Bio-Medical Data Classification Approaches with Limited Annotation

Shota Harada[1,a)] Seiichi Uchida[1,b)]

Abstract: This paper aims to solve the problem of limited annotation in several bio-medical data analysis tasks. Specifically, group-based labeling utilizing constrained clustering and semi-supervised learning are proposed as the approaches. For group-based labeling utilizing constrained clustering, I proposed a new constrained clustering method, where a user attaches annotations to several sample pairs. Annotations are two types: cannot-link and must-link. The pair with cannot-link should not belong to the same cluster, whereas the pair with must-link should belong. These annotations are useful especially for medical data, because medical experts can have a more expected clustering result by a small number of annotations. Moreover, those annotations are treated as soft-constraints and therefore medical experts can attach them without extreme carefulness. For semi-supervised learning in bio-medical data classification tasks, I proposed order-guided disentangled representation learning. This method performs disentangled representation learning with prior knowledge that is effective for learning bio-medical data classification tasks. This method could improve classification performance even with limited annotation by effectively utilizing the prior knowledge through disentangled representation learning.

Keywords: Bio-medical data classification, Soft-constrained clustering, Semi-supervised learning

1. Introduction

1.1 Background

Building a diagnostic support system is one of the important tasks to help medical doctors. Medical doctors diagnose various diseases from various data, such as endoscopic images, electrocardiograms, and brain magnetic resonance imaging. Diagnostic support systems aim to enhance decision-making of medical doctors by using Artificial Intelligence (AI) technologies.

As shown in Figure 1, there are two main strategies to realize diagnosis support systems using AI technology. One is a rule-based method [1], which utilizes heuristic rules manually specified by medical doctors. To construct a rule-based method, we first extract expert knowledge from medical doctors and design decision rules based on the extracted knowledge. The other is a machine learning-based method [2, 3], which utilizes a large amount of actual diagnosis cases to derive decision rules automatically and statistically. In constructing a machine-based method, we first collect data and attach annotations, such as diagnostic results and disease information, to each sample. We train a machine learning-based model using annotated data as actual diagnosis cases.

The development of deep learning techniques has made it possible to implement a machine learning-based diagnostic support system with high accuracy. In recent years, deep learning techniques have been applied to various tasks in many fields, such as image recognition, signal recognition, and natural language processing, and achieving an excellent performance of these tasks. Deep learning techniques have already been applied to many tasks in bio-medical fields [4–6].

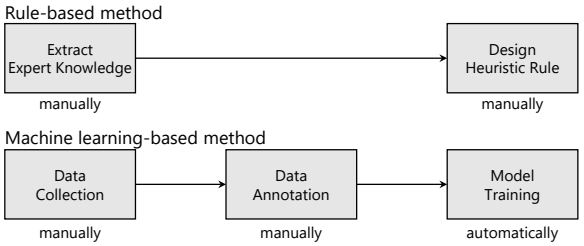


Fig. 1: A standard flow of a diagnostic support system construction.

Deep learning-based methods require a large amount of annotated samples to increase their performance. In general, for a machine-based method, we first collect data and attach annotations, such as diagnostic results and disease information, to each sample. Then, we train a machine learning-based model using annotated data. For deep learning, far more data is necessary than other machine learning methods, to fully utilize its high flexibility.

However, it is difficult to obtain a large amount of annotated samples in bio-medical data analysis tasks because the number of annotators is limited. In the general object image classification task, such as ImageNet classification, it is easy to prepare a large number of annotators using crowdsourcing services such as Amazon Mechanical Turk. In contrast, it is difficult to secure a large number of annotators for bio-medical data analysis tasks because it can deal only with medical experts. Due to the reason, we are often forced to build a deep learning-based diagnostic support

[1] Kyushu University, Fukuoka, Japan
[a)] shota.harada@human.ait.kyushu-u.ac.jp
[b)] uchida@ait.kyushu-u.ac.jp

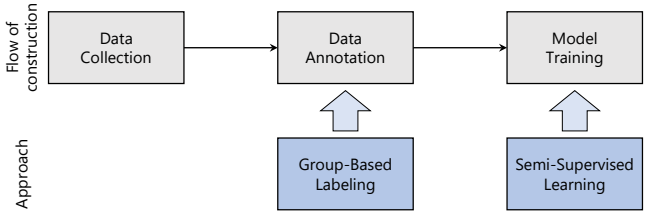


Fig. 2: Approaches that correspond to each step for constructing a machine learning-based method to solve the problem of limited annotation.

system with limited annotations for training.

