STARLET: A Framework for Aggregating Complex Medical Image Annotations

Yonghao Si
Sun Yat-sen University
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, Guangdong Province
siyh3@mail2.sysu.edu.cn

Zhao Chen
The Hong Kong University of Science and Technology
Kowloon, Hong Kong
chenzhao@ust.hk

Libin Zheng*
Sun Yat-sen University
Guangzhou, Guangdong Province
zhenglb6@mail.sysu.edu.cn

Caleb Chen Cao
The Hong Kong University of Science and Technology
Kowloon, Hong Kong
cao@ust.hk

Lei Chen
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, Guangdong Province
leichen@hkust-gz.edu.cn

Jian Yin
Sun Yat-sen University
Guangzhou, Guangdong Province
issjyin@mail.sysu.edu.cn

ABSTRACT

Crowdsourcing has become a pivotal method for gathering annotations for a variety of data types, including text, images, and videos. This paper introduces STARLET, a comprehensive framework devised to aggregate complex medical image annotations, which has not been seriously studied before. Unlike common image annotation tasks, medical images pose challenges including multi-sub-tasks, uncertain object sizes, and multi-category objects. STARLET addresses them by integrating multiple function modules, including annotation clustering, outlier removal, and semantics-enhanced label aggregation. Our experiments demonstrate that STARLET significantly improves the annotation quality, compared to both traditional computer vision and existing crowdsourcing methods.

By effectively enhancing and aggregating the crowdsourced annotations, STARLET provides a medical data refinement tool for advancing AI-based medical applications.

PVLDB Reference Format:
Yonghao Si, Zhao Chen, Libin Zheng*, Caleb Chen Cao, Lei Chen, and Jian Yin. STARLET: A Framework for Aggregating Complex Medical Image Annotations. PVLDB, 14(1): XXX-XXX, 2020.
doi:XX.XX/XXX.XX

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/YHSI5358/anno-code.

1 INTRODUCTION

Over the past few years, crowdsourcing has emerged as a powerful tool for obtaining annotations for raw data across various types, including images [25], text [31], videos [7], etc. This approach offers significant opportunities to enhance the data quality for the development of downstream applications [37]. Recently, the task of medical image annotation has drawn increasing attention in the field, with the objective of enhancing the downstream medical applications like AI-based diagnosis and pathology identification. In contrast to common image annotation tasks, medical images have unique characteristics that present new challenges for the crowdsourcing community.

Typical medical images include cytology slide images, CT images, MRI images, X-ray images, PET scans, ultrasound images, endoscopic images, and so on. In terms of annotation, medical images have three key characteristics which distinguish them from common images: multi-sub-tasks, uncertain object sizes, and multi-category objects. I. Multi-sub-tasks. The most prominent feature of medical images lies in its high resolution and the associated rich semantic information. Thus, a single medical image may contain many target objects of interest inside. Each object indicates a sub-task for annotation, such as those framed with rectangles in Figure 2 and 1. II. Uncertain object size. Objects can have different sizes, which is even true for objects of the same type in a single image. III. Multi-category objects. There can be multiple object types of interest in a single image. For example, Figure 2 shows a cytology slide image which contains 13 rectangular-framed target objects, showing different sizes and labeled as Endocervical Cell, Metaplastic Cell, and Other Cell, respectively.

Acknowledging these three issues, the difficulty of medical image annotation lies in simultaneously determining the number of objects, their areas and categories without prior knowledge. In this paper, the terms “category" and “label" are used interchangeably.

The aforementioned task is analogous to the problems of image segmentation and object detection in the CV (computer vision) field, which have already a set of mature public released models available, such as YOLOv10 for object detection [40], Retinanet for object detection [24], and Segment Anything Model (SAM) for segmentation [20]. However, the aforementioned models are generally trained with common images from the fields like natural scenes, everyday objects, and wildlife. However, the models mentioned above are generally trained on common images from fields such as natural scenes, everyday objects, and wildlife. While it is possible to use these pre-trained models and fine-tune them for specific

Proceedings of the VLDB Endowment, Vol. 14, No. 1 ISSN 2150-8097.
doi:XX.XX/XXX.XX

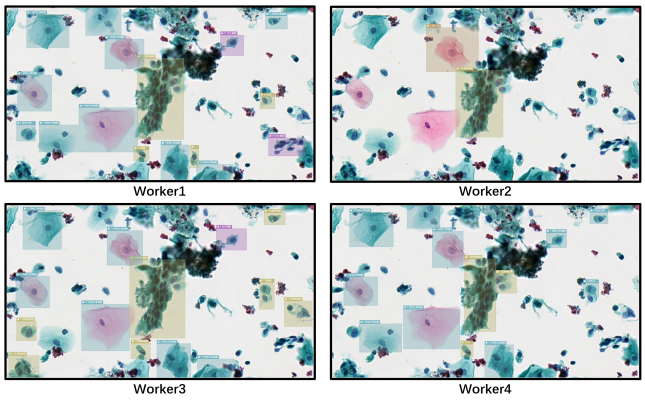


Figure 1: An illustration showcasing the complexity of annotating medical images. Different colored ROIs represent different categories.

Method	Key Characteristics		
	Issue I	Issue II	Issue III
Lin et al. [16]		✓	✓
Ren et al. [34]		✓	✓
Le et al. [21]			✓
Keshavan and Yan [19]			✓
Heim [18]	✓	✓	
STARLET (Ours)	✓	✓	✓

Table 1: Comparison of existing methods in addressing key issues in medical image annotation.

medical annotation tasks, a significant challenge is the lack of well-annotated medical image datasets in the community. Additionally, each new type of medical image (e.g., from cytology to CT images) necessitates a new dataset, leading to ongoing costs.

Recognizing this challenge, the majority of existing works focus on crowd-assisted medical image annotation [16, 18, 19, 21, 34]. However, none of them fully considers the aforementioned three issues (see Table 1). Lin et al. [16] and Ren et al. [34], have underscored the benefits of decomposing complex medical annotation tasks into more manageable sub-tasks. This decomposition strategy has been shown to alleviate cognitive load and enhance annotation efficiency. However, they can only partially address the issues II&III, as they fail to consider the case of having multiple objects in a single image, but instead only decomposing the multilevel annotations of a single object. Le et al. [21], Keshavan and Yan [19] focused on classification tasks in brain and radiological images, respectively, which are simpler annotation types and fail to consider issues I and II. Heim [18] addressed liver segmentation tasks in abdominal CT scans, partially addressing issues I and II without involving multiple categories. We summarize the aforementioned workers on their set-up of the three issues in Table 1. None of them sufficiently consider the three issues.

The difficulty of resolving the aforementioned three issues lie in the complex annotation space. In contrast to classical classification tasks which only requires a resolved label for each input object, for medical image annotation, we need to determine not only the labels, but also the "objects" themselves (i.e., their location area in an image). Figure 1 shows the annotation results from four pathologists for each cytology image, where each annotation is a rectangle (box) with a label. It can be found that in addition to the divergence among workers' annotated labels for the drawn boxes, there is also a significant deviation on their drawn location area for the objects. Indeed, it is not uncommon that a worker only discovers part of the objects per image. Different workers can draw different numbers of boxes for the same image. Then, when there is significant overlap between two bounding boxes from different workers, careful resolution is required. The overlap may indicate either that the two boxes represent a single object (i.e., the workers disagree on its location) or that they represent two distinct objects that slightly overlap in the image.

In response to these challenges, we propose a novel crowdsourcing framework, particularly tailored for medical image annotation. We first propose the Annotation Clustering Module (ACM), which incorporates a worker intent recognition algorithm to effectively determine the number of annotation clusters per image. Regarding the cluster number, a crowdsourcing-oriented spectral clustering algorithm is devised to identify and refine the clusters. Each annotation cluster is a set of annotations (box & label) from different workers pointing the same ground truth object. Based on the clustering results, an outlier removal process is further conducted to identify and eliminate outliers, as workers are also likely to give wrong annotations.

Upon the completion of the clustering phase, we introduce the Annotation Aggregation Module (AAM) to aggregate the annotations inside each cluster, including both the eventual object location and the category. AAM jointly considers the inherent image feature and the external crowdsourcing profile, i.e., the worker quality, as well as the annotation themselves.

We evaluate our framework using two distinct datasets, the cytology medical image annotation dataset inside the medical domain and a general dataset outside the domain. The experiments demonstrate that STARLET exhibits a significantly superior aggregation effect compared to the baseline methods. In terms of box aggregation, our method outperforms various baseline methods by 21.83% to 40.89% in terms of the One2one F1-score (Section 6.2). Regarding category aggregation, our method achieves a F1-score improvement ranging from 1.06% to 2.38%.

To summarize, we present several key contributions towards the complex annotations of medical images, summarized as follows.

(1) **Framework for Aggregating Complex Annotations:** we introduce a novel framework designed specifically to aggregate complex crowdsourcing annotations towards medical images. This framework addresses the challenge of reconciling diverse answers from crowd workers, in order to produce coherent and accurate annotations (Section 3).
(2) **Crowdsourcing-oriented Spectral Clustering Method:** we propose a novel crowdsourcing-oriented spectral clustering method, integrating the crowdsourcing profile with the image's inherent feature to determine the object locations. This method partitions the annotations provided by workers, remove the outliers and refine the clusters iteratively. By leveraging both sources of information, i.e., both the annotations and the inherent image feature, it effectively clusters the annotations pointing the same object.

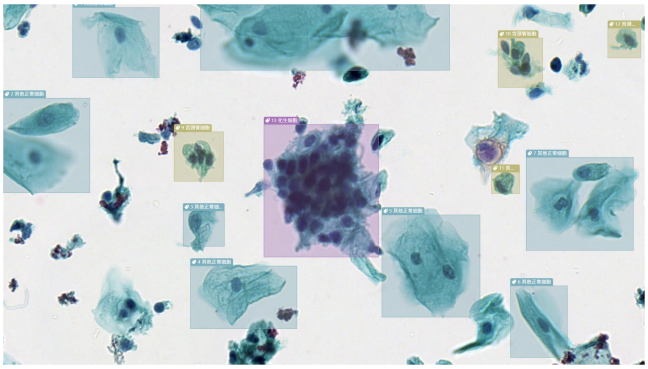


Figure 2: An example of cytological labeling results.

