features than the entire feature extractor similar to ResNet [107]. Affinity graphs can capture non-linear and complex relationships between data points, which might not be possible with simpler representations [16]. Therefore, we use the affinity graph between the source modality and target modality as an external constraint to utilize the similarity hidden space of source modality and target modality. We bridged two modalities with the affinity graph and used non-linear transformation to project the affinity graph to a hidden

vector as an affinity graph priori.

We construct a multimodality feature-feature bipartite graph where the features from two modalities (*e.g.* GEX, ADT or ATAC) are treated as different nodes. In an affinity graph, nodes or vertices correspond to entities, and edges or links represent their relationships. The weight of edges reflects the strength of relationships between entities. Formally, we denote the bipartite graph as $\mathcal{G} = (\mathcal{U}, \mathcal{V}, \mathcal{E})$. $\mathcal{U}$ is the set of n features $u_1, u_2, ..., u_n$ in the source modality, and $\mathcal{V}$ is the set of n features $v_1, v_2, ..., v_n$ in the target modality. $\mathcal{E} \subseteq \mathcal{U} \times \mathcal{V}$ represents the set of edges between $\mathcal{U}$ and $\mathcal{V}$, which describe the relations between cells in different modalities. Various similarity measures, such as Euclidean distance and cosine similarity, can be employed to calculate the relationships between entities. Cosine similarity is a common method to measure pairwise similarity. Specifically, inspired by SimCLR [47], we further extend the idea of affinity graph constraints to form an affinity matrix by connecting input queries and keys to measure pairwise similarity. Given n queries $q_1, q_2, ..., q_n$, then an adjacency matrix W of $n \times n$ can be used to represent the affinity graph, as follows:

$$W = \begin{bmatrix} W_{11} & W_{12} & \cdots & W_{1n} \\ W_{21} & W_{22} & \cdots & W_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ W_{n1} & W_{n2} & \cdots & W_{nn} \end{bmatrix}, \tag{5.1}$$

where W_{ij} is usually a non-negative real number, indicating the similarity between the query q_i and q_j. When W_{ij} is larger, it means that the relationship between the token x_i and the token x_j is closer. The affinity graph is constructed based on the affinity matrix; that is, the affinity graph of q and the affinity graph of the complete q are randomly selected, and then the two affinity graphs are matched by position embedding to obtain finer-grained features.

5.4.2 scAGFormer

In this section, we present the details of our proposed scAGFormer, a transformer-based framework tailored for the multi-modal single-cell prediction task. The comprehensive architecture of our proposed methodology is delineated in Figure 5.2. scAGFormer employs a two-branch transformer-based structure that unifies statistical feature extraction and representation-based learning through the application of correlation loss. Existing raw data predominantly consist of statistical features;

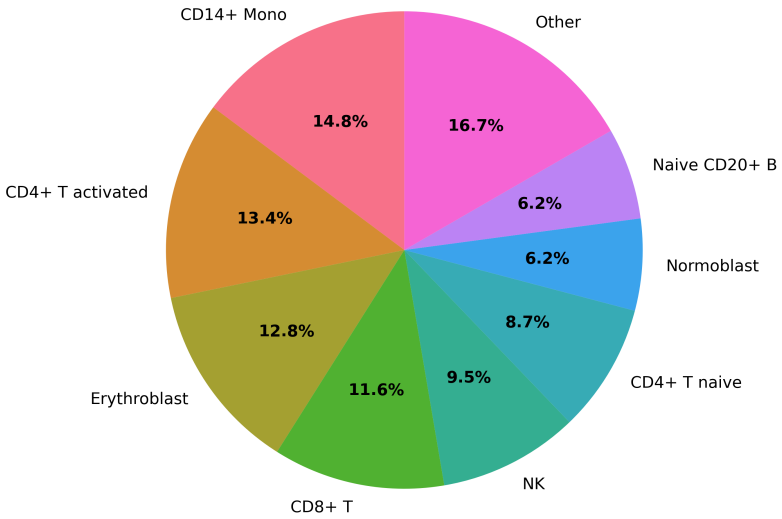

Figure 5.3: Distribution of Cell Types in Donor 1 and Site 1 for NeurIPS 2021 Multimodal Single-Cell Data Integration.

thus, according to [274], statistical feature learning methods like Gradient-based Decision Tree (GBDT) methods are anticipated to excel in tasks related to multi-modal single-cell analysis. Nonetheless, our empirical investigations reveal that, in the absence of additional feature engineering, the correlation between the outputs generated by GBDT and its variants is markedly inferior to that of baseline models. To address this discrepancy, we integrate deep learning-based tabular feature extraction pipelines with representation learning pipelines, employing correlation loss as the unifying element.

To utilize features extracted from each task more comparably, we design and employ an identical block, Interaction Extraction Module (IEM), to extract the statistical features and representation features. Keeping architectural consistency ensures both tasks are modelled in a consistent way [145, 19]. Nowadays, applying Transformer [253] to representation learning has achieved widespread success in computer vision [73] and protein modeling [213, 214]. Previous research has proven the effectiveness and robustness of applying transformer framework in tabular feature processing [6, 119]. To this end, we utilize the statical features and representation features with a transformer layer-based solution.

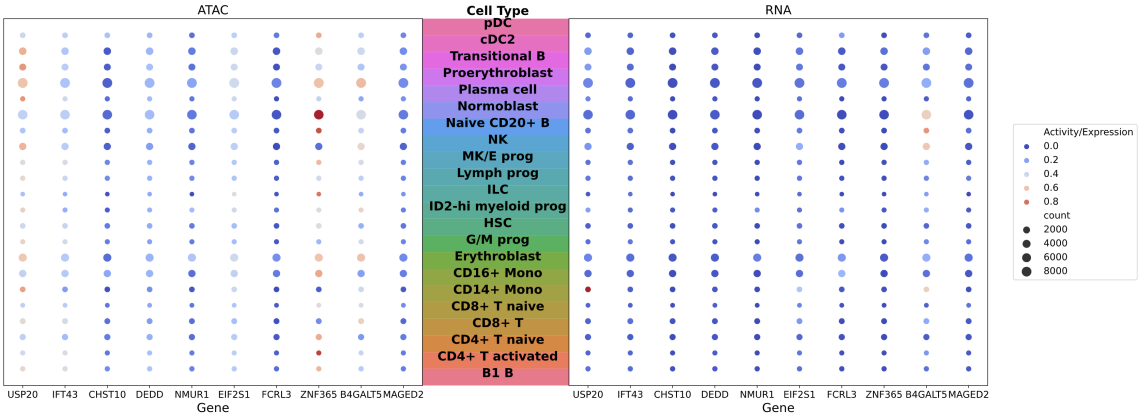

Figure 5.4: Expression of marker genes across the identified cell types in NeurIPS 2021 Multimodal Single-Cell Data Integration Challenge. Each row in a dot plot corresponds to a marker gene, while each column corresponds to a cell type. The size of a dot within the plot reflects the proportion of cells within that type expressing the gene (fraction of cells), and the colour intensity represents the average expression level (mean expression) of that gene in those cells. This dual-parameter approach helps in assessing both the prevalence and the degree of gene expression or chromatin accessibility across different cell types.

5.5 Experiments

5.5.1 Datasets

In this section, we evaluate the effectiveness of scAGFormer against two benchmarks and show scAGFormer outperforms baselines, state-of-the-art methods, and top winners in the competitions, over three metrics on NeurIPS 2021: Multimodal Single-Cell Data Integration [183] and NeurIPS 2022: Multimodal Single-Cell Integration Across Time, Individuals, and Batches [66] datasets.

For NeurIPS 2021: Multimodal Single-Cell Data Integration [183], the dataset statistics of the modality prediction task are presented in Table 5.1. The dimensions are compatible for GEX-ADT and GEX-ATAC, while GEX-ATAC and ATAC-GEX are not symmetric in the dimensions. We use the modality prediction benchmark to evaluate our proposed method. As depicted in Figure 5.3, the pie chart delineates the distribution of diverse cell types in this dataset. Cell types constituting less than 16.7% of the total are amalgamated under the "Other" category. This representation underscores an equitable distribution of cells across the various types. There is no one dominant cell type. In particular, CD14 + Monocytes have the highest proportion among cell types, 1.4 higher than the second highest CD4 + T activated cells and four types of cells have a proportion greater than 10%. Therefore, the wide distribution

	GEX→ADT	ADT→GEX	GEX→ATAC	ATAC→GEX
Source Dim	13,953	134	13,431	116,490
Target Dim	134	13,953	10,000	13,431
Train Cells	66,175	66,175	42,492	42,492
Test Cells	1,000	1,000	1,000	1,000
Train Batches	9	9	10	10
Test Batches	3	3	3	3

Table 5.1: Dataset statistics of modality prediction task for NeurIPS 2021 Multimodal Single-Cell Data Integration. The number of feature dimensions, train/test samples, and batches.

observed in our datasets underscores the potential reach of our research. Successfully completing the task and demonstrating strong performance on this dataset suggest that many other single-cell tasks could similarly be addressed. This implication is profound, as it indicates that our methodologies and findings could be applicable and beneficial across a broad spectrum of single-cell research, offering a versatile toolkit for unraveling the complexities of cellular biology.

For the NeurIPS 2022: Multimodal Single-Cell Integration Across Time, Individuals, and Batches [66], we are provided with inputs where rows correspond to cells and columns correspond to genes and outputs which are surface protein levels for the same cells. It is worth mentioning that the protein level testing data is not available during the completion of this work. Therefore, we simulate the competition scenario by treating the training data from day 4 as our testing set. The processed RNA data are centred and logarithmically transformed, while the normalized protein levels are denoised and scaled by background [149].

5.5.2 Evaluation Metrics

The efficacy of the final protein-level predictions is assessed by employing the root mean square error (RMSE) and mean absolute error (MAE) metrics, where RMSE and MAE are defined by

$$\text{RMSE} = \sqrt{\frac{1}{N}\|\mathbf{M}' - \hat{\mathbf{M}}'\|_F^2}, \tag{5.2}$$

$$\text{MAE} = \frac{1}{N}\|\mathbf{M}' - \hat{\mathbf{M}}'\|_1, \tag{5.3}$$

where $\hat{\mathbf{M}}'$ is the corresponding prediction of $\mathbf{M}'$respectively. However, it is crucial to acknowledge that multimodal data are frequently subjected to the confounding

influences of batch effects and unequal measurement depths. Consequently, the scale of each cell's count can exhibit significant variability, profoundly impacting the RMSE and MAE measurements.

To mitigate these effects and provide a more balanced evaluation, we also incorporate the Pearson correlation coefficient (Corr), a metric that normalizes the input's mean and variance on a cell-by-cell basis. The Corr metric serves as a robust and scale-independent measure to assess predictions. What's more, in the single-cell analysis domain, the coefficient is more informative than the RMSE and MAE metric [54, 53]. Lower scores on the RMSE or MAE scales signify a closer geometric estimation of the protein levels, indicating more accurate predictions. Concurrently, a higher Corr score reflects a stronger statistical similarity to the actual protein levels, implying a more precise match with the true value.

5.5.3 Baselines

We evaluated the performance of scAGFormer against state-of-the-art multimodal prediction models for the task of using gene expression to predict surface protein levels. However, as the competition report [153] and scMoGNN [274] suggested that statistical methods like tSVD with logistic regression (the baseline model), random forest, and GBDT (LightGBM [143]) should also be listed as a competitive model in this task. Therefore, we present the decision tree-based method without further feature engineering to see the model performance between representation-based methods and statistical-based methods. We also applied a popular model structure which applied a shallow neural network after the GBDT [108].

For all statistically based methods, we applied the same pre-processing measure with the centre log-ratio transformation (CLR) [197] to normalize the data and applied tSVD [101] to transform both input and target data into 128 dim. We applied GBDT with CatBoost [203] directly to the transformed 128 features to perform the regression. We then decode the regression results into original feature spaces to evaluate the overall performance.

The selected representation-based baselines are as follows:

- **Cross-modal Autoencoders** [293] (CMAE), incorporates multiple autoencoders to integrate multimodal data and utilizes domain knowledge by adding discriminative loss to the training process to align shared markers or clusters among datasets.

- **BABEL** [276] proposes a general framework for multimodal translation with modality-specific encoders and decoders. Note that initially, BABEL focuses on RNA and ATAC-seq [27] data. In this evaluation, we repurpose BABEL from the RNA to protein setting.

- **scMM** [194] models the multimodal data with a generative setting. We note that the input of scMM is restricted to raw counts by design, and the output predictions are scaled as log-transformed centered.

- **scMoGNN** [274] involves domain knowledge like biological pathways to enhance the GNNs. The original scMoGNN follows a transductive setting. In this work, we implement an inductive setting of scMoGNN for a fair comparison with the baselines. It is worthwhile to mention that scMoGNN has won the first prize in the NeurIPS 2021 Competition (GEX→ADT).

- **scFormer** [64] uses the Transformer to represent genes and cells in an unsupervised manner. We utilize the representation of scFormer and directly connected with a multilayer perceptron (MLP).

- **scGPT** [65] uses GPT [210] to learn the gene and cell representation from over 10 million single-cell data, outside the training set. We apply this pre-trained model[3] directly and perform fine-tuning on the downstream task.