1.2 Motivation

Researchers have proposed various methods to solve the problem of limited annotation in classification tasks [7–11]. This paper focuses on two approaches: group-based labeling and semi-supervised learning, which correspond to each step for constructing a machine learning-based method, as shown in Figure 1.

Figure 2 shows the correspondence between each approach of this paper and each step for constructing a machine learning-based method. Group-based labeling is an approach that corresponds data annotation step, and semi-supervised learning is an approach for model training step.

1.2.1 Group-Based Labeling with Constrained Clustering

Group-based labeling is an efficient strategy for collecting a sufficient amount of annotated samples for machine learning-based methods [9, 10, 12–18]. In group-based labeling, we first gather similar samples by clustering and then have an expert labels the samples of each cluster (i.e., each group). As a result, annotation costs drastically decrease by this cluster-wise annotation process, compared to the sample-wise general annotation process.

To effectively reduce annotation costs with group-based labeling, the purity of clusters is important. If the annotator has a glance at the samples of each cluster and finds that a certain cluster comprises the samples from the same class, they can give the same class label to all the samples at once. Therefore, to annotate efficiently with group-based labeling, we need to obtain clusters with high purity.

Constrained clustering that uses annotations for sample pairs as constraints is a promising approach to improve the purity of clusters for efficient group-based labeling. Annotations for sample pairs are two types: cannot-link and must-link. The cannot-link is attached to the sample pair that should belong to different clusters, whereas the must-link is attached to the sample pair that should belong to the same cluster. Unlike typical unsupervised clustering methods, constrained clustering optimizes clusters while satisfying those links. If medical expert picks a limited number of confusing sample pairs and attach the links to them in advance, we may obtain high purity clusters by applying a constrained clustering.

In the bio-medical data clustering scenario, we should handle two issues. The first issue is that the must-link is not always useful. Let us assume a clustering task of endoscopic images, where each cluster should contain images of a single organ or component of an organ. In this task, the variance of image appearances in a class is large because they are affected by some factors such as camera angle and light source intensity. Since those image appearances are very different, satisfying the must-link of such image pair results in an unexpectedly large cluster with low purity. The second issue is that the annotation cost of attaching the links of bio-medical data is high compared to general data cases because this work requires the cooperation of medical experts. Medical experts still need to look at the data collection and attach links. Even though it is not necessary to attach links to many sample pairs, the medical experts should select the confusing sample pairs carefully and, this is a time-consuming task.

1.2.2 Semi-Supervised Learning for Bio-Medical Data Classification

Semi-supervised learning methods that efficiently utilize unlabeled samples for training a classifier with limited annotations have been reported. To improve the classification performance, the conventional semi-supervised learning methods use the estimation result from the model trained by small amounts of annotated data [19, 20] and other information such as prior knowledge of the target data and information related to the target task [21, 22].

In bio-medical image classification, differences between classes are often unclear. For example, in ulcerative colitis (UC) classification from endoscopic images, the UC classifier needs to focus on subtle features, such as visible vascular patterns and ulcers.

The issue of subtle differences between classes is not an issue in fully-supervised learning, where all samples are annotated with class information. Using given class labels, the UC classifier can automatically find discriminative features by contrastive learning and general classification loss, such as binary cross-entropy loss and hinge loss.

In contrast to fully-supervised learning, the issue of subtle differences between classes is a major problem in semi-supervised learning, where the amount of labeled samples is limited. When the majority of the sample set consists of samples without class labels, the UC classifier cannot find reasonable discriminative features by contrastive learning and general classification loss.

A possible approach is to use prior knowledge, which is various for each actual medical image classification task. An example of prior knowledge for UC classification tasks is that the severity of endoscopic images changes smoothly with their captured order, because endoscopic images are taken sequentially while endoscope moves inside the organs. More specifically, although they are not a video sequence, they are captured quasi-continuously enough to observe the smooth transition of the severity. Thus, if we utilize such cost-free prior knowledge on the training of the UC classifier, we can expect improved classification performance.

1.3 Purpose

1.3.1 Self and Soft-Constrained Clustering for Group-Based Labeling of Bio-Medical Data

As mentioned in Section 1.2.1, in bio-medical data clustering scenarios, we should handle two issues: the first issue is that the links for constrained clustering are not always useful, and the sec-

ond issue is that the cost of attaching links is high.