(3) **Crowd-machine Collaborative Inference for Medical Objects:** we propose a crowd-machine collaborative inference method for medical targets. By comparing the boxes provided by human workers with the outputs of a machine-learning model, we estimate the quality of workers' annotations for each cluster. The annotation quality is in turn leveraged during the inference process, to obtain the eventual location & label for each cluster/object (Section 5).
(4) **Release of Crowdsourced Datasets:** we release two crowdsourced medical annotation datasets to the community: one for cytology image annotation with 446 cytology images and 14,579 annotations, and the other for COCO image annotation [27] with 57 images and 1,930 annotations (Section 6.1).

2 PROBLEM DEFINITIONS

Notation	Description
T	Set of tasks $\{t_1, t_2, \ldots, t_N\}$
A_i^j	Annotations by worker w_j for task t_i
$\hat{A}_i$	Aggregated annotations for the i^{th} task
a_m	m^{th} annotation in image A_i, $a_m = \langle c_m, r_m \rangle$
r_m	m^{th} ROI in the i^{th} task, $r_m = \langle x_1, y_1, x_2, y_2 \rangle$
r_m^{sam}	Output after inputting r_m into SAM as a prompt
$\sigma(\cdot)$	Estimated quality of annotation, e.g., $\sigma(a_m)$ denotes the estimated quality of a_m annotation
$w(\cdot)$	The worker the annotation belongs to, e.g., $w(a_m)$ is the worker a_m belongs to
$\eta(\cdot)$	Estimated quality of worker, e.g., $\eta(w(a_m))$ is the estimated quality of the worker $w(a_m)$

Table 2: Important notations and definitions.

In this section, we define the preliminary concepts for the medical image annotation problem. Annotating a medical image involves drawing the location area and labels for the objects inside. For example, a task may be annotating the malignant cells in a cytology image, identifying and labeling tumors in a CT image, marking abnormalities in an MRI image, and so on.

Definition 1 (Tasks). *A set of tasks are represented as $T = \{t_1, t_2, \ldots, t_N\}$, where t_i signifies annotating the i^{th} image.*

Each task can encompass different numbers of objects inside. In a single image, the objects can also have different sizes. Thus, we need to define the bounding boxes to locate such objects.

Definition 2 (Rectangles of Interest (ROI)). *The objects in the images are located via a set of rectangles, $R = \{R_1, \ldots, R_N\}$, where $R_i = \{r_{i,1}, r_{i,2}, \ldots, r_{i,L_i}\}$ represents the ROIs for the i^{th} task (image). Each $r_{i_l} = \langle x_1, y_1, x_2, y_2 \rangle$ denotes a specific object or feature within task t_i. L_i is the number of objects of interest in the i^{th} image and can vary.*

For instance, in Figure 2, each bounding box corresponds to a ROI, delineating the cells of interest inside the image.

Definition 3 (Categories). *Each ROI is associated with a specific category c from a pre-defined category set $C = \{c_1, c_2, \ldots, c_Q\}$. For example, each ROI in Figure 2 represents a specific cell type, which is the category of the ROI in this application.*

Definition 4 (Ground Truth). *$G = \{G_1, \ldots, G_N\}$, where $G_i = \{g_{i,1}, g_{i,2}, \ldots, g_{i,L_i'}\}$ represents the ground truth annotations for the i^{th} task. Each $g_{i,l} = \langle c_{i,l}, r_{i,l} \rangle$ denotes the ROI for a certain object and its category. L_i' represents the number of ground truth annotations in the i^{th} task.*

Definition 5 (Annotation). *$A = \{A_i^j\}$ signifies the collection of annotations from workers, where $A_i^j = \{a_{i,l}^j\}$ is the set of annotations by worker w_j for task t_i, with $a_{i,l}^j = \langle c_{i,l}^j, r_{i,l}^j \rangle$ being the l^{th} annotation. Analogously $c_{i,l}^j$ represents the category and $r_{i,l}^j$ represents the ROI of its targeted object. $A_i = \{a_{i,l}\}$ represents the collection of all annotations received for task t_i, where $a_{i,l}$ represents one of its annotations (without j, i.e., not specifying the worker).*

We eventually define the problem of aggregating crowdsourced medical image annotations as below.

Definition 6 (Aggregation of Complex Medical Image Annotations). *Given a set of annotations A_i for medical images $i = \{1, 2, \ldots, N\}$, our objective is to produce a consolidated annotation set $\hat{A} = \{\hat{A}_1, \hat{A}_2, \ldots, \hat{A}_N\}$, where $\hat{A}_i = \{\hat{a}_{i,1}, \hat{a}_{i,2}, \ldots, \hat{a}_{i,\hat{L}_i}\}$, such that for each i, $\hat{A}_i$ is as close to G_i as possible. $\hat{L}_i$ represents the number of annotations in the aggregation result on the i^{th} task.*

To evaluate how good $\hat{A}_i$ is, for each $\hat{a}_{i,l} = \{c_{i,l}, r_{i,l}\}$, we need to consider both the correctness of label $c_{i,l}$ and the quality of the area $r_{i,l}$. The evaluation metric would be formally formulated in the experiment Section 6.2, to investigate the performances of our method and existing ones.

For the rest of this paper, whenever the context is clear, we omit the subscript of 'i', since the discussion always focuses on the annotations/ROIs in a single image. For example, we use a_m and a_n to directly denote two annotations in A_i.

We summarize the important notations in Table 2.

3 OVERVIEW OF FRAMEWORK

In this section, we introduce the proposed framework outlined in Figure 3, named STARLET, i.e., cluSTering And aggRegating medicaL imagE annoTations. For a medical image, workers can draw different numbers of ROIs with heterogeneous locations, sizes and

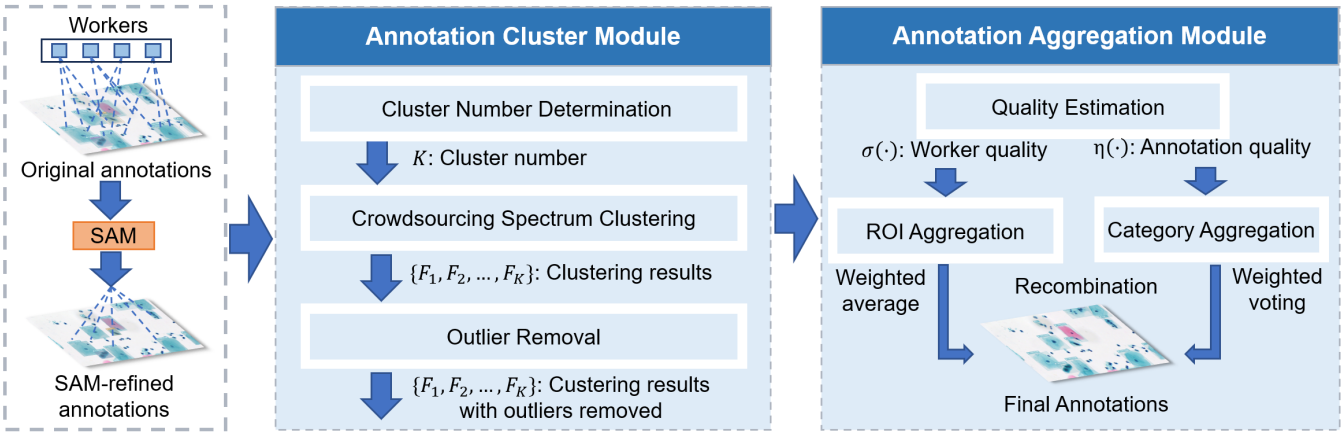


Figure 3: Framework of STARLET.

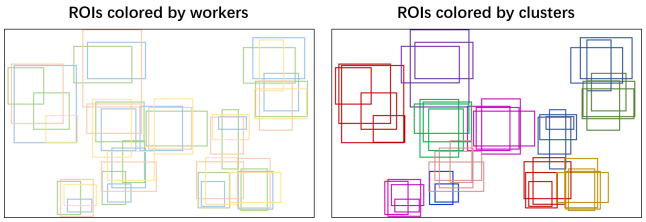


Figure 4: The left depicts the raw ROIs from workers (a color for each worker), while the right shows the expected clusters (a color for each object).

categories, which brings the difficulty for aggregation. Then, STARLET is comprised of two main modules: the Annotation Clustering Module (ACM) and the Annotation Aggregation Module (AAM). The ACM module first clusters the ROIs from all workers to roughly determine the number of objects, with one cluster indicating one object. Then, for each ROI cluster, AAM reaches a consensus for the eventual location area and category for the object.

Annotation Clustering Module: Workers can provide variable numbers of annotations for a single image, depending on their understanding of the image and objects. This means that there is not necessarily a one-to-one correspondence among workers' annotations. For example, a worker plots two boxes in top-left and top-right respectively, while the other plots two in top-right and bottom-right. This means that there exist simultaneously overlap and difference between their found objects. In addition, even for the found same object (e.g. the top-right one), their annotated boxes can vary. This necessitates clustering workers' annotated ROIs so that we could first determine the number of found objects. As illustrated in Figure 4, this clustering process is expected to identify ROIs for the same object. To that end, we propose a novel crowdsourcing-oriented ROI spectral clustering method, which integrates workers' behaviours into the clustering process. We further propose an outlier removal technique, which identifies and eliminates outliers caused by some workers' mis-annotation.

Annotation Aggregation Module: Once clustering is complete, the next step involves aggregating the annotations within each cluster, to derive a consensus ROI and then its category. We propose the aggregation module (AAM). It gauges the quality of workers and ROIs by amalgamating both the image context and workers' answers. The aggregation is then conducted by synthesizing these insights.