5.5.4 Experimental Setups

All experiments are carried out on 8 NVIDIA A100 GPUs. For all the transformer-based structures (scFormer and scGPT), we use the AdamW [182] optimizer with an initial learning rate of $1e^{-4}$. To complete the training, we used mixed precision training [192] with FP16 precision with NVIDIA Apex[4]. For CMAE, BABEL, and scMM, we apply 150 epochs with a batch size of 128. ScMoGNN is hosting a large network, we followed the same setup in the original paper with over 1,500 epochs. To present a fair comparison, we performed a CLR for all experiments. Note that scMM [194] prediction is not compatible with normalized protein levels in the NeurIPS 2022 dataset. Therefore, we do not report its experimental result in Table **??**. To present a fair comparison, we applied tSVD [101] preprocessing to ScMoGNN [274] without ad-hoc manual feature selection.

[3] `https://github.com/bowang-lab/scGPT`
[4] `https://github.com/NVIDIA/apex`

Model	RMSE↓	MAE↓	Corr↑
CMAE	2.000 ± 0.022	1.208 ± 0.008	0.811 ± 0.001
BABEL	1.662 ± 0.007	1.076 ± 0.006	0.873 ± 0.002
scMoGNN	1.663 ± 0.007	1.066 ± 0.005	0.877 ± 0.001
GBDT	1.622 ± 0.002	1.055 ± 0.001	0.677 ± 0.002
GBDT+MLP	1.620 ± 0.002	1.055 ± 0.001	0.697 ± 0.003
scFormer w/MLP	1.661 ± 0.002	1.078 ± 0.008	0.711 ± 0.010
scGPT w/MLP	1.662 ± 0.005	1.077 ± 0.004	0.673 ± 0.001
NeurIPS 2022 CITE-seq[5]	1.622 ± 0.003	1.055 ± 0.002	0.899 ± 0.001
scAGFormer (ours)	**1.577 ± 0.003**	**1.005 ± 0.002**	**0.909 ± 0.002**

Table 5.2: Prediction evaluations based on different metrics (score ± std) in CITE-seq [239]. Note that the scale of scMM [293] prediction is not compatible with that of normalized protein levels.

Since both scFormer and scGPT are not directly designed for these cases, we manually modify the prediction head as a linear transformation to fit into our cases. To present a fair comparison, we also disable domain-specific batch normalization in scFormer and scGPT. We follow the same hyperparameter settings in scGPT [65] to fine-tune the scGPT on the two datasets. We apply the learning rate as $5e^{-5}$ with cosine annealing decay. Each experiment within our study was rigorously conducted in three separate trials, each initialized with a unique random number seed. The results of these trials are presented in terms of their mean and variance.

5.5.5 Main Results

We use bidirectional predictions spanning GEX and ADT, and GEX and ATAC to prove that our method is superior to state-of-the-art multimodal prediction models in the task of using gene expression to predict surface protein levels. The quantitative results are shown in Table 5.2, Table 5.3, and Table 5.4 and the specific analysis is as follows.

First, as delineated in Table 5.3, for bidirectional predictions spanning GEX and ADT, our method consistently registers the smallest errors, underscoring the superiority of lower RMSE, lower MAE, and higher Corr values. Specifically, in bidirectional prediction across GEX and ADT, our scAGFormer outperforms the latest methods that exploit cell embedding or representation as an intermediate step in their overall workflow (i.e., CMAE, BABEL, scMM, scMoGNN, and scGPT) in all metrics. For

[5]Winner solution of NeurIPS 2022: Multimodal Single-Cell Data Integration Competition. Source code from `https://github.com/shu65/open-problems-multimodal`

Model	GEX→ADT			ADT→GEX		
	RMSE↓	MAE↓	Corr↑	RMSE↓	MAE↓	Corr↑
CMAE	0.499 ± 0.011	0.344 ± 0.004	0.811 ± 0.004	0.359 ± 0.001	0.121 ± 0.001	0.692 ± 0.003
BABEL	0.498 ± 0.027	0.338 ± 0.016	0.819 ± 0.012	0.371 ± 0.018	0.138 ± 0.003	0.662 ± 0.038
scMM	0.638 ± 0.008	0.433 ± 0.004	0.670 ± 0.010	0.314 ± 0.056	0.100 ± 0.035	0.278 ± 0.001
scMoGNN	0.421 ± 0.001	0.285 ± 0.003	0.866 ± 0.001	0.324 ± 0.003	0.154 ± 0.002	0.744 ± 0.003
GBDT	0.425 ± 0.002	0.282 ± 0.001	0.658 ± 0.001	0.323 ± 0.001	0.153 ± 0.001	0.575 ± 0.002
GBDT+MLP	0.403 ± 0.002	0.288 ± 0.001	0.658 ± 0.001	0.330 ± 0.001	0.148 ± 0.002	0.577 ± 0.002
scFormer w/MLP	0.481 ± 0.003	0.320 ± 0.003	0.658 ± 0.001	0.327 ± 0.001	0.155 ± 0.001	0.677 ± 0.002
scGPT w/MLP	0.461 ± 0.005	0.308 ± 0.003	0.780 ± 0.007	0.328 ± 0.003	0.154 ± 0.001	0.677 ± 0.004
scAGFormer (ours)	**0.351 ± 0.008**	**0.244 ± 0.008**	**0.885 ± 0.002**	**0.291 ± 0.008**	**0.077 ± 0.002**	**0.811 ± 0.006**

Table 5.3: Experimental results on GEX→ADT and ADT→GEX modality prediction in NeurIPS 2021 Multimodal Single-Cell Data Integration Challenge. The reported numbers are mean ± standard deviation. Note that the scMoGNN is the winner model of the original task.

Model	GEX→ATAC			ATAC→GEX		
	RMSE↓	MAE↓	Corr↑	RMSE↓	MAE↓	Corr↑
CMAE	0.192 ± 0.001	0.037 ± 0.001	0.005 ± 0.010	0.247 ± 0.004	0.087 ± 0.002	0.301 ± 0.013
BABEL	0.182 ± 0.001	0.047 ± 0.001	0.285 ± 0.013	0.243 ± 0.002	0.093 ± 0.001	0.381 ± 0.001
scMM	0.216 ± 0.001	0.047 ± 0.001	0.091 ± 0.002	0.374 ± 0.001	0.156 ± 0.028	0.188 ± 0.001
scMoGNN	0.203 ± 0.008	0.169 ± 0.002	0.282 ± 0.019	0.230 ± 0.007	0.103 ± 0.007	0.412 ± 0.001
GBDT	0.178 ± 0.001	0.062 ± 0.001	0.055 ± 0.001	0.236 ± 0.002	0.117 ± 0.001	0.195 ± 0.001
GBDT+MLP	0.178 ± 0.001	0.062 ± 0.001	0.101 ± 0.001	0.225 ± 0.002	0.112 ± 0.002	0.211 ± 0.001
scFormer w/MLP	0.178 ± 0.001	0.058 ± 0.001	0.234 ± 0.004	0.235 ± 0.002	0.112 ± 0.002	0.195 ± 0.001
scGPT w/MLP	0.178 ± 0.001	0.060 ± 0.001	0.215 ± 0.006	0.238 ± 0.001	0.115 ± 0.001	0.195 ± 0.001
scAGFormer (ours)	**0.138 ± 0.009**	**0.030 ± 0.004**	**0.303 ± 0.010**	**0.211 ± 0.011**	**0.077 ± 0.005**	**0.477 ± 0.010**

Table 5.4: Experimental results on GEX→ATAC and ATAC→GEX modality prediction in NeurIPS 2021 Multimodal Single-Cell Data Integration Challenge. The reported numbers are mean ± standard deviation. Note that the scMoGNN is the winner model of the orignal task.

example, in the important Corr metric, our method v.s. the second best method is 0.885 ± 0.002 versus 0.866 ± 0.001 (scMoGNN) and 0.811 ± 0.006 versus 0.744 ± 0.003 (scMoGNN) respectively. This is because scAGFormer is a method capable of predicting modes independent of cell embedding, which can alleviate the huge heterogeneity and fluidity of cellular systems. Moreover, comparing scAGFormer with statistics-based methods (i.e., GBDT-based methods), we also find that scAGFormer achieves the best results in all metrics, which proves the effectiveness of merging feature-driven decision trees and neural networks.

Then, the same observation is made in the bidirectional prediction of cross-domain GEX and ATAC in Table 5.4, that is, our method always records minor errors. From GEX to ATAC, our algorithm is improved by 0.018 compared to the optimal BABEL in the Corr metric, and the improvement is still obvious. In the opposite direction, from ATAC to GEX, the MAE metric of scAGFormer is 0.010 lower than CMAE, mainly because multi-modal data are often confounded by batch effects and unequal measurement depths. In the Corr metric that better reflects the similarity to the

actual protein, scAGFormer is 0.065 higher than the second-best scMoGNN, which means that the protein it generates matches the real value more accurately.

Finally, in Figure 5.4, the visual representation juxtaposes ATAC activity with RNA expression of specific genes across diverse cellular classifications. The left-hand panel delineates ATAC activity, whilst the right-hand panel showcases RNA expression magnitudes. Distinct hues and dimensions signify the levels of activity/expression and their corresponding counts. The central legend systematically categorises the various cell types, offering an integrated analysis of the inherent cellular dynamics. From the analysis of the figure, it is evident that for the gene ZNF365, the ATAC activity is markedly high in Normoblast cells, reaching over 8000 counts and an activity level near 0.8. For RNA expression, the genes B4GALT5 and USP20 stand out, with B4GALT5 showing higher expression levels in Normoblast, Naive CD20+ B, and NK cells. Meanwhile, the USP20 gene exhibits a notably high RNA expression level in the CD14+ Mono cell type. This visual comparison elucidates the differences in gene regulation across various cell types, highlighting specific gene-cell type interactions. The ability of our model to complete tasks on such complex datasets bodes well for the applicability of our methods to other single-cell tasks.

In summation, our scAGFormer demonstrated consistently superior performance in comparison to both baseline and contemporary state-of-the-art models across all error and correlation metrics.

5.5.6 Ablation Studies - Impact of Representational Feature Learning Module

In this section, we discuss the results of an ablation study designed to assess the contribution of the Statistical Feature Learning Module (Stat Module) and the Representation Feature Learning Module (Rep Module) to the scAGFormer model, as implemented in the multimodal single-cell data integration challenge at NeurIPS 2021. The results of this examination are systematically tabulated in Table 5.5 and Table 5.6. The ablation study investigates the two-branch configuration of the scAGFormer, focusing particularly on the roles of the Statistical Feature Learning Module (Stat Module) and its combination with the Representation Feature Learning Module (Rep Module). The scAGFormer utilizes a dual-branch structure, which necessitates an analysis of its two separate modules. Also, we view the combination of GBDT with MLP as analogous to our integration of the Stat Module with the Rep Module, with GBDT acting as the Stat Module and MLP as the Rep Module. In our ablation study, we evaluate the performance of scAGFormer in conjunction with GBDT and

Stat Module		Rep Module		GEX→ADT			ADT→GEX		
GBDT	Ours	MLP	Ours	RMSE	MSE	Corr	RMSE	MSE	Corr
✓				0.425 ± 0.002	0.282 ± 0.001	0.658 ± 0.001	0.323 ± 0.001	0.153 ± 0.001	0.575 ± 0.002
	✓			0.413 ± 0.004	0.271 ± 0.002	0.664 ± 0.002	0.343 ± 0.003	0.155 ± 0.001	0.577 ± 0.002
✓		✓		0.403 ± 0.002	0.288 ± 0.001	0.658 ± 0.001	0.330 ± 0.001	0.148 ± 0.002	0.577 ± 0.002
✓			✓	0.403 ± 0.002	0.287 ± 0.002	0.775 ± 0.002	0.331 ± 0.002	0.147 ± 0.001	0.751 ± 0.010
	✓	✓		0.388 ± 0.003	0.266 ± 0.004	0.751 ± 0.004	0.311 ± 0.008	0.113 ± 0.003	0.770 ± 0.004
	✓		✓	$\mathbf{0.351 \pm 0.008}$	$\mathbf{0.244 \pm 0.008}$	$\mathbf{0.885 \pm 0.002}$	$\mathbf{0.291 \pm 0.008}$	$\mathbf{0.077 \pm 0.002}$	$\mathbf{0.811 \pm 0.006}$

Table 5.5: Impact of Representational Feature Learning Module in the Trajectory from GEX→ ADT and ADT→ GEX in NeurIPS 2021 Multimodal Single-Cell Data Integration Challenge. The reported numbers are mean ± standard deviation.

Stat Module		Rep Module		GEX→ATAC			ATAC→GEX		
GBDT	Ours	MLP	Ours	RMSE	MSE	Corr	RMSE	MSE	Corr
✓				0.178 ± 0.001	0.062 ± 0.001	0.055 ± 0.001	0.236 ± 0.002	0.117 ± 0.001	0.195 ± 0.001
	✓			0.178 ± 0.001	0.062 ± 0.002	0.057 ± 0.003	0.238 ± 0.003	0.118 ± 0.003	0.194 ± 0.002
✓		✓		0.178 ± 0.001	0.062 ± 0.001	0.101 ± 0.001	0.225 ± 0.002	0.112 ± 0.002	0.211 ± 0.001
✓			✓	0.153 ± 0.003	0.055 ± 0.002	0.133 ± 0.004	0.223 ± 0.002	0.113 ± 0.003	0.311 ± 0.003
	✓	✓		0.157 ± 0.001	0.055 ± 0.003	0.122 ± 0.003	0.243 ± 0.003	0.118 ± 0.002	0.213 ± 0.002
	✓		✓	$\mathbf{0.138 \pm 0.009}$	$\mathbf{0.030 \pm 0.004}$	$\mathbf{0.303 \pm 0.010}$	$\mathbf{0.211 \pm 0.011}$	$\mathbf{0.077 \pm 0.005}$	$\mathbf{0.477 \pm 0.010}$

Table 5.6: Impact of Representational Feature Learning Module in the Trajectory from GEX→ ATAC and ATAC→ GEX in NeurIPS 2021 Multimodal Single-Cell Data Integration Challenge. The reported numbers are mean ± standard deviation.

a three-layers MLP, focusing on how each stream contributes to the model's overall effectiveness.