I propose a soft constrained clustering method suitable for bio-medical data clustering tasks to handle the first issue. The advantage of the proposed method is that it allows a violation of must-link. This property is useful when the links attached by medical experts are noisy or too hard to satisfy. In addition, to resolve the second issue, I propose a self and soft-constrained clustering method. Self-constraints are must-links, given automatically by defined as cost-free prior knowledge that temporally adjacent endoscopic images tend to belong to the same class. In other words, the self-constraints forces adjacent endoscopic images to be classified into the same class. At the same time, the self-constraints are treated as soft-constraints to allow the adjacent images to be different classes.

1.3.2 Semi-Supervised Learning with Disentangled Representation Learning Using Prior Knowledge

In semi-supervised learning for bio-medical data classification, it is often difficult to find reasonable discriminative features because the differences between classes are subtle. Prior knowledge related to each classification may help reduce this difficulty.

For example, in UC classification tasks, we can easily obtain two information, the location in a colon (e.g., left colon) and image capturing order. As mentioned in Section 1.2.2, image capturing order is effective to train an UC classifier. However, the problem is that temporally adjacent images also tend to belong to the same location class. Therefore, if learning using the image capturing order is directly applied to UC classification tasks, it may find discriminative features for location classification because the features for classifying UC images are subtle. Therefore, to effectively utilize image capturing order for learning the UC classifier, it is necessary to separate the location-dependent and UC-dependent features from endoscopic images.

I propose order-guided disentangled representation learning that utilizes disentangled representation learning and image capturing order of endoscopic images. Disentangled representation learning is a representation learning to divide various factors into a feature space [23–29]. The proposed method separates the location-dependent and UC-dependent features into the feature space by introducing the disentangled representation learning. In addition, to compensate for the lack of UC labels, the proposed method uses the image capturing order for learning the UC classifier.

1.4 Contributions

The contributions of this paper are follows:

- I propose a soft constrained clustering method suitable for bio-medical data clustering. It demonstrates the effectiveness of the proposed clustering method by endoscopic image clustering. The experiments are performed under several conditions, and in all cases, the proposed clustering method is superior. In addition, it proposes a self and soft-constrained clustering method that uses self-constraints as must-links, where self-constraints are defined as prior knowledge. The proposed method uses the self-constraints defined from the temporal continuity of endoscopic images and outperforms several state-of-the-art soft-constrained clustering methods in the experiment of endoscopic image clustering without annotation.
- I propose an order-guided disentangled representation learning method that uses disentangled representation learning and image capturing order of endoscopic images. It demonstrates the effectiveness of the proposed semi-supervised learning method for UC classification tasks. The experiment indicates that the proposed method outperforms the existing semi-supervised learning methods.

2. Self and Soft-Constrained Clustering for Group-Based Labeling of Bio-Medical Data

2.1 Background

Constructing a large-scale labeled image set is important for achieving sufficient performance with deep neural networks in image classification tasks and often requires a huge effort by annotators. In general object image classification tasks, image labeling is still tractable by using crowdsourcing services, such as Amazon Mechanical Turk, which employs a large number of annotators. However, such services cannot be utilized for labeling medical images, such as endoscopic images, because expert knowledge is required to label medical images. Therefore, constructing a medical image dataset is still intractable and is a serious bottleneck for medical image classification tasks.

The purpose of this thesis is to propose a novel clustering method for group-based labeling, which relaxes the labeling difficulty. In group-based labeling methods, we first gather similar images by clustering and then have an expert label the images of each cluster, as shown in Figure 3(a). If the annotator has a glance at the images of each cluster and finds that a certain cluster is comprised of the images from the same class, it is possible for him or her to give the same class label to all the images *at once.* (In the figure, a cluster is labeled as "Organ A" by this procedure.) Even if a cluster contains several outliers (i.e., images of the minor classes in the cluster), they can be easily found and discarded because their appearance is different from the inliers (i.e., the majority class samples in the cluster). As a result, by using group-based labeling, the labeling process can be dramatically accelerated, especially when the purity of each cluster is reasonably high.

To improve the purity of each cluster for efficient group-based labeling, constrained clustering is a promising approach. Different from typical unsupervised clustering methods, constrained clustering optimizes clusters while considering a limited number of constraints. Specifically, two types of constraint, called must-link and cannot-link, are given to image pairs, as shown in Figure 3(b). A pair of images with a must-link should belong to the same cluster, whereas a pair of images with a cannot-link should belong to different clusters. If medical experts pick a limited number of confusing image pairs and attach those links in advance, we can expect that the resulting clusters will have high purity.