Algorithm 1: Framework

Input : A
Output: $\hat{A}$

1 Pre-processing: $A^{sam} \leftarrow$ use SAM to refine A;
2 **for** $img \leftarrow all\_imgs$ **do**
3 Determine the number of clusters K (Section 4.1);
4 $C_1, C_2, \ldots, C_K \leftarrow$ Cluster A based on K (Section 4.2);
5 Remove outliers for each cluster (Section 4.3);
6 Estimate worker and label quality w.r.t. A & A^{sam} (Section 5);
7 $\hat{A} \leftarrow$ ROI & category aggregation based on the estimated quality (Section 5);
8 **Return** $\hat{A}$

STARLET is delineated in Algorithm 1. It entails the following steps:

- **Line 1: Pre-processing**. Utilize the CV model, i.e., SAM [20], to refine the worker annotations, enhancing the annotations by leveraging the inherent image features (Section 4.1).
- **Lines 2-5: ROI clustering**. The ACM module further consists of three steps. It ascertains the number of clusters within the image data (Section 4.1), then conducts spectral clustering on the annotations (Section 4.1), and finally eliminate outliers within each cluster (Section 4.4). Since workers are also likely to produce wrong ROI annotations, outlier removal is necessary to enhance the stability and reliability of clusters.
- **Lines 6-7: ROI aggregation**. It applies AAM to estimate the quality of both the workers and their annotations, and then conducts the aggregation for both the object location and category (Section 5).

4 ANNOTATION CLUSTERING MODULE

In this section, we provide a detailed description of our Annotation Clustering Module. Overall, our clustering process is iterative,

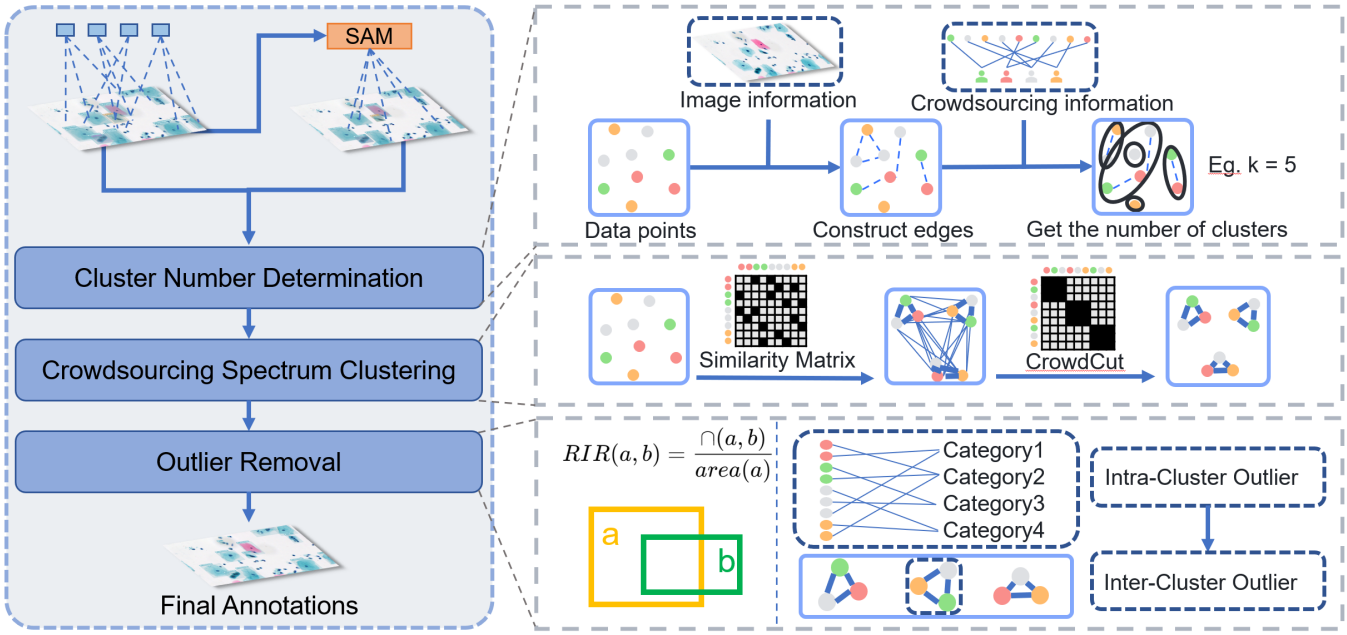


Figure 5: Clustering process.

consisting of three main steps in each iteration. Firstly, we determine the number of clusters (Section 4.1). Secondly, we perform clustering using our proposed Crowdsourcing-oriented Spectral Clustering method (Section 4.2). Thirdly, we filter and remove outliers (Section 4.4).

4.1 Cluster Number Determination Method

Preparation: the Segment Anything Model (SAM) [20] is an impressive segmentation model renowned for its robust segmentation capabilities, widely applied across various computer vision domains. In our work, as a pre-processing step we also leverage SAM to refine worker annotations. Specifically, for a single image A_i, by feeding every worker ROI r_m and the image itself as prompts into SAM, it utilizes the raw image features to produce a SAM-refined ROI r_m^{sam} and the corresponding confidence $conf_m^{sam}$, leading to A_i^{sam}.

Algorithm 2: Determine Number of Clusters

Input : A_i: Set of annotations in T_i, τ: Similarity threshold
Output: Number of clusters K

1 Initialize connection matrix $\mathcal{M}$ with zeros;
2 $K \leftarrow 0$;
3 **for** *each pair* (a_m, a_n) *in* A_i **do**
4 $\mathcal{M}_{m,n} \leftarrow \text{IoU}(a_m, a_n)$;
5 **for** *each* a_m *in* A_i **do**
6 **if** a_m *is not visited* **then**
7 $U \leftarrow \text{UniqueWorkerDFS}(a_m, \mathcal{M}, \tau)$;
8 $K \leftarrow K + 1$;
9 **Return** K

Algorithm 3: UniqueWorkerDFS

Input : *node*: Start node, $\mathcal{M}$: Affinity matrix, τ: threshold
Output: U: List of visited nodes

1 Initialize an empty list U and a stack S;
2 Add *node* to U and S respectively;
3 **while** S *is not empty* **do**
4 Pop *currentNode* from S;
5 **for** *each neighbor of currentNode in* $\mathcal{M}$ **do**
6 **if** $M_{currentNode,neighbor} > \tau$ *and neighbor* $\notin U$ **then**
7 Add *neighbor* to U, push *neighbor* onto S;
8 **for** *each pair* $(a_m, a_n) \in U$ *annotated by the same worker* **do**
9 $s_m, s_n \leftarrow$ the accumulated affinity of rows of a_m, a_n in $\mathcal{M}$;
10 Pop $a_{\min(s_m,s_n)}$ from U;
11 **Return** U

For the cluster number, either an over-fine-grained or a coarse-grained clustering would lead to misinformation for the latter aggregation steps. Algorithm 2 addresses this issue:

- **Lines 1-2: Initialization**. The algorithm starts by initializing a matrix $\mathcal{M}$ with zeros and setting the cluster counter K to 0.
- **Lines 3-4: Construct Affinity Matrix**. For each pair of annotations (a_m, a_n) in the set of worker annotations A_i, the algorithm calculates an affinity score between their ROIs, via their Intersection over Union (IoU). Note that in this section, when we refer to an annotation, we mean its ROI.
- **Lines 5-8: Connected Component Analysis**. The algorithm conducts a connected component analysis on the matrix $\mathcal{M}$. Each group of connected annotations forms a

cluster, denoted as U. To do so, treating an ROI as a node and placing edges between two ROIs with affinity scores larger than τ, we then can use the vanilla depth-first search (DFS) to traverse the un-visited ROIs one-by-one. However, it may absorb two ROIs from the same worker in a single cluster. This apparently violates the worker's intention (a worker would not repeatedly draw ROIs for the same object). We then propose UniqueWorkerDFS in Algorithm 3, which prevents this over-connection. Basically, it first performs a DFS traversal by popping the current node, examining its neighbors, and adding those with affinity scores above τ to U and S (Lines 3-7 of Alg. 3). Then, for each pair of annotations from the same worker in U, it retains only the one with the higher accumulated affinity score, which ensures each cluster has at most one annotation per worker (Lines 8-10 of Alg. 3).

- **Line 9: Returning the Number of Clusters**. Finally, the algorithm returns the total number of clusters identified, providing a measure of the granularity of the clustering.

Given that the dominant operations involve $O(N^2)$ computations for both constructing the affinity matrix and performing DFS traversals, the overall time complexity of Algorithm 2 is $O(N^2)$. This matches the time complexity of the DFS method used for connected component analysis.

4.2 Crowdsourcing Spectrum Clustering Method

In this section, given the estimated cluster number, we form clusters via spectral clustering. Note that we do not directly use the cluster membership results from last section, while only referencing its cluster number. This is because 1) Algorithm 2 produces coarse-grained partitions. While we believe its cluster number, the cluster partition itself is less qualified, which would be demonstrated in our experiments. 2) To ensure the cluster quality, we propose the fine-grained spectral clustering algorithm in this section, while however needs an estimated number of clusters. Then, the previous method is efficient and adopted to attain the number solely.

We propose the crowdsourcing-oriented spectral clustering method, which is customized to leverage the crowdsourcing context information compared to classical methods like RatioCut [17] and NCut [36]. Unlike these approaches that solely minimize the lost edges of a cut, our method further integrates the worker quality and annotation quality.

In the following, we first show our novel formulated cluster objective in Section 4.2.1, which is, however, NP-hard. To find a heuristic, we first deduct its equivalent formulation (Section 4.2.2), and then its relaxation together with the heuristic highlight (Section 4.2.3). The formal algorithm is detailed in Section 4.3.