Our findings indicate that the individual contributions of the Stat Module and the Rep Module are both substantial, yet their integration is crucial for optimal performance. When deployed individually, the Stat Module (both ours and GBDT) serves as an effective baseline, providing competent predictive power. However, it falls short in capturing the complexity interaction of the data, as evidenced by higher RMSE and MSE scores and lower correlation coefficients across the tasks when compared to the integrated approach. The synergetic effect of combining both the Stat Module and the Rep Module cannot be overstated. This configuration achieves the highest performance across all tasks, with the most substantial gains observed in correlation scores. Moreover, the combined model significantly outperforms each module used in isolation, underlining the effectiveness of a dual-branch structure.

It is clear that the integration of both the Statistical Feature Learning Module (Stat Module) and the Representation Feature Learning Module (Rep Module) within our approach results in the highest performance for the tasks of GEX→ADT and ADT→GEX. The combined model (scAGFormer with both modules) substantially outperforms the standalone GBDT model, emphasizing the additional value of the Rep Module in enhancing predictive accuracy and correlation with actual data. When only the GBDT is used for the Stat Module, we observe performance met-

Table 5.7: Abalation Study on Loss Functions in the trajectory from GEX → ADT and ADT → GEX in NeurIPS 2021 Multimodal Single-Cell Data Integration Challenge. The reported numbers are mean ± standard deviation.

Rep Loss			Feat Loss			GEX→ADT			ADT→GEX		
MAE	MSE	Corr	MAE	MSE	Corr	RMSE	MSE	Corr	RMSE	MSE	Corr
✓			✓			0.370 ± 0.010	0.263 ± 0.003	0.806 ± 0.004	0.299 ± 0.005	0.100 ± 0.005	0.711 ± 0.009
	✓		✓			0.371 ± 0.010	0.261 ± 0.003	0.796 ± 0.004	0.301 ± 0.005	0.100 ± 0.005	0.711 ± 0.005
		✓	✓			0.354 ± 0.006	0.245 ± 0.003	$\mathbf{0.866 \pm 0.004}$	0.295 ± 0.006	0.079 ± 0.003	0.801 ± 0.010
✓				✓		0.369 ± 0.008	0.260 ± 0.005	0.799 ± 0.006	0.296 ± 0.005	$\mathbf{0.077 \pm 0.002}$	0.707 ± 0.005
	✓			✓		0.372 ± 0.004	0.261 ± 0.003	0.796 ± 0.004	0.296 ± 0.005	$\mathbf{0.077 \pm 0.002}$	0.717 ± 0.009
		✓		✓		$\mathbf{0.351 \pm 0.008}$	$\mathbf{0.244 \pm 0.008}$	0.885 ± 0.002	$\mathbf{0.291 \pm 0.008}$	$\mathbf{0.077 \pm 0.002}$	$\mathbf{0.811 \pm 0.006}$
✓					✓	0.358 ± 0.005	$\mathbf{0.244 \pm 0.008}$	0.808 ± 0.010	0.302 ± 0.005	0.101 ± 0.003	0.766 ± 0.005
	✓				✓	0.358 ± 0.005	0.248 ± 0.005	0.809 ± 0.003	0.311 ± 0.003	0.101 ± 0.003	0.775 ± 0.003
		✓			✓	0.373 ± 0.007	0.263 ± 0.003	0.806 ± 0.004	0.315 ± 0.006	0.119 ± 0.003	0.785 ± 0.010

Table 5.8: Abalation Study on Loss Functions in the trajectory from GEX → ATAC and ATAC → GEX in NeurIPS 2021 Multimodal Single-Cell Data Integration Challenge. The reported numbers are mean ± standard deviation.

Rep Loss			Feat Loss			GEX→ATAC			ATAC→GEX		
MAE	MSE	Corr	MAE	MSE	Corr	RMSE	MSE	Corr	RMSE	MSE	Corr
✓			✓			0.148 ± 0.003	0.050 ± 0.002	0.233 ± 0.005	0.195 ± 0.004	0.078 ± 0.003	0.448 ± 0.004
	✓		✓			0.147 ± 0.003	0.050 ± 0.002	0.235 ± 0.003	$\mathbf{0.195 \pm 0.004}$	0.078 ± 0.003	0.438 ± 0.004
		✓	✓			0.141 ± 0.006	0.031 ± 0.004	$\mathbf{0.303 \pm 0.004}$	0.205 ± 0.004	0.081 ± 0.004	0.447 ± 0.003
✓				✓		0.144 ± 0.005	0.050 ± 0.002	0.233 ± 0.005	$\mathbf{0.195 \pm 0.004}$	0.081 ± 0.003	0.440 ± 0.006
	✓			✓		0.141 ± 0.004	0.040 ± 0.005	0.237 ± 0.003	0.205 ± 0.004	0.079 ± 0.003	0.440 ± 0.006
		✓		✓		$\mathbf{0.138 \pm 0.009}$	$\mathbf{0.030 \pm 0.004}$	$\mathbf{0.303 \pm 0.010}$	0.211 ± 0.011	$\mathbf{0.077 \pm 0.005}$	0.477 ± 0.008
✓					✓	0.141 ± 0.004	0.041 ± 0.002	0.255 ± 0.004	0.199 ± 0.005	0.080 ± 0.002	0.457 ± 0.001
	✓				✓	0.141 ± 0.004	0.041 ± 0.003	0.255 ± 0.004	0.198 ± 0.003	0.080 ± 0.002	0.456 ± 0.003
		✓			✓	0.141 ± 0.006	0.031 ± 0.004	$\mathbf{0.303 \pm 0.010}$	0.205 ± 0.004	0.078 ± 0.003	$\mathbf{0.478 \pm 0.006}$

rics with high RMSE and MSE, and lower correlation (Corr) values. This indicates that while GBDT contributes the model's predictive capabilities, it does not offer the comprehensive feature understanding provided by the Rep Module. Integrating an MLP with GBDT results in some improvement in performance metrics; however, the correlation still remains relatively low, suggesting that the MLP does not fully compensate for the absence of the dedicated Rep Module.

5.5.7 Ablation Studies - Analysis on Loss Functions

In this section, we explore the ablation study on loss functions. Concerning the loss functions, MAE, MSE, and Corr are compared for both Representational Feature Learning Loss (Rep loss) and Feature Learning Loss (Feat Loss). We argued that precise statistical features derived from statistical learning modules, together with correlations identified by representation learning modules, are crucial for solving multi-modal prediction task. Since statistical methods are cannot capture the complex interactions between different representations [17], we expect correlation loss would be more helpful to representation learning module.

As depicted in Tables 5.7 and 5.8, the trajectory from GEX to ADT exhibits the

Module			GEX→ADT			ADT→GEX		
Bi-GRU	G-Dropout	Skip Connection	RMSE	MAE	Corr	RMSE	MAE	Corr
✓			0.377 ± 0.003	0.266 ± 0.003	0.711 ± 0.003	0.313 ± 0.001	0.155 ± 0.001	0.671 ± 0.002
	✓		0.378 ± 0.002	0.266 ± 0.002	0.677 ± 0.002	0.311 ± 0.004	0.131 ± 0.003	0.671 ± 0.002
		✓	0.358 ± 0.003	0.246 ± 0.002	0.812 ± 0.008	0.315 ± 0.006	0.131 ± 0.005	0.771 ± 0.005
✓	✓		0.377 ± 0.002	0.266 ± 0.002	0.711 ± 0.003	0.311 ± 0.002	0.131 ± 0.001	0.672 ± 0.002
✓		✓	0.355 ± 0.007	0.244 ± 0.005	0.867 ± 0.010	0.293 ± 0.004	0.111 ± 0.005	0.787 ± 0.004
	✓	✓	0.357 ± 0.007	0.244 ± 0.005	0.877 ± 0.005	0.311 ± 0.002	0.131 ± 0.002	0.771 ± 0.002
✓	✓	✓	$\mathbf{0.351 \pm 0.008}$	$\mathbf{0.244 \pm 0.008}$	$\mathbf{0.885 \pm 0.002}$	$\mathbf{0.291 \pm 0.008}$	$\mathbf{0.077 \pm 0.002}$	$\mathbf{0.811 \pm 0.006}$

Table 5.9: Analysis of Interaction Extraction Module

lowest RMSE and MSE given Rep Loss is correlation loss and Feat Loss is MSE. The best RMSE and MSE versus the second best results are 0.351 ± 0.008 v.s 0.354 ± 0.006 and 0.244 ± 0.008 v.s 0.245 ± 0.008 respectively. Simultaneously, the correlation coefficients (Corr) are second highest which is 0.885 ± 0.002 compared to 0.866 ± 0.004 when the Rep Loss is a correlation loss and the Feat Loss is an MAE loss which corroborate our claims. We can easily find out that once correlation loss applied on representation module (Corr as Rep Loss), the performance has got a significant improvement on all the task in GEX→ADT and ADT→GEX. Since the prediction of ADT→GEX modality involves a small source dimension (134) and an extremely large target dimension $(13,953)$, learning the correct correlation becomes significantly challenging. We observe that the applying Rep loss with correlation loss is highly effective when combined with accuracy-based loss in Feat Loss. Applying correlation loss to both ways results in high performance on correlation-related metrics, although it compromises prediction accuracy. What's more, apply correlation loss to both module does not have the highest correlation score in all tasks except the ATAC→GEX (0.478 ± 0.006 vs 0.477 ± 0.008). This implies that without accurate statistical features, the reliability of feature learning is compromised.

5.5.8 Ablation Studies - Analysis of Interaction Extraction Module

We further analyze three functional modules within IEM, which are the Bidirectional Gated Recurrent Units (Bi-GRUs), Gaussian Dropout (G-Drop) layer, and skip connection. We examine the impact of three modules by substituting all Bi-GRUs with linear transformation, substituting G-Drop with vanilla dropout, and introducing a skip connection from the initial G-Drop layer directly to the IEM's output, The results of this examination are presented in Table 5.9.

Analysing Table 5.9, the most complete version of the IEM, with Bi-GRUs, G-Drop, and a skip connection integrated, registers the lowest RMSE of 0.351 ± 0.008 and MSE of 0.244 ± 0.008, along with the highest correlation of 0.885 ± 0.002. Skip

connection is a well-known option to improve the representation learning [112, 306]. We can easily find that although skip-connection improved the overall performance in both accuracy and correlation, it also increased the standard deviation. G-Dropout is designed to be more robust to the noise in the input data. We can easily see the effectiveness of stabilizing the performance of all different tasks when introducing G-Dropout in Table 5.9.

5.6 Summary

Our study introduces the single-cell Affinity Graph transFormer (scAGFormer), a novel cell-agnostic framework for multi-modal single-cell analysis. In our exploration of single-cell multimodal prediction methodologies, the synergy between the Statistical and Representational Modules emerges as a pivotal factor in transcending the limitations of traditional techniques. The advantage of statistical features in conjunction with representational features schemes over even large models like scGPT with strong representational capabilities alone means that statistical features in this domain still have a very significant advantage. Our model's ability to operate independently of predefined cell embeddings allows it to handle outlier cell types and novel cellular conditions more effectively, showcasing its robustness and adaptability. To refine the predictive accuracy of our models further, it is crucial to strengthen collaborative ties with experimental biology. Collaborative efforts could facilitate training our machine learning models on rich, annotated datasets, incorporating novel biological markers or detailed treatment responses. This enhancement would improve the model's utility in clinical settings, where accurate predictions can significantly impact patient outcomes. Moreover, feedback from biological experiments could guide refinements in our model's architecture and feature selection processes, ensuring that our approach remains aligned with the latest scientific findings.

5.7 Limitations

While scAGFormer demonstrates superior performance compared to existing methods for multi-modal single-cell analysis, it is important to acknowledge its limitations. One key limitation is the interpretability of the learned representations. Although scAGFormer's cell-agnostic approach offers advantages in handling novel cell types and conditions, the complexity of the model architecture may hinder the direct biological interpretation of the learned features. Improving the interpretability of the

model's internal representations could facilitate deeper insights into the underlying biological mechanisms and enhance the model's utility for hypothesis generation and experimental validation.

Furthermore, the scalability and computational efficiency of scAGFormer when applied to extremely large-scale single-cell datasets remain to be fully explored. As the size and complexity of single-cell multi-omics datasets continue to grow, it will be important to assess the model's ability to handle such data volumes while maintaining its predictive performance and practical feasibility.

Lastly, while scAGFormer has been evaluated on multiple benchmark datasets, its generalizability to a wider range of single-cell technologies, biological systems, and disease contexts warrants further investigation. Extensive validation across diverse experimental settings and biological questions will be necessary to establish the model's robustness and broad applicability in the rapidly evolving field of single-cell multi-omics research.