However, for practical medical image clustering scenarios, the conventional constrained clustering methods should handle the following two issues. The first issue is that the links (i.e. constraints) are not always useful. Let us assume a clustering task

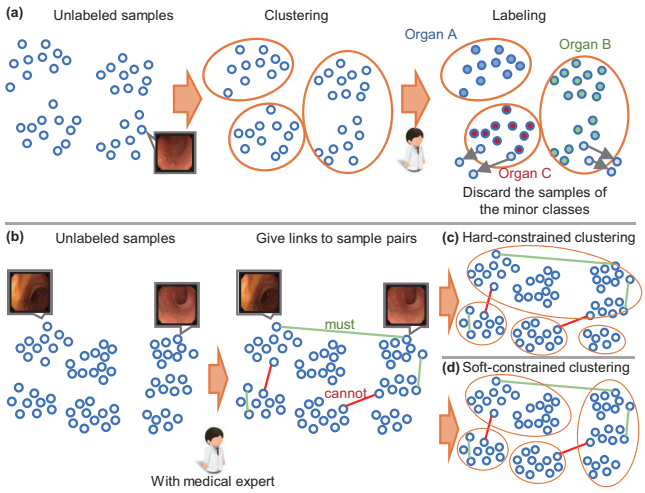


Fig. 3: (a) Group-based labeling by clustering. (b) Constraints given to sample pairs to control the clustering results. (c) Results of hard-constrained clustering. A must-link attached to the sample pair from the same organ leads too large a cluster because of the large difference in sample appearance. (d) The proposed soft-constrained clustering reduces the effect of the inappropriate constraints.

of endoscopic images, where each cluster should contain images of a single organ or component of an organ. In this task, a medical expert attaches a must-link to a pair of images of the same organ, *even though their appearances may be very different due to, for example, a difference in camera angle.* Since those images are very distant in the feature space, satisfying the must-link results in an unexpectedly large cluster with low purity, as shown in Figure 3(c).

The second issue is that attaching the links requires extra effort by medical experts. They still need to look at the image collection and attach links, especially to confusing image pairs. Even though it is not necessary to attach links to a large number of image pairs, the medical experts should select the confusing image pairs carefully and, this is a time-consuming task.

To resolve the first issue, I propose a novel soft-constrained clustering for medical image clustering tasks. The difference between the proposed method and conventional constrained clustering methods is that it allows the violation of must-link constraints to ignore the constraint between distant samples, as shown in Figure 3(d). Since the proposed soft-constrained clustering method is formulated as a single optimization problem, its solution is rather simpler than those of the conventional (hard) constrained clustering methods. To validate the effectiveness of the proposed soft-constrained clustering method, I collected the endoscopic image dataset from Kyoto Second Red Cross Hospital. In the experiment of the endoscopic images, the proposed soft-constrained clustering method outperformed several state-of-the-art soft-constrained clustering methods.

To handle the second issue, I propose a self and soft-constrained clustering method, where prior knowledge relevant to the target images is used as natural constraints. Specifically for classifying endoscopic images into organ classes, we can use the order in the image sequence as prior knowledge because consecutive images often belong to the same organ (e.g., esophagus, stomach), as shown in Figure 4(a). By putting a soft must-link

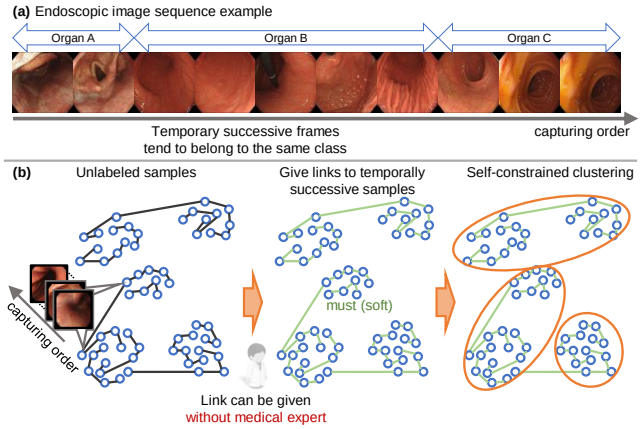


Fig. 4: (a) An endoscopic image sequence that captures parts of three organs. (b) The self and soft-constrained clustering wherein the order of the images in the sequence is used as soft must-links.

between consecutive images, we can expect to obtain high-purity clustering results without extra effort by experts, as shown in Figure 4(b).