4.2.1 Context-integrated Objective Formulation. We consider all ROIs on an image as nodes of a graph, with the edge weights between any two nodes denoted as $sim(\cdot)$. For example, $sim(r_m, r_n)$ represents the edge weight between the m-th ROI and the n-th ROI in a single task, calculated as follows:

$$sim(r_m, r_n) = \exp(\text{IoU}(r_m, r_n)) \cdot \mathbb{I}(w(r_m), w(r_n)), \tag{1}$$

where $w(\cdot)$ denotes the specific worker to which the ROI belongs, and $\mathbb{I}(w(r_m), w(r_n))$ indicates whether r_m and r_n belong to the same worker (1 if true, otherwise 0).

For each image, we compute the similarity matrix S, where $S_{m,n} = sim(r_m, r_n)$, and subsequently calculate the diagonal degree matrix D, where

$$D_{m,m} = \sum_{n=1}^{|A_i|} S_{m,n}. \tag{2}$$

Additionally, we construct the matrix M as the context-integrated version of the degree matrix, which incorporates worker quality and annotation quality. M is a diagonal matrix, with the m-th element on the diagonal as

$$M_{m,m} = \sum_{a_n \in A_i} S_{m,n} \cdot (1 + e^{\prod_{l \in \{m,n\}} \sigma(a_l)\eta(w(a_l))}). \tag{3}$$

Here, $\sigma(a_l)$ represents the quality estimate of the annotation a_l, which is measured by their closeness to the ROIs given by A_i^{sam}. This quality estimate is defined by the Generalized Intersection over Union (GIoU) [35] as follows:

$$\begin{aligned}\sigma(a_l) &= \text{GIoU}(r_l, r_l^{sam}) \\ &= \text{IoU}(r_l, r_l^{sam}) - \frac{\text{area}(B) - \text{area}(r_l \cup r_l^{sam})}{\text{area}(B)},\end{aligned} \tag{4}$$

where B denotes the minimum bounding rectangle enclosing both regions, and area$(\cdot)$ represents the area of a given ROI.

Additionally, $\eta(w(a_l))$ denotes the quality estimate of the worker to which the annotation a_l belongs, which is computed as the average quality of his/her annotations.

$$\eta(w_j) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|A_i^j|} \sum_{a_m \in A_i^j} \sigma(a_m). \tag{5}$$

We then form clusters by partitioning the graph with the objective as

$$CrowdCut(F_1, F_2, \ldots, F_K) = \frac{1}{2} \sum_{k=1}^{K} \frac{cut(F_k, \overline{F}_k)}{CrowdWeight(F_k)}, \tag{6}$$

where

$$CrowdWeight(F_k) = \sum_{a_m \in F_k} M_{m,m}. \tag{7}$$

F_k denotes a cluster. $CrowdWeight$ calculates the quality-weighted strength of the ROIs in a cluster regarding M. The goal above is to minimize the ratio of the number of edges cut to the total $CrowdWeight$ of the clusters. This ensures that clusters are formed not just by proximity in the graph but also by the overall quality of the annotations they contain. The problem is NP-hard. Introducing any balancing condition as a denominator turns the min-cut problem to NP-hard; see Wagner and Wagner [39] for the details.

4.2.2 Objective Equivalence. To attain the equivalent formulation, we begin with some auxiliary definitions.

Lemma 4.1 (A property of the Laplacian Matrix [15, 30]). *For the Laplacian matrix $\mathbb{L}$, assuming its dimension is n, and the value at the i-th row and j-th column is $w_{i,j}$, it satisfies the following property:*

For any vector $f \in \mathbb{R}^n$, we have $f^T \mathbb{L} f = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2$.

The detailed proof of this lemma can be found in previous studies [15, 30].

We define the Laplacian matrix of an image as $\mathbb{L} = D - S$. In addition, for a certain clustering result, we introduce the indicator vectors $h_k \in \{h_1, h_2, \ldots, h_K\}$ for $k = 1, 2, \ldots, K$. Each vector h_k is a $|A_i|$-dimensional vector, where $|A_i|$ is the number of ROIs collected for the image to be clustered, $h_k = \left[h_{1,k}, h_{2,k}, \ldots, h_{|A_i|,k}\right]$. We define $h_{m,k}$ as follows:

$$h_{m,k} = \begin{cases} 0 & a_m \notin F_k \\ \frac{1}{\sqrt{CrowdWeight(F_k)}} & a_m \in F_k \end{cases} \tag{8}$$

which indicates the cluster membership of the ROIs. Then we set the matrix $H \in \mathbb{R}^{|A_i| \times K} = \{h_k\}$ as the composition of such vectors.

Based on the above definitions, we have the following equivalence theorem.

Theorem 1: Minimizing the cut objective in Equation (6) is equivalent to the following optimization objective:

$$\begin{aligned} &\arg \min_{F_1, F_2, \ldots, F_K} \operatorname{tr}(H^T \mathbb{L} H) \\ &\text{s.t.} \quad H^T M H = I, \quad H \text{ as defined in Eq.8 .} \end{aligned} \tag{9}$$

where tr(·) denotes the trace of a matrix.

Proof Sketch. According to the definition of H, $\mathbb{L}$ and M, together with a lemma from the existing literature, we deduct $h_k^T \mathbb{L} h_k = \frac{cut(F_k, \overline{F_k})}{CrowdWeight(F_k)}$, and then $H^T M H = I$. On top of these two properties, the objective transformation is conducted. □

Proof. **Proving $\mathbf{h_k^T \mathbb{L} h_k = \frac{cut(F_k, \overline{F_k})}{CrowdWeight(F_k)}}$.** From Lemma 4.1, we have

$$h_k^T \mathbb{L} h_k = \frac{1}{2} \sum_{m=1} \sum_{n=1} S_{m,n} (h_{m,k} - h_{n,k})^2$$

Therefore,

$$\begin{aligned} h_k^T \mathbb{L} h_k &= \frac{1}{2} \Bigg(\sum_{\substack{a_m \in F_k \\ a_n \notin F_k}} S_{m,n} \left(\frac{1}{\sqrt{CrowdWeight(F_k)}} - 0 \right)^2 \\ &\quad + \sum_{\substack{a_m \notin F_k \\ a_n \in F_k}} S_{m,n} \left(0 - \frac{1}{\sqrt{CrowdWeight(F_k)}} \right)^2 \Bigg) \\ &= \frac{1}{2} \left(\sum_{\substack{a_m \in F_k \\ a_n \notin F_k}} S_{m,n} \frac{1}{CrowdWeight(F_k)} + \sum_{\substack{a_m \notin F_k \\ a_n \in F_k}} S_{m,n} \frac{1}{CrowdWeight(F_k)} \right) \\ &= \frac{1}{2} \left(cut(F_k, \overline{F_k}) \frac{1}{CrowdWeight(F_k)} + cut(\overline{F_k}, F_k) \frac{1}{CrowdWeight(F_k)} \right) \\ &= \frac{cut(F_k, \overline{F_k})}{CrowdWeight(F_k)} \end{aligned} \tag{10}$$

Proving $\mathbf{H^T M H = I}$. To simplify the notation, let $(H^T M H)_{pq}$ denote the element in the p-th row and q-th column of the matrix $H^T M H$. Clearly, we have:

$$(H^T M H)_{p,q} = \sum_{m=1}^{|A_i|} \sum_{n=1}^{|A_i|} h_{m,p} M_{m,n} h_{n,q}$$

Because M is a diagonal matrix, the product of this formula is not 0 only when $m = n$, so in fact:

$$(H^T M H)_{p,q} = \sum_{m=1}^{|A_i|} h_{m,p} M_{m,m} h_{m,q}$$

Based on the definition of the h vectors, $h_{m,p}$ and $h_{m,q}$ are nonzero only when the ROI belongs to cluster F_p or F_q. Moreover,

$$h_{m,p} = \frac{1}{\sqrt{CrowdWeight(F_p)}}$$

$$h_{m,q} = \frac{1}{\sqrt{CrowdWeight(F_q)}}$$

Thus, the above expression can be rewritten as:

$$(H^T M H)_{p,q} = \sum_{a_m \in F_p \cap F_q} \frac{1}{\sqrt{CrowdWeight(F_p)}} M_{m,m} \frac{1}{\sqrt{CrowdWeight(F_q)}}$$

Combining this with the definition of M, we obtain:
For $p = q$,

$$\begin{aligned} (H^T M H)_{p,p} &= \sum_{a_m \in F_p} \frac{1}{\sqrt{CrowdWeight(F_p)}} M_{m,m} \frac{1}{\sqrt{CrowdWeight(F_p)}} \\ &= \sum_{a_m \in F_p} \frac{M_{m,m}}{CrowdWeight(F_p)} \end{aligned} \tag{11}$$

since $CrowdWeight(F_p) = \sum_{a_m \in F_p} M_{m,m}$, the above sum equals 1.

Therefore,

$$(H^T M H)_{p,p} = 1$$

For $p \neq q$,

$$(H^T M H)_{p,q} = 0$$

since no ROI simultaneously belongs to different clusters F_p and F_q. Thus, we have proven that $H^T M H = I$.

Proving the objective equivalence. From Equation (10) and $H^T M H = I$, we can derive

$$\begin{aligned} CrowdCut(F_1, F_2, \ldots, F_K) &= \sum_{k=1}^{K} h_k^T \mathbb{L} h_k \\ &= \sum_{k=1}^{K} (H^T \mathbb{L} H)_{k,k} \\ &= \operatorname{tr}(H^T \mathbb{L} H) \end{aligned} \tag{12}$$

where tr(·) denotes the trace of a matrix. And H satisfies $H^T M H = I$.