Chapter 6

Conclusions

As discussed in Chapter 3 and Chapter 4, affinity graphs are a particularly effective option in semi-supervised and self-supervised learning due to several key advantages they offer in capturing and utilizing the underlying structures within the data:

- **Capturing Relationships and Similarities**. Affinity graphs represent data points as nodes and the pairwise similarities or affinities between them as edges. This structure naturally captures the relationships between data points based on their feature similarities. By encoding these relationships, affinity graphs help the model understand how data points are connected, even in the absence of labels. This is crucial in semi-supervised and self-supervised settings, where labeled data is sparse or unavailable. The graph structure allows the model to propagate information from labeled to unlabeled data effectively, leveraging the intrinsic data manifold for better learning.

- **Enhanced Representation Learning**. By compelling similar nodes to exhibit congruent representations, they enforce a structured regularization on the feature extraction process, fostering representations that are not only discriminative but also highly transferrable across different tasks. This mechanism is pivotal, for instance, in the domain of computer vision, where the alignment of comparable image patches in the embedding space promotes the extraction of salient and transferable visual features.

- **Improved Clustering and Segmentation**. Affinity graphs enable the implementation of clustering techniques that group similar data points together. This is particularly useful in segmentation tasks where the goal is to partition the data into meaningful segments. By leveraging the graph structure, models can achieve better segmentation results as they can more accurately delineate

the boundaries between different clusters or segments. This is crucial in applications like medical imaging, where precise segmentation of anatomical structures is essential.

- **Scalability and Flexibility**. Affinity graphs can be scaled to large datasets by using efficient graph construction and processing techniques. They can also be integrated with various types of neural networks, including convolutional and transformer-based architectures. This scalability and flexibility make affinity graphs suitable for a wide range of applications and data types, from small-scale datasets to large, high-dimensional data like whole-slide images in pathology. This scalability and flexibility make affinity graphs suitable for a wide range of applications and data types, from small-scale datasets to large, high-dimensional data like whole-slide images in pathology.

- **Effective Regularization**. Affinity graphs impose regularization constraints on the learning process by enforcing smoothness over the graph. This means that the model is encouraged to produce similar outputs for similar inputs, as defined by the graph. This regularization helps prevent overfitting, especially in scenarios with limited labeled data. It ensures that the model generalizes better to unseen data, which is a key challenge in semi-supervised and self-supervised learning.

We are now able to answer the research questions in Section 1.2 based on the content of this thesis:

RQ1: *What mathematical representations can be utilized to efficiently elucidate and encode the implicit constraints within input data structures?* In this thesis, we propose to use affinity graph as the representation to efficiently elucidate and encode the implicit constraints. An affinity graph is a mathematical representation that captures relationships between elements in data. It is represented as an adjacency matrix W of size $n \times n$, where n is the number of elements. Each entry W_{ij} is typically a non-negative real number in the matrix represents the similarity or affinity between elements i and j. Higher values of W_{ij} indicate a closer relationship or stronger affinity between elements i and j. The affinity between two elements is often calculated using the dot product between their vector representations where the vector representations are typically learned by the model during training.

RQ2: *How this representation (in RQ1) applied into the learning process?*

Affinity graphs are used to model the intrinsic structure and constraints present in the data. The edge weights of the affinity graph capture the strength of relationships between data points. By incorporating affinity graphs into the learning process, the implicit spatial structures within the input data itself can be leveraged to guide representation learning and improve annotation efficiency.

In self-supervised learning on whole slide images, an affinity graph constraint is proposed that establishes connections between different Transformer layers. This inter-layer constraint allows implicit structures to be propagated and refined across the model architecture, enabling the capture of more detailed features with limited annotations.

For semi-supervised learning in medical image segmentation, an affinity-graph-guided contrastive learning framework is introduced. Affinity graphs are constructed between pseudo-labels of student and teacher networks to align their outputs and capture the inherent structure of the data. This implicit structural constraint enhances the discriminative ability of the learned representations.

RQ3: *Can this representations (in RQ1) extract reliable implicit constraints from multiple source?*

Besides the diverse modalities of medical images, we also applied the affinity graph into bio-informatics area to perform multi-modal predicition on single-cell data. We proposed the scAGFormer model to use an affinity graph prior to learn interactions between different modalities (e.g. gene expression and protein levels) without relying on cell embeddings. This allowed it to outperform existing methods on modality prediction tasks.

Therefore, the thesis demonstrates affinity graphs can capture different types of implicit constraints, including spatial relationships in images, semantic similarities between patches/regions, and cross-modality interactions in single-cell data. We believe the affinity graph is able to extract reliable implicit constraints from multiple source.

6.1 Discussion

This thesis offers a thorough investigation of affinity graph constraints (AGC) within self-supervised, semi-supervised, and multi-modal learning frameworks. The methodologies introduced here mark significant progress in annotation-efficient learning.

However, several promising avenues remain to further advance this research:Expanding Beyond Transformer Architectures: Chapter 3 reveals that AGC is closely integrated with Transformer-based models, potentially limiting its application to other neural network architectures. Future research could adapt AGC for use with CNNs or hybrid models, thereby increasing its versatility across different frameworks.

Should affinity graphs be used with all transformer architectures? The primary motivation for employing affinity graphs within transformer-based frameworks is indeed the pursuit of fine-grained alignment, which serves as a crucial mechanism for enhancing the learning process in a variety of tasks. The core advantage of affinity graphs lies in their ability to model intricate relationships and dependencies between data points, features, or modalities, enabling precise and contextual learning. The concept of fine-grained alignment, as applied across different domains, highlights the versatility and incremental innovation within transformer architectures and attention mechanisms.

In semi-supervised learning, the alignment between pseudo labels generated by student and teacher networks serves as a critical regularization mechanism. Affinity graphs facilitate this alignment by creating a structural representation of pseudo-label relationships. The edges in the affinity graph encode the similarity between pseudo labels, and optimizing the graph structure ensures consistency and coherence between the outputs of the student and teacher networks. This approach not only enhances the discriminative capability of the model but also minimizes noise and errors in pseudo-label propagation, especially when labeled data is sparse.

Multi-modal learning tasks often involve disparate data representations, such as statistical features (e.g., gene expression levels) and representational features (e.g., embeddings from a neural network). Here, affinity graphs enable the alignment of these heterogeneous representations by capturing the intricate interactions and dependencies between modalities. For instance, in multi-modal single-cell analysis, affinity graphs are employed to align statistical and representation features, bridging the gap between these fundamentally different feature types. This alignment ensures that the model captures meaningful cross-modality relationships, leading to better integration and prediction.

I believe there is an incremental improvement over different attention mechanism to improve the scalablity. Early transformers relied on global self-attention, which captures relationships across all tokens. However, this approach is computationally expensive for large inputs, such as images or long sequences. Sparse attention [55] and local attention mechanisms [1], which focus on subsets of tokens, provide a more

efficient alternative. Affinity graphs align naturally with these methods by encoding only the most relevant relationships, reducing computational overhead while retaining critical information.

Cross-attention mechanisms, as seen in models like BLIP [160, 159] and DALL-E [212] align inputs from different modalities. Affinity graphs enhance cross-attention by providing a structural prior that guides the alignment process, ensuring that the relationships between modalities are faithfully captured.

Attention-Free Transformers: Emerging models like Performer [56] and Linformer [265] attempt to replace the standard attention mechanism with linear approximations or kernel methods for efficiency. While these approaches reduce complexity, they still benefit from the fine-grained alignment facilitated by affinity graphs, which can act as an external regularizer to compensate for the reduced capacity of simplified attention mechanisms.

Theoretical exploration and framework of Affinity Graph and beyond. Affinity graph has achieved great success in semi-supervised learning and self-supervised learning. However, such success initially stemmed from an in-depth analysis of the scalability and reliability of the Transformer structure. There are still valuable open questions for the structure inside the Transformer to answer:

- What is the theoretical framework for the Transformer Block?
- Why aligning key and query embedding works in Transformer?
- Why scaling Transformer in natural language works while failed in computer vision?[1]

Figure 6.1 has demonstrated what is the Transformer Block. The Transformer Block has been considered as a fundamental component of Transformer. The overall Transformer model is just repeat or extend the Transformer Block. There is no doubt that the theoretical framework for the Transformer Block is a valuable and meaningful contribution to the machine learning community. Recent research has setup a basic theoretical framework for Transformer [89]. However, such work still does answer any of the raised questions. The theoretical exploration and the framework for affinity graph could further contribute to the theoretical framework of Transformer from representation learning and supervised signal utilization perspective.

[1]Recently Google has scale the ViT to 22 Billion parameters [68]. ViT-22B is just an incremental improvement. Therefore, scaling a visual Transformer to a larger model size is as successful as scaling a Transformer in natural language processing tasks.

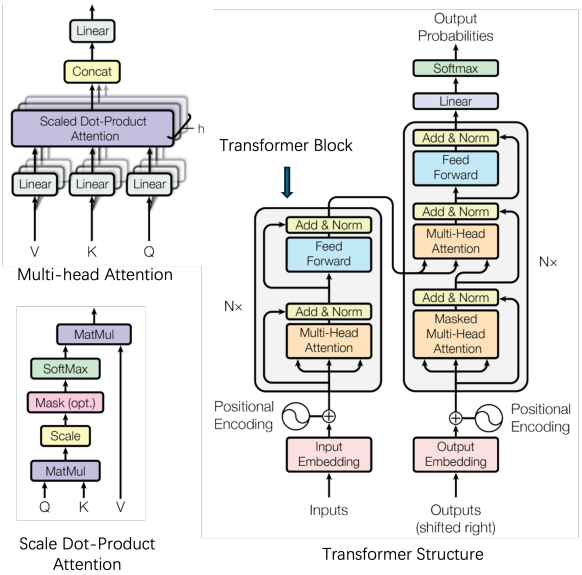

Figure 6.1: The Transformer block. The Transformer structure and its component are taken from original paper [253].

Alignment as learning paradigm. The affinity graph defined in Section 2.6 is just the beginning of the application of the affinity graph. The form of the affinity graph could be applied beyond the form of a Transformer. For example, our preliminary exploration mentioned in Chapter 5 applied an affinity graph to bridge the representation of two modalities in multi-modality learning. From the perspective of representation learning, graph representations are just one way to build affinity alignment between two subjects and the affinity graph is a well-defined representation module to construct the alignment of two subjects. Under this form, the aligned two subjects could be queries and keys in the multi-head self-attention, and then scale up to the Transformer model. Therefore, I believe that starting with affinity graphs and expanding on alignment as a new learning paradigm will be a more promising research direction.

Affinity graph in Multi-modal Large-language Models. Large-language models (LLMs) are a deep learning architecture featured by its vast scale, encompassing billions of parameters and trained on extensive datasets [26], enabling it to capture intricate patterns in language, model complex linguistic relationships, and generalize across a wide range of natural language tasks with greater accuracy and contextual understanding [310, 193]. OpenAI has first scaled the Transformer structure into billion-scale parameters on GPT-3 [26] and has shown great success in many downstream tasks. Motivated by the success of GPT-3, a multi-modal large language

model that trained language model with visual encoder has emerged as a new structure to handle complicated visual features [174, 173, 175, 301, 160]. However, how to effectively construct a reliable representation for both visual context and natural language input is still an open question. BLIP [160] and BLIP 2 [159] have discussed and explored a joint training framework to utilize the embeddings from visual encoders and the language models to achieve multi-modal processing tasks like visual question answering (VQA) [4], video summarization [99] and visual grounding [142]. In Chapter 5, the affinity graph has shown its superiority in bridging the modalities in the single-cell analysis domain. It's worth applying an affinity graph to bridge two modalities in a larger domain, such as the multi-modal LLMs. An affinity graph could also serve as a potential option to bridge the visual encoder and the language model to construct the multi-modal language model.

Bibliography

[1] Ignacio Aguilera-Martos, Andrés Herrera-Poyatos, Julián Luengo, and Francisco Herrera. Local attention mechanism: Boosting the transformer architecture for long-sequence time series forecasting. *arXiv preprint arXiv: 2410.03805*, 2024.

[2] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[3] Alexander Andreopoulos and John K Tsotsos. Efficient and generalizable statistical models of shape and appearance for analysis of cardiac mri. *Medical image analysis*, 12(3):335–357, 2008.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[5] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *IJCNN*, pages 1–8, 2020.

[6] Sercan Ö. Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. *ArXiv*, abs/1908.07442, 2019.

[7] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3478–3488, 2021.

[8] Chiori Azuma, Tomoyoshi Ito, and Tomoyoshi Shimobaba. Adversarial domain adaptation using contrastive learning. *Engineering Applications of Artificial Intelligence*, 123:106394, 2023.

[9] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. 32, 2019.

[10] Alexei Baevski, Henry Zhou, Abdel rahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *ArXiv*, abs/2006.11477, 2020.

[11] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. A cookbook of self-supervised learning. *arXiv preprint arXiv: Arxiv-2304.12210*, 2023.

[12] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *International Conference on Learning Representations*, 2021.

[13] W Bao, Y Jin, C Huang, and W Peng. Ct image classification of invasive depth of gastric cancer based on 3d-dpn structure. In *The 11th International Workshop on Computer Science and Engineering (WCSE 2021)*, pages 115–121, 2021.

[14] Adrien Bardes, Jean Ponce, and Yann LeCun. VICRegl: Self-supervised learning of local visual features. In *NIPS*, 2022.

[15] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 2017.