Therefore, our optimization objective can be formulated as:

$$\begin{aligned} &\arg \min_{F_1, F_2, \ldots, F_K} \operatorname{tr}(H^T \mathbb{L} H) \\ &\text{s.t.} \quad H^T M H = I, \quad H \text{ as defined in Eq.8} \end{aligned} \tag{13}$$

□

4.2.3 Objective Relaxation. The formulation attained in last subsection is still NP-hard. To find a heuristic for an efficient solution, we adopt a general relaxation method [38], which is to abandon the discreteness condition of h_i's (its binary values by Eq. (8)). Instead we allow h_i to take any value in $\mathbb{R}^{|A_i|}$. This leads to the relaxed optimization problem.

$$\begin{aligned} \arg \min_{H \in \mathbb{R}^{|A_i| \times K}} & \operatorname{tr}(H^T \mathbb{L} H) \\ \text{s.t.} \quad & H^T M H = I. \end{aligned} \tag{14}$$

We can further reparameterize this optimization problem into a standard matrix eigenvalue decomposition problem. Specifically, let $H = M^{-\frac{1}{2}} R$, then:

$$H^T \mathbb{L} H = R^T M^{-\frac{1}{2}} \mathbb{L} M^{-\frac{1}{2}} R,$$
$$H^T M H = R^T R = I.$$

Thus, the optimization objective becomes:

$$\begin{aligned} \arg \min_{R \in \mathbb{R}^{|A_i| \times K}} & \operatorname{tr}(R^T M^{-\frac{1}{2}} \mathbb{L} M^{-\frac{1}{2}} R) \\ \text{s.t.} \quad & R^T R = I \end{aligned} \tag{15}$$

Note that this is a standard trace minimization problem. According to the Rayleigh entropy theorem [29], the solution R of the above objective is the eigenvectors corresponding to the first K smallest eigenvalues of the matrix $\mathbb{L}$ by columns. After obtaining R through eigenvectors, H can be obtained as $H = M^{-\frac{1}{2}} R$.

Now the solution H denotes the strength of a_m's over the K clusters. Since H is no longer binary, to attain the clustering membership, a naive way is to set the cluster of a_m as $\max_k h_{m,k}$. However, this ignores the strength similarity of a_m's. Thus, we choose the commonly used method in spectral clustering [38] to decide the clusters. The matrix H with dimension $|A_i| \times K$ is treated as the feature matrix of a_m's $\in A_i$, and we perform K-means clustering on the rows of matrix H, which represent the a_m elements, according to these features.

4.3 Algorithm Description

According to previous subsections, the proposed Crowdsourcing-oriented Spectral Clustering, is detailed in Algorithm 4.

Algorithm 4: Crowdsourcing Spectral Clustering

Input : A_i: Set of annotations in T_i, K: Number of clusters
Output: Cluster assignment $F_1, F_2, \ldots, F_K$

1 Construct matrices S, D and M according to Equations (1), (2) and (3), respectively.
2 Construct the Laplacian $\mathbb{L} \leftarrow S - D$;
3 Compute the normalized Laplacian $\mathbb{L}_{sym} \leftarrow M^{-\frac{1}{2}} \mathbb{L} M^{-\frac{1}{2}}$;
4 $R \leftarrow$ eigenvectors corresponding to the minimum K eigenvalues of the $\mathbb{L}_{sym}$ matrix by columns;
5 Recover matrix $H \leftarrow M^{-\frac{1}{2}} R$;
6 $F_1, F_2, \ldots, F_K \leftarrow$ the results of clustering H by rows using the K-means method;
7 **Return** $F_1, F_2, \ldots, F_K$

- **Line 1: Initialization**. Construct the similarity matrix S, the degree matrix D and the normalized degree matrix M according to Equations (1), (2) and (3), respectively.
- **Lines 2-3: Laplacian Construction and Normalization**. The Laplacian matrix $\mathbb{L}$ is derived by subtracting the degree matrix D from the similarity matrix S. The normalized Laplacian $\mathbb{L}_{sym}$ is then computed by applying normalization using M, specifically $M^{-\frac{1}{2}} \mathbb{L} M^{-\frac{1}{2}}$.
- **Lines 4-5: Eigen Decomposition**. Eigenvectors corresponding to the smallest K eigenvalues of the normalized Laplacian matrix $\mathbb{L}_{sym}$ are extracted. These eigenvectors form the matrix R, which is used to construct the matrix H.
- **Line 6: Clustering**. The rows of matrix H are clustered using the K-means method, resulting in K clusters. These clusters are the final output of the algorithm, representing the partitioned annotation sets.
- **Line 7: Return Cluster Assignment**. Finally, the algorithm returns clustering results, i.e., $F_1, F_2, \ldots, F_K$.

The time complexity of Algorithm 4 is $O(N^3)$, which is mainly due to the steps for the eigenvectors during the clustering process. Though the complexity is cubic, here N is the number of ROIs in a figure. According to our real data, N normally would not exceed 40, which would not cause much time even with the cubic time cost.

4.4 Outlier Removal Method

The medical image annotators are also likely to make faults and give wrong ROIs, especially when the image itself is difficult or the annotator gets tired. Such wrong ROIs typically do not fit any cluster and could be detected as outliers.

The previous section produces a set of clusters as $\{F_1, F_2, \ldots, F_K\}$, where $F_k = \{a_1, a_2, \ldots, a_{|F_k|}\}$, and $\sum_{k=1}^{K} |F_k| = |A_i|$. To identify outliers, we first define a Relative Intersection Ratio (RIR) score as

$$RIR(r_m, r_n) = \frac{|r_m \cap r_n|}{area(r_m)}.$$

For a specific ROI r_m within a cluster F_k, we determine whether it is an outlier in terms of two affinity scores.

Score 1: Intra-Cluster affinity. We define the affinity score of an annotation r_m within its own cluster as

$$p_1^m = \max_{r_n \in F_k, n \neq m} (1 - \mathrm{GIoU}(r_m, r_n)),$$

where r_n represents other ROIs in F_k.

Score 2: Inter-Cluster affinity. We further compute its affinity to other clusters as:

$$p_2^m = \frac{1}{K-1} \sum_{k'=1, k' \neq k}^{K} \sum_{a_n \in F_{k'}} \frac{RIR(r_m, r_n)}{|F_{k'}|}.$$

Intuitively, a ROI that excessively matches other clusters while failing to match the ROIs within its own cluster is considered as an outlier. We then model the event that ROI r_m is an outlier with a Bernoulli distribution $Ber(p_1 \cdot p_2)$. Once ROI is identified as an outlier w.r.t. the Bernoulli event, it is then removed.

The time complexity of the above process is $O(K \cdot N^2)$, N is the number of ROIs on a single image, and satisfies that $K \ll N$. Therefore, the time complexity is acceptable.

5 ANNOTATION AGGREGATION MODULE

In this section, we aggregate both the ROIs and categories/labels given by the annotations per cluster. Recall that we introduce σ (Equation (4)) and η (Equation (5)) to represent the annotation quality and worker quality, respectively. In addition, r_m's and r_m^{sam}'s are the worker ROIs and SAM-refined ROIs, respectively.

Step 1: Aggregation of ROIs. The aggregated ROI for cluster F_k is resolved by considering both the worker ROIs and the SAM-drawn ROIs. We define r_k^{worker} and r_k^{sam} as the aggregated worker ROI and SAM-drawn ROI for the cluster F_k, which are computed as

$$r_k^{worker} = \frac{\sum\limits_{a_m \in F_k} \sigma(a_m)\eta(w(a_m))r_m}{\sum\limits_{a_m \in F_k} \sigma(a_m)\eta(w(a_m))},$$

$$r_k^{sam} = \frac{\sum\limits_{a_m \in F_k} \sigma(a_m)\eta(w(a_m))r_m^{sam}}{\sum\limits_{a_m \in F_k} \sigma(a_m)\eta(w(a_m))}.$$

Intuitively, they are computed as the average of worker and SAM's ROIs, with the weights given by the product of η and σ. The overall confidence of SAM is computed analogously:

$$\text{CONF}_k^{sam} = \frac{\sum\limits_{a_m \in F_k} \sigma(a_m)\eta(w(a_m))\text{conf}_m^{sam}}{\sum\limits_{a_m \in F_k} \sigma(a_m)\eta(w(a_m))}.$$

A naive method is to generate the eventual ROI for F_k by a weighted average of r_k^{worker} and r_k^{sam} with weights as CONF_k^{sam} and $1 - \text{CONF}_k^{sam}$. However, the weight ignores the relations between the two ROIs, and could be too aggressive. Thus, we obtain a new weight by smoothing CONF_k^{sam} with the GIoU score between the two ROIs:

$$\text{weight}_k = \frac{1}{2}(\text{CONF}_k^{sam} + \text{GIoU}(r_k^{worker}, r_k^{sam})).$$

The final aggregated ROI for F_k is then

$$\hat{r_k} = (1 - \text{weight}_k) \cdot r_k^{worker} + \text{weight}_k \cdot r_k^{sam}$$

Step 2: Aggregation of labels.

For category aggregation, we choose the label which has the maximum votes weighted by both σ and η.

$$\hat{c_k} = \arg\max_{c \in C} \sum_{a_m \in F_k} \sigma(a_m)\eta(w(a_m))\mathbb{I}(c, c_m)$$

where c_m is the category given by a_m.

The complexity for aggregating ROIs and categories involves iterating over the annotations in each cluster, leading to a linear complexity of $O(N)$, and N is the number of annotations in a figure.

6 EXPERIMENTS

In this section, we first describe the datasets (Section 6.1), the baseline methods & evaluation metrics (Section 6.2), and the parameter settings (Section 6.3), respectively. Then, for the results, we compare our method to other aggregation methods in Section 6.4, to pure CV-based methods in Section 6.6. Additionally, we present the results of ablation experiments in Section 6.5. Finally, we provide a detailed analysis of the computational performance, including the time taken to run our method and other methods, in Section 6.7.

Our datasets and code are publicly accessible[1].

6.1 Datasets

Dataset	Workers	Tasks	ROIs	GT ROIs
Cytology image dataset	4	446	14,579	6,402
COCO image dataset	13	57	1,930	245

Table 3: Datasets Statistics

At present, a noticeable absence persists in the publicly available crowdsourced annotation datasets for the medical field. While numerous datasets cater to basic labeling tasks across other domains, the gap between the fields cannot be ignored due to the difference in the image characteristics and employed workers: 1) medical images usually have a higher resolution and contain rich semantics; 2) they require expert physicians for annotation. Acknowledging this gap, we embark on an initiative to develop authentic datasets for evaluation purposes, the cytology annotation dataset as introduced below.

Cytology Image Dataset: Four physicians are enlisted to annotate cervical cytology images, with an example shown in Figure 1. The annotation results underwent comprehensive review by an expert physician, whose assessments are considered as the Ground Truth. This dataset encompasses 33 distinct cell categories, alongside an "unclassified" category, with detailed statistics provided in Table 3. In total, the dataset comprises 446 cytology images, incorporating 14,579 ROIs, and a cumulative count of 6,402 cell regions as per the evaluations conducted by expert physicians.

Nevertheless, to validate the generalizability of the proposed approach, we still curate a comprehensive and complex-labeled crowdsourcing dataset based on images of the COCO dataset [27].

COCO Image Dataset: We engage 13 workers to annotate a subset of images from the COCO dataset, as detailed in Table 3. This subset comprises a total of 57 images across 6 categories—people, dogs, cars, sheep, cows, and cats—with 1,930 ROIs meticulously collected.

6.2 Baselines and Evaluation

To the best of our knowledge, we are not aware of any current work on the aggregation for multi-size ROIs with diverse categories, let alone one specially tailored for medical images. Nevertheless, we compare our method STARLET to six state-of-art aggregation methods for complex labels as close to our annotation target as possible:

- **BVHP [3]:** The BVHP approach models worker skill and image difficulty, treating the clustering process as a facility localization problem [11]. It aggregates the annotated ROIs but does not consider their diverse categories.
- **BAU [4]:** The BAU method identifies the best workers by modeling their capabilities, and subsequently use their annotated ROIs as the aggregation results.
- **SAD [4]:** The SAD method selects ROIs based on their distance to other ROIs, choosing those with the shortest distances as the aggregated result.

[1] https://github.com/YHSI5358/anno-code

- **MAS [4]:** The MAS method constructs a hierarchical Bayesian probabilistic model with a multidimensional scaling likelihood function, integrating worker probability and task difficulty to select the optimal ROIs as the aggregation result.
- **PSR-MAS [5]:** The PSR-MAS method employs a Partition-Selection-Recombination process to select the optimal ROI for each partition as the aggregation result.
- **SMAS [6]:** The SMAS method represents the unknown Ground Truth as an embedding vector. By retaining 10% of the dataset as the golden dataset, it uses semi-supervised learning to estimate the embedding distance between the worker's ROIs and the real ROI, selecting the optimal ROIs as the aggregation result. In our experiment, we similarly randomly select 10% of our dataset as the golden dataset to apply the method for comparison.
- **RetinaNet [26]:** RetinaNet is a pure CV-based method, used for object detection. In this study, we train it on a broader dataset of cytological images with annotations, consisting of 4,534 images and 35,126 ROIs.
- **Only-SAM [20]:** As mentioned in Section 4.1, SAM is also a state-of-art CV-based method. In this experiment, we first feed workers' ROIs as prompts to SAM, whose output (i.e., r_m^{sam}'s) is directly treated as the eventual inferred ROIs.

All of the above methods, except for **BVHP** and the two CV methods, use hierarchical clustering [41] to obtain the ROI clusters, while we propose a customized spectral clustering method in STARLET. To demonstrate its significance, we use the same hierarchical clustering method (as the baselines) to substitute our clustering component of STARLET in the ablation experiment (Section 6.5).

Except for the two CV-based baselines, the above baselines only aggregate ROIs but do not reach a consensus on their label. However, each aggregated ROI represents a set of ROIs associated possibly different categories. To adapt them to the studied problem, we then employ several existing category aggregation methods to supplement the above baselines:

- **Majority Voting (MV):** The majority voting method counts the number of votes for each category within the cluster and selects the category with the highest number of votes as the result.
- **CATD [22]:** The CATD method models worker confidence, assuming that workers with more answers provide higher quality.
- **D&S [8]:** The D&S method models workers' behaviours using confusion matrices. It employs the expectation-maximization [10] method to iteratively update the model parameters and the inferred category.
- **PM [23]:** The PM method models the worker using a probability (from 0 to 1) that represents the accuracy of the worker's answer, and then iteratively updates the aggregation results and the probability for each worker.
- **LFC [33]:** LFC is an extension of the **D&S** method. It further estimates worker quality by assuming it is known a priori through a Beta distribution.
- **ZenCrowd [9]:** The ZenCrowd method uses a probabilistic graphical model, which attempts to maximize the probability of workers' answers, called the *likelihood*.

Evaluation metric. An annotation consists of a framed area and an associated category, so does the eventual aggregated result. Therefore, the evaluation consists of two parts: the consistency between the aggregated ROI area and the ground truth, and the correctness of its category.

Evaluation of ROI: we use Braylan's F1-score [5], which is widely used in existing works [4–6], to assess the quality of the aggregated ROIs, which is formulated based on the *Intersection over Union* (IoU) between ROIs as described in the following.

We define $\hat{A}$ as the set of annotations inferred/aggregated from the annotations of an image, and G as the set of ground truth ROIs of the objects. We compute the scores of precision and recall per object as follows:

$$P(\hat{A}, G) = \left\{\max(\text{IoU}(r_m, g_n) \mid g_n \in G\}) \mid r_m \in \hat{A}\right\} \tag{16}$$

$$R(\hat{A}, G) = \left\{\max(\text{IoU}(r_m, g_n) \mid r_m \in \hat{A}\}) \mid g_n \in G\right\} \tag{17}$$

Braylan's F1-score is then calculated as:

$$\text{Braylan's F1-score} = \frac{2 \times \text{mean}(P(\hat{A}, G)) \times \text{mean}(R(\hat{A}, G))}{\text{mean}(P(\hat{A}, G)) + \text{mean}(R(\hat{A}, G))} \tag{18}$$

This metric provides a comprehensive assessment by simultaneously considering both precision and recall. However, it may also incur that a ROI in $\hat{A}$ is repeatedly considered for different ROIs in G. For instance, if ROI $a \in \hat{A}$ overlaps with both ROIs b and c in G, matching a to both of them is apparently improper, which means an answer is repeatedly used for multiple objects. To address this issue, we also report the One2one F1-score, where each ROI $\in \hat{A}$ is mapped exclusively to a single GT ROI.

Specifically, in One2one F1-score, each ROI in $\hat{A}$ is matched to the ROI in G that has the highest IoU value. This ensures that each aggregated ROI is paired with the best possible ground truth ROI, thereby avoiding the issue of multiple mappings for a single ROI.

Evaluation of category: We use several metrics to evaluate the category quality, including precision, recall, F1-score. These metrics are defined as follows:

Precision measures the proportion of correctly classified ROIs among all aggregated ROIs:

$$\text{Precision} = \frac{\text{Number of correctly classified ROIs}}{\text{Total number of aggregated ROIs}} \tag{19}$$

Recall assesses the proportion of correctly classified ROIs among all ground truth ROIs:

$$\text{Recall} = \frac{\text{Number of correctly classified ROIs}}{\text{Total number of ground truth ROIs}} \tag{20}$$

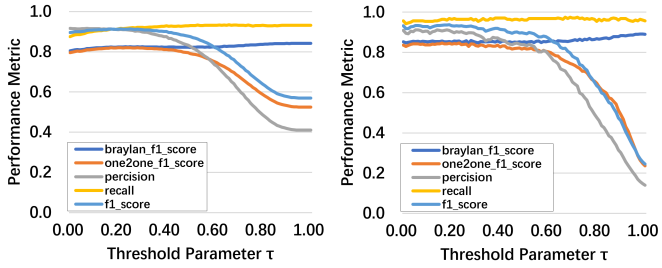
F1-score is the harmonic mean of precision and recall, providing a single metric that balances both aspects:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{21}$$

6.3 Parameter Settings

There is only one parameter for STARLET, i.e., the similarity threshold τ in Algorithm 2, which helps to determine the cluster number.

The value of this parameter is set to 0.2 on default. We also report the effect and sensitivity of τ in Figure 6.



(a) Cytology dataset. (b) COCO dataset.

Figure 6: The effect and sensitivity of τ in Cytology dataset (left) and COCO dataset (right).

In the Figure 6, recall that Braylan's F1-score and One2one F1-score are for ROI quality, while others are for the labeling quality. It can be seen that as τ keeps increasing, the precision and One2one F1-score as well as the F1-score of STARLET slowly decrease, while its Braylan's F1-score and recall are almost unchanged (even increase slightly). This is because τ, as the threshold value for calculating the similarity matrix, has a large impact on the number of clusters. When the value τ is too large, it would lead to a large number of clusters. This corresponds to more predictions inside the images, increasing the recall while decreasing the precision and One2one F1-score. It is worth noting that Braylan's F1-score is almost unaffected by the change of τ, in contrast to the changing One2one F1-score. This also justifies the effectiveness of the proposed One2one F1-score as a metric.

According to the results, we choose $\tau = 0.2$, which shows satisfying performances for all the metrics.

6.4 Basic Results

This section presents the comparison results among the crowdsourcing-based methods, which is further divided into ROI quality evaluation and label evaluation, shown in Table 4, Table 5 & 6 respectively.

	Cytology Dataset		COCO Dataset	
Method	Braylan's F1	One2one F1	Braylan's F1	One2one F1
BVHP	0.6511	0.6401	0.8040	0.7919
BAU	0.6179	0.6012	0.8349	0.8264
SAD	0.6004	0.5847	0.7692	0.7594
MAS	0.6198	0.6050	0.8223	0.8146
PSR-MAS	0.6543	0.6428	0.8328	0.8233
SMAS	0.6877	0.6762	0.8132	0.8034
STARLET (ACM)	**0.8282**	**0.8238**	**0.8535**	**0.8431**

Table 4: ROI evaluation in Cytology and COCO Datasets.