[16] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15:1373–1396, 2003.

[17] Yoshua Bengio, Olivier Delalleau, and Clarence Simard. Decision trees do not generalize to new variations. *Computational Intelligence*, 26(4):449–467, 2010.

[18] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, 2018.

[19] Hakan Bilen and Andrea Vedaldi. Universal representations: The missing link between faces, text, planktons, and cat breeds. *ArXiv*, abs/1701.07275, 2017.

[20] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84:102680, 2023.

[21] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[22] Gerda Bortsova, Florian Dubost, Laurens Hogeweg, Ioannis Katramados, and Marleen De Bruijne. Semi-supervised medical image segmentation via learning consistency under transformations. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pages 810–818. Springer, 2019.

[23] Behzad Bozorgtabar, Dwarikanath Mahapatra, Guillaume Vray, and Jean-Philippe Thiran. Salad: Self-supervised aggregation learning for anomaly detection on x-rays. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 468–478. Springer, 2020.

[24] Adrian Brady, Risteárd Ó Laoide, Peter McCarthy, and Ronan McDermott. Discrepancy and error in radiology: concepts, causes and consequences. *The Ulster medical journal*, 81(1):3, 2012.

[25] Amy Brand, Liz Allen, Micah Altman, Marjorie Hlava, and Jo Scott. Beyond authorship: Attribution, contribution, collaboration, and credit. *Learned Publishing*, 28(2), 2015.

[26] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[27] Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nature Methods*, 10(12):1213–1218, 2013.

[28] Elizabeth Bullitt, Donglin Zeng, Guido Gerig, Stephen Aylward, Sarang Joshi, J Keith Smith, Weili Lin, and Matthew G Ewend. Vessel tortuosity and brain tumor malignancy: a blinded study1. *Academic radiology*, 12(10):1232–1240, 2005.

[29] Philippe Burlina, William Paul, Philip Mathew, Neil Joshi, Katia D Pacheco, and Neil M Bressler. Low-shot deep learning of diabetic retinopathy with potential applications to address artificial intelligence bias in retinal diagnostics and rare ophthalmic diseases. *JAMA ophthalmology*, 138(10):1070–1077, 2020.

[30] Vivien A. Cabannes, Léon Bottou, Yann LeCun, and Randall Balestriero. Active self-supervised learning: A few low-cost relationships are all you need. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16228–16237, 2023.

[31] Jian-Feng Cai, Emmanuel J. Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.*, 20:1956–1982, 2008.

[32] Qi Cai, Yu Wang, Yingwei Pan, Ting Yao, and Tao Mei. Joint contrastive learning with infinite possibilities. *ArXiv*, abs/2009.14776, 2020.

[33] Junyue Cao, Darren A Cusanovich, Vijay Ramani, Delasa Aghamirzaie, Hannah A Pliner, Andrew J Hill, Riza M Daza, Jose L McFaline-Figueroa, Jonathan S Packer, Lena Christiansen, et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, 361(6409):1380–1385, 2018.

[34] Aaron Carass, Snehashis Roy, Amod Jog, Jennifer L Cuzzocreo, Elizabeth Magrath, Adrian Gherman, Julia Button, James Nguyen, Ferran Prados, Carole H

Sudre, et al. Longitudinal multiple sclerosis lesion segmentation: resource and challenge. *NeuroImage*, 148:77–102, 2017.

[35] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[36] Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *ICCV*, pages 9630–9640, 2021.

[37] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *NeurIPS*, 33:12546–12558, 2020.

[38] Changyou Chen, Jianyi Zhang, Yi Xu, Liqun Chen, Jiali Duan, Yiran Chen, S. Tran, Belinda Zeng, and Trishul M. Chilimbi. Why do we need large batchsizes in contrastive learning? a gradient-bias perspective. In *Neural Information Processing Systems*, 2022.

[39] Cheng Chen, Kangneng Zhou, Zhiliang Wang, and Ruoxiu Xiao. Generative consistency for semi-supervised cerebrovascular segmentation from tof-mra. *IEEE Transactions on Medical Imaging*, 42(2):346–353, 2022.

[40] Chengkuan Chen, Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Andrew J Schaumberg, and Faisal Mahmood. Fast and scalable search of whole-slide images via self-supervised deep learning. *Nature Biomedical Engineering*, 6(12):1420–1434, 2022.

[41] Gaoxiang Chen, Jintao Ru, Yilin Zhou, Islem Rekik, Zhifang Pan, Xiaoming Liu, Yezhi Lin, Beichen Lu, and Jialin Shi. Mtans: multi-scale mean teacher combined adversarial network with shape-aware embedding for semi-supervised brain lesion segmentation. *NeuroImage*, 244:118568, 2021.

[42] Jun Chen, Heye Zhang, Raad Mohiaddin, Tom Wong, David Firmin, Jennifer Keegan, and Guang Yang. Adaptive hierarchical dual consistency for semi-supervised left atrium segmentation on cross-domain data. *IEEE Transactions on Medical Imaging*, 41(2):420–433, 2021.

[43] Kaifeng Chen, Daniel Salz, Huiwen Chang, Kihyuk Sohn, Dilip Krishnan, and Mojtaba Seyedhosseini. Improve supervised representation learning with masked image modeling. *ArXiv*, abs/2312.00950, 2023.

[44] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin P. Murphy, and Alan Loddon Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *PAMI*, 40:834–848, 2016.

[45] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *CVPR*, pages 16144–16155, 2022.

[46] Song Chen, Blue B Lake, and Kun Zhang. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature Biotechnology*, 37(12):1452–1457, 2019.

[47] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020.

[48] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *ArXiv*, abs/2002.05709, 2020.

[49] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[50] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, pages 15750–15758, 2021.

[51] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9620–9629, 2021.

[52] Yuhao Chen, X. Tan, Borui Zhao, Zhaowei Chen, Renjie Song, Jiajun Liang, and Xuequan Lu. Boosting semi-supervised learning by exploiting all unlabeled data. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7548–7557, 2023.

[53] C.-L. Cheng, Shalabh, and Gaurav Garg. Coefficient of determination for multiple measurement error models. *J. Multivar. Anal.*, 126:137–152, 2014.

[54] Davide Chicco, Matthijs J. Warrens, and Giuseppe Jurman. The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *PeerJ Computer Science*, 7, 2021.

[55] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

[56] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking attention with performers. *arXiv preprint arXiv: 2009.14794*, 2020.

[57] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *ArXiv*, abs/2007.00224, 2020.

[58] Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198, 2022.

[59] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26:1045–1057, 2013.

[60] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

[61] Carlos Coello-Coello. Theoretical and numerical constraint-handling techniques used with evolutionary algorithms: a survey of the state of the art. *Computer Methods in Applied Mechanics and Engineering*, 2002.

[62] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gómez-Cabrero, Alejandra Cervera, Andrew W McPherson, Michał Wojciech Szcześniak, Daniel J. Gaffney, Laura L. Elo, Xuegong Zhang, and Ali Mortazavi. A survey of best practices for rna-seq data analysis. *Genome Biology*, 17, 2016.

[63] Stefan Cornelissen, Joost A van der Putten, TGW Boers, Jelmer B Jukema, Kiki N Fockens, JJGHM Bergman, Fons van der Sommen, and PHN de With. Evaluating self-supervised learning methods for downstream classification of neoplasia in barrett's esophagus. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 66–70. IEEE, 2021.

[64] Haotian Cui, Chloe Wang, Hassaan Maan, Nan Duan, and Bo Wang. scFormer: A universal representation learning approach for single-cell data using transformers. *bioRxiv*, pages 2022–11, 2022.

[65] Haotian Cui, Chloe Wang, Hassaan Maan, and Bo Wang. scGPT: Towards building a foundation model for single-cell multi-omics using generative ai. *bioRxiv*, pages 2023–04, 2023.

[66] Andrew Benz Peter Holderrieth Jonathan Bloom Christopher Lance Ashley Chow Ryan Holbrook Daniel Burkhardt, Malte Luecken. Open problems - multimodal single-cell integration, 2022.

[67] Olivier Dehaene, Axel Camara, Olivier Moindrot, Axel de Lavergne, and Pierre Courtiol. Self-supervision closes the gap between weak and strong supervision in histology. *arXiv preprint arXiv:2012.03583*, 2020.

[68] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd van Steenkiste, Gamaleldin F. Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Patrick Collier, Alexey Gritsenko, Vighnesh Birodkar, Cristina Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetić, Dustin Tran, Thomas Kipf, Mario Lučić, Xiaohua Zhai, Daniel Keysers, Jeremiah Harmsen, and Neil Houlsby. Scaling vision transformers to 22 billion parameters. *arXiv preprint arXiv: 2302.05442*, 2023.

[69] Chaitanya Devaguptapu, Sumukh Aithal, Shrinivas Ramasubramanian, Moyuru Yamada, and Manohar Kaul. Semantic graph consistency: Going beyond patches for regularizing self-supervised vision transformers. *arXiv preprint arXiv: 2406.12944*, 2024.

[70] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019.

[71] Haohua Dong, Yutaro Iwamoto, Xianhua Han, Lanfen Lin, Hongjie Hu, Xiujun Cai, and Yen-Wei Chen. Case discrimination: Self-supervised feature learning for the classification of focal liver lesions. In *Innovation in Medicine and Healthcare: Proceedings of 9th KES-InMed 2021*, pages 241–249. Springer, 2021.

[72] Nanqing Dong and Irina Voiculescu. Federated contrastive learning for decentralized unlabeled medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 378–387. Springer, 2021.

[73] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020.

[74] Alexey et al. Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[75] Jingcheng Du, Peilin Jia, Yulin Dai, Cui Tao, Zhongming Zhao, and Degui Zhi. Gene2vec: distributed representation of genes based on co-expression. *BMC Genomics*, 20, 2018.

[76] Zhana Duren, Xi Chen, Mahdi Zamanighomi, Wanwen Zeng, Ansuman T Satpathy, Howard Y Chang, Yong Wang, and Wing Hung Wong. Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *Proceedings of the National Academy of Sciences*, 115(30):7723–7728, 2018.

[77] Gökçen Eraslan, Lukas M. Simon, Maria Mircea, Nikola S Mueller, and Fabian J Theis. Single-cell rna-seq denoising using a deep count autoencoder. *Nature Communications*, 10, 2019.

[78] Pascal Esser, Satyaki Mukherjee, and Debarghya Ghoshdastidar. Representation learning dynamics of self-supervised models. *arXiv preprint arXiv: 2309.02011*, 2023.

[79] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231, 1996.

[80] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark A. DePristo, Katherine Chou, Claire Cui, Greg S. Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature Medicine*, 25:24 – 29, 2019.

[81] Alex Fedorov, Eloy Geenjaar, Lei Wu, Thomas P DeRamus, Vince D Calhoun, and Sergey M Plis. Tasting the cake: evaluating self-supervised generalization on out-of-distribution multimodal mri data. *arXiv preprint arXiv:2103.15914*, 2021.

[82] Alex Fedorov, Lei Wu, Tristan Sylvain, Margaux Luck, Thomas P DeRamus, Dmitry Bleklov, Sergey M Plis, and Vince D Calhoun. On self-supervised multimodal representation learning: an application to alzheimer's disease. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1548–1552. IEEE, 2021.

[83] Rory A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Human Genetics*, 7:179–188, 1936.

[84] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[85] Yaroslav Ganin, E. Ustinova, Hana Ajakan, Pascal Germain, H. Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. Domain-adversarial training of neural networks. In *Journal of machine learning research*, 2015.

[86] Jianliang Gao, Ling Tian, Tengfei Lv, Jianxin Wang, Bo Song, and Xiaohua Hu. Protein2vec: Aligning multiple ppi networks with representation learning. *IEEE/ACM transactions on computational biology and bioinformatics*, 18(1):240–249, 2019.

[87] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *Conference on Empirical Methods in Natural Language Processing*, 2021.

[88] Matej Gazda, Ján Plavka, Jakub Gazda, and Peter Drotar. Self-supervised deep convolutional neural network for chest x-ray classification. *IEEE Access*, 9:151972–151982, 2021.

[89] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *arXiv preprint arXiv: 2312.10794*, 2023.

[90] Boying Gong, Yun Zhou, and Elizabeth Purdom. Cobolt: integrative analysis of multimodal single-cell sequencing data. *Genome Biology*, 22(1):1–21, 2021.

[91] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[92] Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, and Piotr Bojanowski. Self-supervised pretraining of visual features in the wild. *ArXiv*, abs/2103.01988, 2021.

[93] Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised constrastive learning. In *ICML*, 2021.

[94] Simon Graham, Hao Chen, Qi Dou, Pheng-Ann Heng, and Nasir M. Rajpoot. Mild-net: Minimal information loss dilated network for gland instance segmentation in colon histology images. *Medical Image Analysis*, 52:199–211, 2018.

[95] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *NeurIPS*, 33:21271–21284, 2020.

[96] Haocheng Gu, Hao Cheng, Anjun Ma, Yang Li, Juexin Wang, Dong Xu, and Qin Ma. scGNN 2.0: a graph neural network tool for imputation and clustering of single-cell rna-seq data. *Bioinformatics*, 38(23):5322–5325, 2022.

[97] Jiuxiang Gu, Jason Kuen, Shafiq R. Joty, Jianfei Cai, Vlad I. Morariu, Handong Zhao, and Tong Sun. Self-supervised relationship probing. In *Neural Information Processing Systems*, 2020.