6.4.1 ROI quality. As shown in Table 4, our method STARLET achieves the highest performance in terms of both Braylan's F1 and One2one F1 metrics. Specifically, in Cytology dataset, it outperforms the second-best method, SMAS, by 20.43% in Braylan's F1 and 21.83% in One2one F1. Notably, the One2one F1-scores are generally slightly lower than Braylan's F1-scores. The smaller the difference between these two scores, the higher the robustness of a method. In Cytology dataset, STARLET has a difference of only 0.5% between the two scores, whereas other methods exhibit differences ranging from 1.7% to 2.8%. This further highlights the superiority of our approach.

6.4.2 Labeling quality. Table 5 and 6 show the labeling quality evaluation results, where each column is a clustering method and each row is a label aggregation method. Each cell in the table then represents an integration of certain clustering and aggregation method. We select the top three clustering methods as reported in Table 4, i.e., STARLET (ACM), SMAS, and BVHP. STARLET (ACM) and STARLET (AAM) denote the clustering and aggregation modules in STARLET, respectively. The result of STARLET is then presented at the bottom-right corner cell.

In the Cytology dataset, STARLET as a composition of AAM and ACM, achieves the highest precision of 0.9151, while the precision of other methods remain no larger than 0.9. Meanwhile, STARLET also achieves the highest Recall and F1-score, having an advantage of 1.05% and 1.06% over the second best, respectively. Furthermore, ACM remarkably boosts the performance of other pure category aggregation methods, which obtains 63.45% and 29.51% improvements in F1-score on average, compared to their coupling with BVHP and SMAS.

In the COCO dataset, STARLET also achieves the highest F1-score of 0.9416, making a large margin of 5% over other methods. Similar effects can be observed in terms of recall (a margin of 11%). In contrast, STARLET fails to rank top in terms of precision, while left behind other label aggregation methods boosted by ACM. Nevertheless, its gap to the best in terms of precision is around 2%, much smaller than its advantage shown over the other two metrics. Overall, STARLET is still the best and most robust across all the metrics.

6.5 Ablation Study

This section conducts an ablation study to discern the individual contributions of each module within STARLET to the overall performance. Table 7 delineates the performance of the complete approach alongside the one subsequent to the removal of each module/component, where the symbol '-' denotes removal of a component.

In Table 7, "STARLET-Crowdsourcing-oriented Spectral Clustering" signifies the substitution of our crowdsourcing-oriented spectral clustering with native spectral clustering, utilizing RatioCut [17] for clustering. For AAM removal (STARLET-AAM), the associated mark "Mean(Median)+MV" denotes the substitution. 'Mean(Median)+MV" denotes the ROI aggregation by simply their average (median) value, and the label aggregation by simply majority voting:

- **Mean:** For each cluster, we average the coordinates corresponding to the upper left and lower right corners of the ROIs, denoted as $\langle x_1, y_1, x_2, y_2 \rangle$.
- **Median:** We calculate the median of each coordinate value, and re-organize the four medians into a ROI.

Method	Precision			Recall			F1-score		
	BVHP	SMAS	STARLET (ACM)	BVHP	SMAS	STARLET (ACM)	BVHP	SMAS	STARLET (ACM)
CATD	0.6129	0.8645	**0.8914**	0.5127	0.5569	**0.8988**	0.5583	0.6774	**0.8951**
D&S	0.6024	0.8795	**0.9006**	0.5039	0.5665	**0.9082**	0.5488	0.6891	**0.9044**
MV	0.6030	0.9028	**0.9080**	0.5043	0.5815	**0.9056**	0.5493	0.7074	**0.9068**
PM	0.6086	0.8635	**0.8908**	0.5100	0.5562	**0.8982**	0.5550	0.6766	**0.8945**
LFC	0.6056	0.8928	**0.8952**	0.5066	0.5751	**0.9027**	0.5517	0.6996	**0.8989**
ZenCrowd	0.6028	0.8872	**0.8974**	0.5021	0.5715	**0.9049**	0.5479	0.6952	**0.9011**
STARLET(AAM)	0.6072	0.9001	**0.9151**	0.5095	0.6173	**0.9177**	0.5541	0.7323	**0.9164**

Table 5: Results of Category Aggregation in Cytology Datasets.

Method	Precision			Recall			F1-score		
	BVHP	SMAS	STARLET (ACM)	BVHP	SMAS	STARLET (ACM)	BVHP	SMAS	STARLET (ACM)
CATD	0.6085	0.7008	**0.9377**	0.4982	0.6584	**0.8612**	0.5479	0.6789	**0.8978**
D&S	0.6071	0.6894	**0.9298**	0.4946	0.6477	**0.8541**	0.5451	0.6679	**0.8903**
MV	0.6094	0.6970	**0.9300**	0.4991	0.6548	**0.8558**	0.5488	0.6752	**0.8914**
PM	0.6105	0.6932	**0.9318**	0.5011	0.6512	**0.8431**	0.5504	0.6715	**0.8852**
LFC	0.6209	0.6780	**0.9316**	0.4972	0.6370	**0.8607**	0.5522	0.6569	**0.8947**
ZenCrowd	0.6203	0.6962	**0.9300**	0.4899	0.6785	**0.8558**	0.5474	0.6872	**0.8914**
STARLET(AAM)	0.6261	0.7112	**0.9103**	0.4946	0.7011	**0.9751**	0.5526	0.7061	**0.9416**

Table 6: Results of Category Aggregation in COCO Datasets.

Method	Braylan's F1	One2one F1	Precision	Recall	F1-score
STARLET-Cluster Number Determination	0.7404	0.7242	0.9142	0.7609	0.8305
STARLET-Crowdsourcing-oriented Spectral Clustering	0.8097	0.7957	0.8880	0.8896	0.8888
STARLET-Outlier Removal	0.8188	0.8134	0.8912	0.9103	0.9006
STARLET-AAM(Mean+MV)	0.7158	0.7113	0.9080	0.9056	0.9068
STARLET-AAM(Median+MV)	0.7220	0.7176	0.9080	0.9056	0.9068
STARLET	**0.8282**	**0.8238**	**0.9151**	**0.9177**	**0.9164**

Table 7: Results of ablation experiments.

As shown in Table 7, the removal or replacement of any module/method results in a decrease in all performance metrics. Specifically, after removing AAM, Braylan's F1 and One2one F1 are declined by 12.82% and 12.89%, respectively, compared to the full STARLET. This finding reinforces the significance of integrating the image feature into the aggregation process, highlighting the efficacy of leveraging the inherent image information as AAM does.

Similarly, modifying the clustering component leads to a decline. Either the removal of the cluster number determination module or the latter spectral clustering module brings performance degradation, This demonstrates the need for a two-stage cluster process, first deciding the cluster number and then deciding the membership, as the proposed ACM does.

6.6 Comparison with CV Method

We also compare STARLET to two CV-based methods on Cytology dataset, as described in Section 6.2, i.e., RetinaNet and SAM.

Method	Braylan's F1	One2one F1
RetinaNet	0.3965	0.3839
Only-SAM	0.7074	0.6899
Ours	**0.8282**	**0.8238**

Table 8: Comparison results with CV methods.

As shown in Table 8, even with a large amount of ground truth data for training, the detection results of the RetinaNet model are significantly lower. When using only the SAM method, despite using workers' ROIs as prompts, its performance is still inferior to our proposed method, with gaps of 17.08% and 19.41% in terms of the two metrics respectively. This demonstrates that purely relying on CV-based methods is not enough to tackle the task of medical image annotation.

6.7 Time Cost Statistics

To assess the efficiency of STARLET, we measure the average processing time per image for all the methods, on a system with an Intel Core i5-14600K processor and 32GB of RAM. The results are summarized in Table 9.

We report the time cost in Table 9, where all inference methods are combined with all clustering methods, with the totally integrated methods as a grid. It can be found that the integrations with TARLET (ACM) or STARLET (AAM) have slightly higher time costs compared to other methods. However, their runtime remains acceptable for aggregating medical images, as a moderate computational overhead. STARLET's balance between accuracy and processing time supports its practical use in scenarios demanding high-quality results.

Method	BVHP	SMAS	STARLET (ACM)
CATD	0.0333	0.1847	0.1650
D&S	0.0366	0.1880	0.1683
MV	0.0317	0.1831	0.1634
PM	0.0321	0.1835	0.1638
LFC	0.0365	0.1880	0.1683
ZenCrowd	0.0343	0.1857	0.1660
STARLET(AAM)	0.0434	0.1948	0.1751

Table 9: Average processing times (in seconds) for different methods.

7 RELATED WORK

Aggregation of Simple Annotations. In the field of crowdsourcing, simple annotation aggregation methods typically refer to approaches that determine relationships between answers through exact matching, such as in single-choice, multiple-choice, and true/false tasks. These methods are characterized by fixed options for each task. Zheng et al.[42] conducted a comprehensive survey of various aggregation algorithms for simple annotations. These methods are primarily used in unsupervised settings, iteratively inferring answers by estimating worker quality and task difficulty. Although these methods are commonly employed for inferring simple annotation types, they are not suitable for aggregating complex annotation tasks.

Aggregation of Complex Annotations. Complex annotation tasks are considered more intricate compared to simple annotation tasks, as they generally involve tasks where exact matching of answers is not possible. For example, tasks like named entity recognition or object detection often yield highly variable answers from different annotators. In the field of general image annotation, Liu et al.[28] used weighted estimation to aggregate annotations. Feng et al. [12, 13] utilized image features (HOG and LBP feature extraction) to fine-tune and aggregate each annotation. Braylan et al. [4] modeled annotation distance matrices and employed hierarchical Bayesian methods for quality estimation to select the final annotation. However, these selection-based methods are fundamentally limited by the annotators' abilities. And research [5, 6] addressed this by partitioning the task and then splitting and combining annotations to construct new results. While this approach is not constrained by the annotators' capabilities, there is still room for improvement in the partitioning methods.

Crowdsourcing in Medical Field. Crowdsourcing in the medical domain poses unique challenges due to the complexity and sensitivity of medical data. Early works, such as those by Foncubierta-Rodríguez and Müller [14], explored the potential of crowdsourcing for medical research, emphasizing the need for specialized knowledge and training for annotators. In medical image annotation, complex tasks involve multi-sub-task annotations, uncertain object sizes, and multi-category objects. Bhatti et al. [1] has underscored the benefits of decomposing these tasks into manageable sub-tasks to enhance efficiency.

In the realm of medical image crowdsourcing, numerous prior studies [18, 32] have commonly employed majority voting aggregation methods. In the work by Keshavan et al. [19], the XGBoost algorithm is leveraged to estimate worker weights, facilitating the adoption of weighted voting strategies. Moreover, Brady et al. [2] take into account the task's complexity, incorporating an estimation of worker accuracy based on task difficulty considerations, and conduct experiments using the Diabetic Retinopathy (DR) dataset.

8 CONCLUSION

In this study, we present STARLET, a novel framework for aggregating complex medical image annotations through crowdsourcing. Our method effectively handles the intricacies of medical image annotation, such as varying object sizes, multi-category objects, and the need for high precision. Experimental results validate that STARLET not only surpasses existing computer vision methods but also enhances the aggregation accuracy by utilizing crowdsourcing-oriented spectral clustering and quality-enhancing modules. The integration of these modules ensures a holistic and accurate annotation process, demonstrating the framework's potential for improving medical image analysis and supporting downstream medical applications. Future work will focus on further refining the framework and exploring its application to other complex annotation tasks in the medical field.

REFERENCES

[1] Shahzad Sarwar Bhatti, Xiaofeng Gao, and Guihai Chen. 2020. General framework, opportunities and challenges for crowdsourcing techniques: A Comprehensive survey. *J. Syst. Softw.* 167 (2020), 110611.

[2] Christopher John Brady, Lucy Iluka Mudie, Xueyang Wang, Eliseo Guallar, and David Steven Friedman. 2017. Improving consensus scoring of crowdsourced data using the Rasch model: development and refinement of a diagnostic instrument. *Journal of medical Internet research* 19, 6 (2017), e7984.

[3] Steve Branson, Grant Van Horn, and Pietro Perona. 2017. Lean Crowdsourcing: Combining Humans and Machines in an Online System. In *CVPR*. IEEE Computer Society, 6109–6118.

[4] Alexander Braylan and Matthew Lease. 2020. Modeling and Aggregation of Complex Annotations via Annotation Distances. In *WWW*. ACM / IW3C2, 1807–1818.

[5] Alexander Braylan and Matthew Lease. 2021. Aggregating Complex Annotations via Merging and Matching. In *KDD*. ACM, 86–94.

[6] Alexander Braylan, Madalyn Marabella, Omar Alonso, and Matthew Lease. 2023. A General Model for Aggregating Annotations Across Simple, Complex, and Multi-Object Annotation Tasks. *J. Artif. Intell. Res.* 78 (2023), 901–973.

[7] Yu Chen, Sheng Zhang, Yibo Jin, Zhuzhong Qian, Mingjun Xiao, Wenzhong Li, Yu Liang, and Sanglu Lu. 2024. Crowdsourcing Upon Learning: Energy-Aware Dispatch With Guarantee for Video Analytics. *IEEE Trans. Mob. Comput.* 23, 4 (2024), 3138–3155.

[8] Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 28, 1 (1979), 20–28.

[9] Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. 2012. ZenCrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *WWW*. ACM, 469–478.

[10] Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)* 39, 1 (1977), 1–22.

[11] Donald Erlenkotter. 1978. A Dual-Based Procedure for Uncapacitated Facility Location. *Oper. Res.* 26, 6 (1978), 992–1009.

[12] Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiaojun Wu. 2017. Face Detection, Bounding Box Aggregation and Pose Estimation for Robust Facial Landmark Localisation in the Wild. In *CVPR Workshops*. IEEE Computer Society, 2106–2115.

[13] Zhen-Hua Feng, Josef Kittler, William J. Christmas, Patrik Huber, and Xiaojun Wu. 2017. Dynamic Attention-Controlled Cascaded Shape Regression Exploiting Training Data Augmentation and Fuzzy-Set Sample Weighting. In *CVPR*. IEEE Computer Society, 3681–3690.

[14] Antonio Foncubierta-Rodríguez and Henning Müller. 2012. Ground truth generation in medical imaging: a crowdsourcing-based iterative approach. In *CrowdMM@ACM Multimedia*. ACM, 9–14.

[15] Robert Grone and Russell Merris. 1994. The Laplacian spectrum of a graph II. *SIAM Journal on discrete mathematics* 7, 2 (1994), 221–229.

[16] Anne Grote, Nadine S. Schaadt, Germain Forestier, Cédric Wemmert, and Friedrich Feuerhake. 2019. Crowdsourcing of Histological Image Labeling and Object Delineation by Medical Students. *IEEE Trans. Medical Imaging* 38, 5 (2019), 1284–1294.

[17] Lars W. Hagen and Andrew B. Kahng. 1992. New spectral methods for ratio cut partitioning and clustering. *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.* 11, 9 (1992), 1074–1085.

[18] Eric Heim. 2018. *Large-scale medical image annotation with quality-controlled crowdsourcing.* Ph.D. Dissertation. University of Heidelberg, Germany.

[19] Anisha Keshavan, Jason D. Yeatman, and Ariel Rokem. 2019. Combining Citizen Science and Deep Learning to Amplify Expertise in Neuroimaging. *Frontiers Neuroinformatics* 13 (2019), 29.

[20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. 2023. Segment Anything. In *ICCV*. IEEE, 3992–4003.

[21] Khiem H. Le, Tuan V. Tran, Hieu H. Pham, Hieu T. Nguyen, Tung T. Le, and Ha Q. Nguyen. 2023. Learning From Multiple Expert Annotators for Enhancing Anomaly Detection in Medical Image Analysis. *IEEE Access* 11 (2023), 14105–14114.

[22] Qi Li, Yaliang Li, Jing Gao, Lu Su, Bo Zhao, Murat Demirbas, Wei Fan, and Jiawei Han. 2014. A Confidence-Aware Approach for Truth Discovery on Long-Tail Data. *Proc. VLDB Endow.* 8, 4 (2014), 425–436.

[23] Qi Li, Yaliang Li, Jing Gao, Bo Zhao, Wei Fan, and Jiawei Han. 2014. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *SIGMOD Conference*. ACM, 1187–1198.

[24] Xiang Li, Chengqi Lv, Wenhai Wang, Gang Li, Lingfeng Yang, and Jian Yang. 2023. Generalized Focal Loss: Towards Efficient Representation Learning for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 3 (2023), 3139–3153.

[25] Yuan-Hong Liao, Amlan Kar, and Sanja Fidler. 2021. Towards Good Practices for Efficiently Annotating Large-Scale Image Classification Datasets. In *CVPR*. Computer Vision Foundation / IEEE, 4350–4359.

[26] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal Loss for Dense Object Detection. In *ICCV*. IEEE Computer Society, 2999–3007.

[27] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *ECCV (5) (Lecture Notes in Computer Science)*, Vol. 8693. Springer, 740–755.

[28] Shu Liu, Cewu Lu, and Jiaya Jia. 2015. Box Aggregation for Proposal Decimation: Last Mile of Object Detection. In *ICCV*. IEEE Computer Society, 2569–2577.

[29] Helmut Lütkepohl. 1997. *Handbook of matrices.* John Wiley & Sons.

[30] Bojan Mohar. 1997. Some applications of Laplace eigenvalues of graphs. In *Graph symmetry: Algebraic methods and applications*. Springer, 225–275.

[31] An Thanh Nguyen, Byron C. Wallace, Junyi Jessy Li, Ani Nenkova, and Matthew Lease. 2017. Aggregating and Predicting Sequence Labels from Crowd Annotations. In *ACL (1)*. Association for Computational Linguistics, 299–309.

[32] Silas Nyboe Ørting, Andrew Doyle, Arno van Hilten, Matthias Hirth, Oana Inel, Christopher R. Madan, Panagiotis Mavridis, Helen Spiers, and Veronika Cheplygina. 2020. A Survey of Crowdsourcing in Medical Image Analysis. *Hum. Comput.* 7 (2020), 1–26.

[33] Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning From Crowds. *J. Mach. Learn. Res.* 11 (2010), 1297–1322.

[34] Xuhua Ren, Sahar Ahmad, Lichi Zhang, Lei Xiang, Dong Nie, Fan Yang, Qian Wang, and Dinggang Shen. 2020. Task Decomposition and Synchronization for Semantic Biomedical Image Segmentation. *IEEE Trans. Image Process.* 29 (2020), 7497–7510.

[35] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian D. Reid, and Silvio Savarese. 2019. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In *CVPR*. Computer Vision Foundation / IEEE, 658–666.

[36] Jianbo Shi and Jitendra Malik. 2000. Normalized Cuts and Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 8 (2000), 888–905.

[37] Ulyana Tkachenko, Aditya Thyagarajan, and Jonas Mueller. 2023. ObjectLab: Automated Diagnosis of Mislabeled Images in Object Detection Data. *CoRR* abs/2309.00832 (2023).

[38] Ulrike von Luxburg. 2007. A tutorial on spectral clustering. *Stat. Comput.* 17, 4 (2007), 395–416.

[39] Dorothea Wagner and Frank Wagner. 1993. Between Min Cut and Graph Bisection. In *MFCS (Lecture Notes in Computer Science)*, Vol. 711. Springer, 744–750.

[40] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. 2024. YOLOv10: Real-Time End-to-End Object Detection. *CoRR* abs/2405.14458 (2024).

[41] Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* 58, 301 (1963), 236–244.

[42] Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. 2017. Truth Inference in Crowdsourcing: Is the Problem Solved? *Proc. VLDB Endow.* 10, 5 (2017), 541–552.