[98] Ran Gu, Jingyang Zhang, Guotai Wang, Wenhui Lei, Tao Song, Xiaofan Zhang, Kang Li, and Shaoting Zhang. Contrastive semi-supervised learning for domain adaptive segmentation across similar anatomical structures. *IEEE Transactions on Medical Imaging*, 42:245–256, 2022.

[99] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*, pages 505–520. Springer, 2014.

[100] Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Zongwei Zhou, Michael B Gotway, and Jianming Liang. Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning. *IEEE transactions on medical imaging*, 40(10):2857–2868, 2021.

[101] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

[102] Kai Han, Lu Liu, Yuqing Song, Yi Liu, Chengjian Qiu, Yangyang Tang, Qiaoying Teng, and Zhe Liu. An effective semi-supervised approach for liver ct image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 26(8):3999–4007, 2022.

[103] Heng Hao, Sima Didari, Jae Oh Woo, Hankyu Moon, and Patrick Bangert. Highly efficient representation and active learning framework for imbalanced data and its application to covid-19 x-ray classification. *NeurIPS*, 2021.

[104] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *CVPR*, pages 15979–15988, 2021.

[105] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020.

[106] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2019.

[107] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[108] Xinran He, Junfeng Pan, Ou Jin, Tianbing Xu, Bo Liu, Tao Xu, Yanxin Shi, Antoine Atallah, Ralf Herbrich, Stuart Bowers, and Joaquin Quiñonero Candela. Practical lessons from predicting clicks on ads at facebook. In *International Workshop on Data Mining for Online Advertising*, 2014.

[109] Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, et al. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*, 2019.

[110] Brian L. Hie, Bryan D. Bryson, and Bonnie Berger. Efficient integration of heterogeneous single-cell transcriptomes using scanorama. *Nature Biotechnology*, 37:685–691, 2019.

[111] Geoffrey E. Hinton and Sam T. Roweis. Stochastic neighbor embedding. In *Neural Information Processing Systems*, 2002.

[112] Hong-Hai Hoang and Hoang Trinh. Improvement for convolutional neural networks in image classification using long skip connection. *Applied Sciences*, 2021.

[113] William Hogan, Jiacheng Li, and Jingbo Shang. Open-world semi-supervised generalized relation discovery aligned in a real-world setting. *arXiv preprint arXiv: 2305.13533*, 2023.

[114] Wan-Ting Hsieh, Jeremy Lefort-Besnard, Hao-Chun Yang, Li-Wei Kuo, and Chi-Chun Lee. Behavior score-embedded brain encoder network for improved classification of alzheimer disease using resting state fmri. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 5486–5489. IEEE, 2020.

[115] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdel rahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

[116] Stanley Bryan Z Hua, Alex X Lu, and Alan M Moses. Cytoimagenet: A large-scale pretraining dataset for bioimage transfer learning. *arXiv preprint arXiv:2111.11646*, 2021.

[117] Jin Huang, Wentai Zhu, Jing Xiao, Tian Lu, and Weihao Yu. Dynamic graph representation based on temporal and contextual contrasting. 2022.

[118] Wei Huang, Chang Chen, Zhiwei Xiong, Yueyi Zhang, Xuejin Chen, Xiaoyan Sun, and Feng Wu. Semi-supervised neuron segmentation via reinforced consistency learning. *IEEE Transactions on Medical Imaging*, 41(11):3016–3028, 2022.

[119] Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar S. Karnin. Tabtransformer: Tabular data modeling using contextual embeddings. *ArXiv*, abs/2012.06678, 2020.

[120] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

[121] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[122] Nahid Ul Islam, Shiv Gehlot, Zongwei Zhou, Michael B Gotway, and Jianming Liang. Seeking an optimal approach for computer-aided pulmonary embolism detection. In *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12*, pages 692–702. Springer, 2021.

[123] Sebastian Jaszczur, Aakanksha Chowdhery, Afroz Mohiuddin, Lukasz Kaiser, Wojciech Gajewski, Henryk Michalewski, and Jonni Kanerva. Sparse is enough in scaling transformers. *NeurIPS*, 34:9895–9907, 2021.

[124] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *MultiMedia modeling: 26th international conference, MMM*

2020, Daejeon, South Korea, January 5–8, 2020, proceedings, part II 26, pages 451–462. Springer, 2020.

[125] Zhanghexuan Ji, Mohammad Abuzar Shaikh, Dana Moukheiber, Sargur N Srihari, Yifan Peng, and Mingchen Gao. Improving joint learning of chest x-ray and radiology report by word region alignment. In *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12*, pages 110–119. Springer, 2021.

[126] Guo-Zhang Jian, Guo-Shiang Lin, Chuin-Mu Wang, and Sheng-Lei Yan. Helicobacter pylori infection classification based on convolutional neural network and self-supervised learning. In *Proceedings of the 5th International Conference on Graphics and Signal Processing*, pages 60–64, 2021.

[127] Ruijie Jiang, Thuan Q. Nguyen, Prakash Ishwar, and Shuchin Aeron. Supervised contrastive learning with hard negative samples. *ArXiv*, abs/2209.00078, 2022.

[128] Jianbo Jiao, Yifan Cai, Mohammad Alsharid, Lior Drukker, Aris T Papageorghiou, and J Alison Noble. Self-supervised contrastive video-speech representation learning for ultrasound. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*, pages 534–543. Springer, 2020.

[129] Suoqin Jin, Lihua Zhang, and Qing Nie. scai: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. *Genome Biology*, 21:1–19, 2020.

[130] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. volume 43, pages 4037–4058. IEEE, 2020.

[131] Peiguang Jing, Yuting Su, Zhengnan Li, and Liqiang Nie. Learning robust affinity graph representation for multi-view clustering. *Inf. Sci.*, 544:155–167, 2021.

[132] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng.

Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

[133] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[134] Rico Jonschkowski and Oliver Brock. Learning state representations with robotic priors. *Autonomous Robots*, 39:407 – 428, 2015.

[135] Aakash Kaku, Sahana Upadhya, and Narges Razavian. Intermediate layers matter in momentum contrastive self supervised learning. *Advances in Neural Information Processing Systems*, 34:24063–24074, 2021.

[136] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *NeurIPS*, 33:21798–21809, 2020.

[137] Uday Kamal, Mohammad Zunaed, Nusrat Binta Nizam, and Taufiq Hasan. Anatomy-xnet: An anatomy aware convolutional neural network for thoracic disease classification in chest x-rays. *IEEE Journal of Biomedical and Health Informatics*, 26(11):5518–5528, 2022.

[138] Rashed Karim, R James Housden, Mayuragoban Balasubramaniam, Zhong Chen, Daniel Perry, Ayesha Uddin, Yosra Al-Beyatti, Ebrahim Palkhi, Prince Acheampong, Samantha Obom, et al. Evaluation of current algorithms for segmentation of scar tissue from late gadolinium enhancement cardiovascular magnetic resonance of the left atrium: an open-access grand challenge. *Journal of Cardiovascular Magnetic Resonance*, 15(1):105, 2013.

[139] Narayanan Kasthuri, Kenneth Jeffrey Hayworth, Daniel Raimund Berger, Richard Lee Schalek, José Angel Conchello, Seymour Knowles-Barley, Dongil Lee, Amelio Vázquez-Reina, Verena Kaynig, Thouis Raymond Jones, et al. Saturated reconstruction of a volume of neocortex. *Cell*, 162(3):648–661, 2015.

[140] Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue. *Zenodo10*, 5281, 2018.

[141] Sota Kato and Kazuhiro Hotta. Adaptive t-vmf dice loss for multi-class medical image segmentation. *arXiv preprint arXiv:2207.07842*, 2022.

[142] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

[143] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *NIPS*, 2017.

[144] Jing Ke, Yiqing Shen, Xiaoyao Liang, and Dinggang Shen. Contrastive learning based stain normalization across multiple tumor in histopathology. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, pages 571–580. Springer, 2021.

[145] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7482–7491, 2017.

[146] Nawid Keshtmand, Raúl Santos-Rodríguez, and J. Lawry. Understanding the properties and limitations of contrastive learning for out-of-distribution detection. *ICPR Workshops*, 2022.

[147] Jaeill Kim, Duhun Hwang, Eunjung Lee, Jangwon Suh, Jimyeong Kim, and Wonjong Rhee. Enhancing contrastive learning with efficient combinatorial positive pairing. *arXiv preprint arXiv: 2401.05730*, 2024.

[148] Kwang In Kim, James Tompkin, Hanspeter Pfister, and Christian Theobalt. Semi-supervised learning with explicit relationship regularization. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2188–2196, 2015.

[149] Yuri Kotliarov, Rachel Sparks, Andrew J Martins, Matthew P Mulè, Yong Lu, Meghali Goswami, Lela Kardava, Romain Banchereau, Virginia Pascual, Angélique Biancotto, et al. Broad immune activation underlies shared set point

signatures for vaccine responsiveness in healthy individuals and disease activity in patients with lupus. *Nature Medicine*, 26(4):618–629, 2020.

[150] Mark A. Kramer. Autoassociative neural networks. *Computers & Chemical Engineering*, 16:313–328, 1992.

[151] Rayan Krishnan, Pranav Rajpurkar, and Eric J. Topol. Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering*, 6:1346 – 1352, 2022.

[152] Denis Krompass, Stephan Baier, and Volker Tresp. Type-constrained representation learning in knowledge graphs. In *International Workshop on the Semantic Web*, 2015.

[153] Christopher Lance, Malte D. Luecken, Daniel B. Burkhardt, Robrecht Cannoodt, Pia Rautenstrauch, Anna Laddach, Aidyn Ubingazhibov, Zhi-Jie Cao, Kaiwen Deng, Sumeer Ahmad Khan, Qiao Liu, Nikolay Russkikh, G. E. Ryazantsev, Uwe Ohler, Angela Oliveira Pisco, Jonathan Bloom, Smita Krishnaswamy, and Fabian J. Theis. Multimodal single cell data integration challenge: results and lessons learned. *bioRxiv*, 2022.

[154] Dong-Hyun Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. 2013.

[155] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021.

[156] Caizi Li, Li Dong, Qi Dou, Fan Lin, Kebao Zhang, Zuxin Feng, Weixin Si, Xuesong Deng, Zhe Deng, and Pheng-Ann Heng. Self-ensembling co-training framework for semi-supervised covid-19 ct segmentation. *IEEE Journal of Biomedical and Health Informatics*, 25(11):4140–4151, 2021.

[157] Jiajun Li, Tiancheng Lin, and Yi Xu. SSLP: Spatial guided self-supervised learning on pathological images. In *MICCAI*, pages 3–12. Springer, 2021.

[158] Jinpeng Li, Gangming Zhao, Yaling Tao, Penghua Zhai, Hao Chen, Huiguang He, and Ting Cai. Multi-task contrastive learning for automatic ct and x-ray diagnosis of covid-19. *Pattern recognition*, 114:107848, 2021.

[159] Junnan Li, Dongxu Li, S. Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *International Conference on Machine Learning*, 2023.

[160] Junnan Li, Dongxu Li, Caiming Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *International Conference on Machine Learning*, 2022.

[161] Lei Li, Veronika A Zimmer, Julia A Schnabel, and Xiahai Zhuang. Atrialgeneral: domain generalization for left atrial segmentation of multi-center lge mris. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24*, pages 557–566. Springer, 2021.

[162] Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. *ArXiv*, abs/1801.07606, 2018.

[163] Shuailin Li, Chuyu Zhang, and Xuming He. Shape-aware semi-supervised 3D semantic segmentation for medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2020.

[164] Wenbin Li, Zhichen Fan, Jing Huo, and Yang Gao. Modeling inter-class and intra-class constraints in novel class discovery. In *CVPR*, pages 3449–3458, 2023.

[165] Xiang Li, Minglei Li, Pengfei Yan, Guanyi Li, Yuchen Jiang, Hao Luo, and Shen Yin. Deep learning attention mechanism in medical image analysis: Basics and beyonds. *International Journal of Network Dynamics and Intelligence*, 2023.

[166] Xiaomeng Li, Xiaowei Hu, Xiaojuan Qi, Lequan Yu, Wei Zhao, Pheng-Ann Heng, and Lei Xing. Rotation-oriented collaborative self-supervised learning for retinal disease diagnosis. *IEEE Transactions on Medical Imaging*, 40(9):2284–2294, 2021.

[167] Xiaomeng Li, Mengyu Jia, Md Tauhidul Islam, Lequan Yu, and Lei Xing. Self-supervised feature learning via exploiting multi-modal data for retinal disease diagnosis. *IEEE Transactions on Medical Imaging*, 39(12):4023–4033, 2020.

[168] Tiancheng Lin, Zhimiao Yu, Zengchao Xu, Hongyu Hu, Yi Xu, and Chang-Wen Chen. SGCL: Spatial guided contrastive learning on whole-slide pathological images. *Medical Image Analysis*, 89:102845, Oct 2023.

[169] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE International Conference on Computer Vision*, 2017.

[170] Geert Litjens, Robert Toth, Wendy Van De Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram Van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18(2):359–373, 2014.

[171] Andy T. Liu, Shang-Wen Li, and Hung yi Lee. Tera: Self-supervised learning of transformer encoder representation for speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2351–2366, 2020.

[172] Fengbei Liu, Yu Tian, Filipe R Cordeiro, Vasileios Belagiannis, Ian Reid, and Gustavo Carneiro. Self-supervised mean teacher for semi-supervised chest x-ray classification. In *International Workshop on Machine Learning in Medical Imaging*, pages 426–436. Springer, 2021.

[173] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[174] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.

[175] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[176] Jihao Liu, Xin Huang, Jinliang Zheng, Yu Liu, and Hongsheng Li. Mixmim: Mixed and masked image modeling for efficient visual representation learning. *ArXiv*, abs/2205.13137, 2022.

[177] Pei Liu, Bo Fu, Feng Ye, Rui Yang, Bin Xu, and Luping Ji. Revisiting whole-slide image pyramids for cancer prognosis via dual-stream networks. *arXiv preprint arXiv:2206.05782*, 2022.

[178] Yuan Liu, Songyang Zhang, Jiacheng Chen, Kai Chen, and Dahua Lin. Pixmim: Rethinking pixel reconstruction in masked image modeling. *ArXiv*, abs/2303.02416, 2023.

[179] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021.

[180] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, pages 11976–11986, 2022.

[181] Romain Lopez, Jeffrey Regier, Michael Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15:1053 – 1058, 2018.

[182] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[183] Malte Lücken, Daniel Burkhardt, Robrecht Cannoodt, Christopher Lance, Aditi Agrawal, Hananeh Aliee, Ann Chen, Louise Deconinck, Angela Detweiler, Alejandro Granados, Shelly Huynh, Laura Isacco, Yang Kim, Dominik Klein, Bony de Kumar, Sunil Kuppasani, Heiko Lickert, Aaron McGeever, Joaquin Melgarejo, Honey Mekonen, Maurizio Morri, Michaela Müller, Norma Neff, Sheryl Paul, Bastian Rieck, Kaylie Schneider, Scott Steelman, Michael Sterr, Daniel Treacy, Alexander Tong, Alexandra-Chloé Villani, Guilin Wang, Jia Yan, Ce Zhang, Angela Oliveira Pisco, Smita Krishnaswamy, Fabian J. Theis, and Jonathan M. Bloom. A sandbox for prediction and integration of dna, rna, and proteins in single cells. In *NeurIPS Datasets and Benchmarks*, 2021.

[184] Malte D Luecken, Daniel Bernard Burkhardt, Robrecht Cannoodt, Christopher Lance, Aditi Agrawal, Hananeh Aliee, Ann T Chen, Louise Deconinck, Angela M Detweiler, Alejandro A Granados, et al. A sandbox for prediction and integration of dna, rna, and proteins in single cells. In *NIPS Datasets and Benchmarks Track*, 2021.

[185] Xiangde Luo, Jieneng Chen, Tao Song, Yinan Chen, Guotai Wang, and Shaoting Zhang. Semi-supervised medical image segmentation through dual-task consistency. In *AAAI Conference on Artificial Intelligence*, 2020.

[186] Xiangde Luo, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Shichuan Zhang, Nianyong Chen, Guotai Wang, and Shaoting Zhang. Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via

uncertainty rectified pyramid consistency. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2020.

[187] Xiangde Luo, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Shichuan Zhang, Nianyong Chen, Guotai Wang, and Shaoting Zhang. Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 318–329. Springer, 2021.

[188] Jun Ma, Yixin Wang, Xingle An, Cheng Ge, Ziqi Yu, Jianan Chen, Qiongjie Zhu, Guoqiang Dong, Jian He, Zhiqiang He, et al. Toward data-efficient learning: A benchmark for covid-19 ct lung and infection segmentation. *Medical physics*, 48(3):1197–1210, 2021.

[189] Oskar Maier, Bjoern H Menze, Janina Von der Gablentz, Levin Häni, Mattias P Heinrich, Matthias Liebrand, Stefan Winzeck, Abdul Basit, Paul Bentley, Liang Chen, et al. Isles 2015-a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral mri. *Medical image analysis*, 35:250–269, 2017.

[190] Maciej A Mazurowski, Mateusz Buda, Ashirbani Saha, and Mustafa R Bashir. Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on mri. *Journal of magnetic resonance imaging*, 49(4):939–954, 2019.

[191] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

[192] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. *arXiv preprint arXiv:1710.03740*, 2017.

[193] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

[194] Kodai Minoura, Ko Abe, Hyunha Nam, Hiroyoshi Nishikawa, and Teppei Shimamura. A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data. *Cell Reports Methods*, 1(5):100071, 2021.

[195] Takeru Miyato, Shin ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:1979–1993, 2017.

[196] Nooshin Mojab, Vahid Noroozi, Darvin Yi, Manoj P Nallabothula, Abdullah Aleem, S Yu Philip, and Joelle A Hallak. Real-world multi-domain data applications for generalizations to clinical settings. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 677–684. IEEE, 2020.

[197] Matthew P Mulè, Andrew J Martins, and John S Tsang. Normalizing and denoising protein expression data from droplet-based single cell profiling. *Nature Communications*, 13(1):2099, 2022.

[198] Somjit Nath, Rushiv Arora, and Samira Ebrahimi Kahou. Locally constrained representations in reinforcement learning. *arXiv preprint arXiv: 2209.09441*, 2022.

[199] Kee Yuan Ngiam and Wei Khor. Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology*, 20(5):e262–e273, 2019.

[200] Nhut-Quang Nguyen and Thanh-Sach Le. A semi-supervised learning method to remedy the lack of labeled data. In *2021 15th International Conference on Advanced Computing and Applications (ACOMP)*, pages 78–84. IEEE, 2021.

[201] Muhammad Khalid Khan Niazi, Anil V Parwani, and Metin N Gurcan. Digital pathology and artificial intelligence. *The Lancet Oncology*, 20(5):e253–e261, 2019.

[202] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv: 1804.03999*, 2018.

[203] Liudmila Ostroumova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. In *NIPS*, 2017.

[204] Jiahong Ouyang, Qingyu Zhao, Ehsan Adeli, Edith V Sullivan, Adolf Pfefferbaum, Greg Zaharchuk, and Kilian M Pohl. Self-supervised longitudinal neighbourhood embedding. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 80–89. Springer, 2021.

[205] Zixuan Pan, Jianxu Chen, and Yiyu Shi. Masked diffusion as self-supervised representation learner. *arXiv preprint arXiv:2308.05695*, 2023.

[206] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. A unified view of masked image modeling. *ArXiv*, abs/2210.10615, 2022.

[207] Shehan Perera, Pouyan Navard, and Alper Yilmaz. Segformer3d: an efficient transformer for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4981–4988, 2024.

[208] Alexey L. Pomerantsev. Principal component analysis (pca). *Encyclopedia of Autism Spectrum Disorders*, 2014.

[209] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[210] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[211] Md Mostafijur Rahman and Radu Marculescu. Medical image segmentation via cascaded attention decoding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6222–6231, 2023.

[212] A. Ramesh, Mikhail Pavlov, Gabriel Goh, S. Gray, Chelsea Voss, Alec Radford, Mark Chen, and I. Sutskever. Zero-shot text-to-image generation. *International Conference on Machine Learning*, 2021.

[213] Roshan Rao, Jason Liu, Robert Verkuil, Joshua Meier, John F. Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. *bioRxiv*, 2021.

[214] Roshan M Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer protein language models are unsupervised structure learners. *bioRxiv*, 2020.

[215] Colorado J Reed, Xiangyu Yue, Ani Nrusimha, Sayna Ebrahimi, Vivek Vijaykumar, Richard Mao, Bo Li, Shanghang Zhang, Devin Guillory, Sean Metzger, et al. Self-supervised pretraining improves self-supervised pretraining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2584–2594, 2022.

[216] G Anthony Reina, Ravi Panchumarthy, Siddhesh Pravin Thakur, Alexei Bastidas, and Spyridon Bakas. Systematic evaluation of image tiling adverse effects on deep learning semantic segmentation. *Frontiers in neuroscience*, 14:65, 2020.

[217] Mina Rezaei, Haojin Yang, and Christoph Meinel. Deep neural network with l2-norm unit for brain lesions detection. In *Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14–18, 2017, Proceedings, Part IV 24*, pages 798–807. Springer, 2017.

[218] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *ArXiv*, abs/2010.04592, 2020.

[219] Reza Sadeghi, Tanvi Banerjee, and William Romine. Early hospital mortality prediction using vital signals. *Smart Health*, 9:265–274, 2018.

[220] Rabeya Sadia, Jin Chen, and Jie Zhang. Ct image denoising methods for image quality improvement and radiation dose reduction. *Journal of Applied Clinical Medical Physics*, 25, 2024.

[221] Emanuele Sansone. Leveraging hidden structure in self-supervised learning. *ArXiv*, abs/2106.16060, 2021.

[222] Jonathan Sauder and Bjarne Sievers. Self-supervised deep learning on point clouds by reconstructing space. In *Neural Information Processing Systems*, 2019.

[223] Zachary Schwehr and Sriman Achanta. Brain tumor segmentation based on deep learning, attention mechanisms, and energy-based uncertainty prediction. *arXiv preprint arXiv:2401.00587*, 2023.

[224] Constantin Marc Seibold, Simon Reiß, Jens Kleesiek, and Rainer Stiefelhagen. Reference-guided pseudo-label generation for medical semantic segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2171–2179, 2022.

[225] Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In *BIOCOMPUTING 2021: proceedings of the Pacific symposium*, pages 232–243. World Scientific, 2020.

[226] Jiangbo Shi, Tieliang Gong, Chunbao Wang, and Chen Li. Semi-supervised pixel contrastive learning framework for tissue segmentation in histopathological image. *JBHI*, 27(1):97–108, 2022.

[227] Yinghuan Shi, Jian Zhang, Tong Ling, Jiwen Lu, Yefeng Zheng, Qian Yu, Lei Qi, and Yang Gao. Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation. *IEEE transactions on medical imaging*, 41(3):608–620, 2021.

[228] Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American journal of roentgenology*, 174(1):71–74, 2000.

[229] Oriane Siméoni, Gilles Puy, Huy V Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. In *British Machine Vision Conference*, 2021.

[230] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.

[231] Pranav Kumar Singh, Raviteja Chukkapalli, Shravan Chaudhari, Luoyao Chen, Mei Chen, Jinqian Pan, Craig Smuda, and Jacopo Cirrone. Shifting to machine supervision: annotation-efficient semi and self-supervised learning for automatic medical image segmentation and classification. *Scientific Reports*, 14, 2023.

[232] Korsuk Sirinukunwattana, Josien PW Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang, Bogdan J Matuszewski, Elia Bruni, Urko Sanchez, et al. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, 35:489–502, 2017.

[233] David RJ Snead, Yee-Wah Tsang, Aisha Meskiri, Peter K Kimani, Richard Crossman, Nasir M Rajpoot, Elaine Blessing, Klaus Chen, Kishore Gopalakrishnan, Paul Matthews, et al. Validation of digital pathology imaging for primary histopathological diagnosis. *Histopathology*, 68(7):1063–1072, 2016.

[234] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin Dogus Cubuk, Alexey Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *ArXiv*, abs/2001.07685, 2020.

[235] Kechen Song, Liming Huang, Aojun Gong, and Yunhui Yan. Multiple graph affinity interactive network and a variable illumination dataset for rgbt image salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 33:3104–3118, 2023.

[236] Geoffrey A Sonn, Shyam Natarajan, Daniel JA Margolis, Malu MacAiran, Patricia Lieu, Jiaoti Huang, Frederick J Dorey, and Leonard S Marks. Targeted biopsy in the detection of prostate cancer using an office based magnetic resonance ultrasound fusion device. *The Journal of urology*, 189(1):86–92, 2013.

[237] Hari Sowrirajan, Jingbo Yang, Andrew Y Ng, and Pranav Rajpurkar. Moco pretraining improves representation and transferability of chest x-ray models. In *Medical Imaging with Deep Learning*, pages 728–744. PMLR, 2021.

[238] Genevieve L Stein-O'Brien, Raman Arora, Aedin C Culhane, Alexander V Favorov, Lana X Garmire, Casey S Greene, Loyal A Goff, Yifeng Li, Aloune Ngom, Michael F Ochs, et al. Enter the matrix: factorization uncovers knowledge from omics. *Trends in Genetics*, 34(10):790–805, 2018.

[239] Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9):865–868, 2017.

[240] Tim Stuart and Rahul Satija. Integrative single-cell analysis. *Nature Reviews Genetics*, 20:257–272, 2019.

[241] Yanming Sun and Chunyan Wang. Brain tumor detection based on a novel and high-quality prediction of the tumor pixel distributions. *Computers in Biology and Medicine*, 172:108196, 2024.

[242] Nima Tajbakhsh, Holger R. Roth, Demetri Terzopoulos, and Jianming Liang. Guest editorial annotation-efficient deep learning: The holy grail of medical imaging. *IEEE transactions on medical imaging*, 40 10:2526–2533, 2021.

[243] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[244] Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12180–12189, 2020.

[245] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290 5500:2319–23, 2000.

[246] Bethany H Thompson, Gaetano Di Caterina, and Jeremy P Voisey. Pseudo-label refinement using superpixels for semi-supervised brain tumour segmentation. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2022.

[247] Songsong Tian, Lusi Li, Weijun Li, Hang Ran, Xin Ning, and Prayag Tiwari. A survey on few-shot class-incremental learning. *arXiv preprint arXiv: 2304.08130*, 2023.

[248] Yu Tian, Guansong Pang, Fengbei Liu, Yuanhong Chen, Seon Ho Shin, Johan W Verjans, Rajvinder Singh, and Gustavo Carneiro. Constrained contrastive distribution learning for unsupervised anomaly detection and localisation in medical images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 128–140. Springer, 2021.

[249] Phi Vu Tran. Semi-supervised learning with self-supervised networks. *ArXiv*, abs/1906.10343, 2019.

[250] Tuan Truong, Sadegh Mohammadi, and Matthias Lenga. How transferable are self-supervised features in medical image classification tasks? In *Machine Learning for Health*, pages 54–74. PMLR, 2021.

[251] Vullnet Useini, Stephanie Tanadini-Lang, Quentin Lohmeyer, Mirko Meboldt, Nicolaus H Andratschke, Ralph P. Braun, and Javier Barranco Garcia. Automatized self-supervised learning for skin lesion screening. *Scientific Reports*, 14, 2023.

[252] Nik Vaessen and David A. van Leeuwen. The effect of batch size on contrastive self-supervised speech representation learning. *ArXiv*, abs/2402.13723, 2024.

[253] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

[254] Angel Villar-Corrales and Sven Behnke. Unsupervised image decomposition with phase-correlation networks. *ArXiv*, abs/2110.03473, 2021.

[255] Yen Nhi Truong Vu, Richard Wang, Niranjan Balachandar, Can Liu, Andrew Y Ng, and Pranav Rajpurkar. Medaug: Contrastive learning leveraging patient metadata improves representations for chest x-ray interpretation. In *Machine Learning for Healthcare Conference*, pages 755–769. PMLR, 2021.

[256] Changlin Wan, Wennan Chang, Yu Zhang, Fenil Shah, Xiaoyu Lu, Yong Zang, Anru Zhang, Sha Cao, Melissa L Fishel, Qin Ma, et al. Ltmg: a novel statistical modeling of transcriptional expression states in single-cell rna-seq data. *Nucleic Acids Research*, 47(18):e111–e111, 2019.

[257] Guotai Wang, Xinglong Liu, Chaoping Li, Zhiyong Xu, Jiugen Ruan, Haifeng Zhu, Tao Meng, Kang Li, Ning Huang, and Shaoting Zhang. A noise-robust framework for automatic segmentation of covid-19 pneumonia lesions from ct images. *IEEE Transactions on Medical Imaging*, 39(8):2653–2663, 2020.

[258] Haoqing Wang, Yehui Tang, Yunhe Wang, Jianyuan Guo, Zhiwei Deng, and Kai Han. Masked image modeling with local multi-scale reconstruction. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2122–2131, 2023.

[259] Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15878–15887, 2023.

[260] Juexin Wang, Anjun Ma, Yuzhou Chang, Jianting Gong, Yuexu Jiang, Ren Qi, Cankun Wang, Hongjun Fu, Qin Ma, and Dong Xu. scGNN is a novel graph neural network framework for single-cell rna-seq analyses. *Nature Communications*, 12(1):1882, 2021.

[261] Kaiping Wang, Bo Zhan, Chen Zu, Xi Wu, Jiliu Zhou, Luping Zhou, and Yan Wang. Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning. *Medical Image Analysis*, 79:102447, 2022.

[262] Renzhen Wang, Yichen Wu, Huai Chen, Lisheng Wang, and Deyu Meng. Neighbor matching for semi-supervised learning. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 439–449. Springer, 2021.

[263] Shanshan Wang, Cheng Li, Rongpin Wang, Zaiyi Liu, Meiyun Wang, Hongna Tan, Yaping Wu, Xinfeng Liu, Hui Sun, Rui Yang, Xin Liu, Jie Chen, Hui-Chong Zhou, Ismail Ben Ayed, and Hairong Zheng. Annotation-efficient deep learning for automatic medical image segmentation. *Nature Communications*, 12, 2020.

[264] Sida I. Wang and Christopher D. Manning. Fast dropout training. In *International Conference on Machine Learning*, 2013.

[265] Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv: 2006.04768*, 2020.

[266] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *International Conference on Machine Learning*, 2020.

[267] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases). In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 43462–3471, 2017.

[268] Xiaoyan Wang, Yiwen Yuan, Dongyan Guo, Xiaojie Huang, Ying Cui, Ming Xia, Zhenhua Wang, Cong Bai, and Shengyong Chen. Ssa-net: Spatial self-attention network for covid-19 pneumonia infection segmentation with semi-supervised few-shot learning. *Medical image analysis*, 79:102459, 2022.

[269] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *CVPR*, pages 3024–3033, 2021.

[270] Yixin Wang, Yao Zhang, Jiang Tian, Cheng Zhong, Zhongchao Shi, Yang Zhang, and Zhiqiang He. Double-uncertainty weighted method for semi-supervised learning. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 542–551. Springer, 2020.

[271] Yixin Wang, Yao Zhang, Jiang Tian, Cheng Zhong, Zhongchao Shi, Yang Zhang, and Zhiqiang He. Double-uncertainty weighted method for semi-supervised learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2020.

[272] Ziyang Wang, Jian-Qing Zheng, and Irina Voiculescu. An uncertainty-aware transformer for MRI cardiac semantic segmentation via mean teachers. In *Annual Conference on Medical Image Understanding and Analysis*, pages 494–507, 2022.

[273] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013.

[274] Hongzhi Wen, Jiayuan Ding, Wei Jin, Yiqi Wang, Yuying Xie, and Jiliang Tang. Graph neural networks for multimodal single-cell data integration. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4153–4163, 2022.

[275] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *CVPR*, pages 2840–2848, 2017.

[276] Kevin E Wu, Kathryn E Yost, Howard Y Chang, and James Zou. BABEL enables cross-modality translation between multiomic profiles at single-cell resolution. *Proceedings of the National Academy of Sciences*, 118(15):e2023070118, 2021.

[277] Yicheng Wu, Zongyuan Ge, Donghao Zhang, Minfeng Xu, Lei Zhang, Yong Xia, and Jianfei Cai. Mutual consistency learning for semi-supervised medical image segmentation. *Medical Image Analysis*, 81:102530, 2022.

[278] Yicheng Wu, Zhonghua Wu, Qianyi Wu, ZongYuan Ge, and Jianfei Cai. Exploring smoothness and class-separation for semi-supervised medical image segmentation. *ArXiv*, abs/2203.01324, 2022.

[279] Yicheng Wu, Minfeng Xu, Zongyuan Ge, Jianfei Cai, and Lei Zhang. Semi-supervised left atrium segmentation with mutual consistency training. *ArXiv*, abs/2103.02911, 2021.

[280] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, pages 3733–3742, 2018.

[281] Zongze Wu, Ming Yin, Yajing Zhou, Xiaozhao Fang, and Shengli Xie. Robust spectral subspace clustering based on least square regression. *Neural Processing Letters*, 48:1359 – 1372, 2017.

[282] Tete Xiao, Xiaolong Wang, Alexei A. Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. *ArXiv*, abs/2008.05659, 2020.

[283] Zhenda Xie, Zigang Geng, Jingcheng Hu, Zheng Zhang, Han Hu, and Yue Cao. Revealing the dark secrets of masked image modeling. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14475–14485, 2022.

[284] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: a simple framework for masked image modeling. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9643–9653, 2021.

[285] Zhaohan Xiong, Qing Xia, Zhiqiang Hu, Ning Huang, Cheng Bian, Yefeng Zheng, Sulaiman Vesal, Nishant Ravikumar, Andreas Maier, Xin Yang, et al. A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. *Medical Image Analysis*, 67:101832, 2021.

[286] Xuanang Xu, Thomas Sanford, Baris Turkbey, Sheng Xu, Bradford J Wood, and Pingkun Yan. Shadow-consistent semi-supervised learning for prostate ultrasound segmentation. *IEEE Transactions on Medical Imaging*, 41(6):1331–1345, 2021.

[287] Zhe Xu, Yixin Wang, Donghuan Lu, Lequan Yu, Jiangpeng Yan, Jie Luo, Kai Ma, Yefeng Zheng, and Raymond Kai-yu Tong. All-around real label supervision: Cyclic prototype consistency learning for semi-supervised medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 26(7):3174–3184, 2022.

[288] Hongwei Xue, Peng Gao, Hongyang Li, Yu Jiao Qiao, Hao Sun, Houqiang Li, and Jiebo Luo. Stare at what you see: Masked image modeling without reconstruction. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22732–22741, 2022.

[289] Yihao Xue, Siddharth Joshi, Eric Gan, Pin-Yu Chen, and Baharan Mirzasoleiman. Which features are learnt by contrastive learning? on the role of simplicity bias in class collapse and feature suppression. In *International Conference on Machine Learning*, 2023.

[290] Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence*, 4(10):852–866, 2022.

[291] Gene-Ping Yang, Yue Gu, Sashank Macha, Qingming Tang, and Yuzong Liu. On-device constrained self-supervised learning for keyword spotting via quantization aware pre-training and fine-tuning. *ICASSP*, 2024.

[292] H Mehta Yang, T Duan, D Ding, A Bagul, C Langlotz, K Shpanskaya, et al. Chexnet: radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[293] Karren Dai Yang, Anastasiya Belyaeva, Saradha Venkatachalapathy, Karthik Damodaran, Abigail Katcoff, Adityanarayanan Radhakrishnan, GV Shivashankar, and Caroline Uhler. Multi-domain translation between single-cell imaging and sequencing data using autoencoders. *Nature Communications*, 12(1):31, 2021.

[294] Pengshuai Yang, Zhiwei Hong, Xiaoxu Yin, Chengzhan Zhu, and Rui Jiang. Self-supervised visual representation learning for histopathological images. In *MICCAI*, pages 47–57. Springer, 2021.

[295] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 35:8934–8954, 2021.

[296] Yu Yao, Mingming Gong, Yuxuan Du, Jun Yu, Bo Han, Kun Zhang, and Tongliang Liu. Which is better for learning with noisy labels: The semi-supervised method or modeling label noise? In *International Conference on Machine Learning*, 2023.

[297] Zizhen Yao, Hanqing Liu, Fangming Xie, Stephan Fischer, Ricky S Adkins, Andrew I Aldridge, Seth A Ament, Anna Bartlett, M Margarita Behrens, Koen Van den Berge, et al. A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. *Nature*, 598(7879):103–110, 2021.

[298] Mehmet Can Yavuz and Berrin A. Yanikoglu. Vcl-pl:semi-supervised learning from noisy web data with variational contrastive learning. *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 740–747, 2022.

[299] Michael Yeung, Evis Sala, Carola-Bibiane Schönlieb, and Leonardo Rundo. Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Computerized Medical Imaging and Graphics*, 95, 2021.

[300] Ming Yin, Shengli Xie, Zongze Wu, Yun Zhang, and Junbin Gao. Subspace clustering via learning an adaptive low-rank graph. *IEEE Transactions on Image Processing*, 27:3716–3728, 2018.

[301] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

[302] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In *Proceedings of MICCAI*, pages 605–613, 2019.

[303] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*, pages 12310–12320. PMLR, 2021.

[304] Ramy A Zeineldin, Mohamed E Karar, Oliver Burgert, and Franziska Mathis-Ullrich. Multimodal cnn networks for brain tumor segmentation in mri: a brats 2022 challenge solution. In *International MICCAI Brainlesion Workshop*, pages 127–137. Springer, 2022.

[305] Hongrun et al. Zhang. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *CVPR*, pages 18802–18812, 2022.

[306] Ning Zhang, Yu Cao, Benyuan Liu, and Yan Luo. Improved multimodal representation learning with skip connections. *Proceedings of the 25th ACM international conference on Multimedia*, 2017.

[307] Tao Zhang, Tianqing Zhu, Jing Li, Mengde Han, Wanlei Zhou, and Philip S. Yu. Fairness in semi-supervised learning: Unlabeled data help to reduce discrimination. *IEEE Transactions on Knowledge and Data Engineering*, 34:1763–1774, 2020.

[308] Yifan Zhang, Jingqin Yang, Zhiquan Tan, and Yang Yuan. Relationmatch: Matching in-batch relationships for semi-supervised learning. *arXiv preprint arXiv: 2305.10397*, 2023.

[309] Zhenxi Zhang, Chunna Tian, Harrison X Bai, Zhicheng Jiao, and Xilan Tian. Discriminative error prediction network for semi-supervised colon gland segmentation. *Medical image analysis*, 79:102458, 2022.

[310] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[311] Xi Zhao and ShuiSheng Zhou. Fast mixing of hard negative samples for contrastive learning and use for covid-19. In *Proceedings of the 4th International Conference on Big Data Technologies*, pages 6–12, 2021.

[312] Evgenii Zheltonozhskii, Chaim Baskin, Avi Mendelson, Alexander M. Bronstein, and Or Litany. Contrast to divide: Self-supervised pre-training for learning with noisy labels. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 387–397, 2020.

[313] Hongwei Zheng, Linyuan Zhou, Han Li, Jinming Su, Xiaoming Wei, and Xiaoming Xu. Bem: Balanced and entropy-based mix for long-tailed semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22893–22903, 2024.

[314] Hong-Yu Zhou, Shuang Yu, Cheng Bian, Yifan Hu, Kai Ma, and Yefeng Zheng. Comparing to learn: Surpassing imagenet pretraining on radiographs by comparing image representations. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 398–407. Springer, 2020.

[315] Xiaojin Zhu and Andrew B. Goldberg. Introduction to semi-supervised learning. In *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 2009.

[316] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2020.

[317] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *ICCV*, pages 6002–6012, 2019.