I show that the self and soft-constrained clustering method, which utilizes the sequence-based constraints based on prior knowledge as must-link constraints, improves the purity of clusters. To validate the sequence-based constraints, I clustered the endoscopic images that were collected from Kyoto Second Red Cross Hospital. The proposed soft-constrained clustering method utilizing the sequence-based constraints improved clustering performance without label information, and the proposed method outperformed several soft-constrained clustering methods utilizing sequence-based constraints.

I demonstrated the effectiveness of the proposed method by applying it to group-based labeling for the construction of an organ-labeled dataset in endoscopy, which is very useful in practical clinical applications. In endoscopy, to make the examination reliable, the secondary reading is performed, where the images are checked by another doctor after the primary reading. However, checking all the images one by one requires a huge effort. If we can train an organ recognition model with the organ-labeled dataset, classifying all the images by this model allow that doctors can skip the step of determining the location where each image was taken. Moreover, by using this model, it is possible to construct a user interface for confirming whether the entire organ was taken by endoscopy. This is called deviation monitoring and is useful content for the education of non-experts in endoscopy. Therefore, the construction of an organ-labeled dataset is important for clinical diagnosis, and it is also important to reduce the effort of this construction with the proposed methods. In addition, the proposed method is the clustering method for performing group-based labeling with little effort, so it can be sufficiently applied not only to organ image tasks but also to other tasks.

The contributions of this work are summarized as follows:

- I propose a soft-constrained clustering method that is formulated as a single optimization problem. The method obtains a suitable solution even if the data distribution is multimodal because it does not use a hard-constrained clustering.
- The proposed soft-constrained clustering method outperformed several conventional methods in an endoscopic im-

age clustering task that included a small number of correct labels. This experiment was performed under several conditions, and in all cases, the proposed method was superior.
- I introduce temporal ordering information of the consecutive images as must-link constraints for soft-constrained clustering. We show that soft-constrained clustering of endoscopic images can be conducted without medical experts.
- I show that clustering performance improves when sequence-based constraints are used in the endoscopic image clustering without constraints derived from the correct labels.

2.2 Related Work

2.2.1 Constrained Clustering

The most famous constrained clustering method is COP-K-Means [30], which modifies the assignment step of ordinary K-means to satisfy must-link and cannot-link constraints. In addition, Shental *et al.* proposed a constrained clustering method based on a Gaussian mixture model [31]. In this method, the constraints are given as is-equivalent constraints and not-equivalent constraints, and the samples given is-equivalent (not-equivalent) constraints belong to the same (different) components. Li *et al.* devised a constrained spectral clustering method by developing a new embedding that introduces pairwise constraints to the spectral embedding [32]. They also proposed a constrained clustering method using the kernels that satisfy the pairwise constraints [33]. Recently, Le *et al.* proposed a binary optimization for constrained K-means (BOCK) in which the optimization problem is formulated as a single binary linear programming problem [34]. Unlike the algorithm in COP-K-Means, which updates the clustering result to satisfy the constraint, BOCK directly obtains a clustering result that satisfies the constraint. These methods perform hard-constrained clustering, so it is difficult for them to obtain reasonable clusters that satisfy the constraints especially when constraints are given between distant samples. Thus, they are not suitable for medical image clustering.

Soft-constrained clustering methods such as CVQE [35], and LCVQE [36] have been proposed to relax these constraints. For example, CVQE penalizes the violation of must-link constraints by adding a penalty based on the distance between the two nearest cluster centers of these two points. LCVQE reduces computational costs by modifying the penalty term in the objective function of CVQE. Ares *et al.* [37] proposed a method that extends the batch K-means method to a soft-constrained clustering method. Similar to these methods, the PCK-means [38], and MPCK-means [39] methods design the penalty function $0-1$ Loss that imposes a constant value on the objective function as a penalty when the constraint is violated. These soft-constrained clustering algorithms first solve the hard-constrained clustering problem and then modify the cluster assignment of each sample. This two-step optimization can eliminate a small number of erroneous constraints, but it cannot deal with multimodal distributions.

Recently, novel frameworks of constrained clustering by using deep learning techniques have been proposed. In particular, Hsu *et al.* proposed a neural-network-based end-to-end clustering framework for pairwise constraints [40]. In this framework, the neural network outputs the cluster probabilities of the input sample and learns to reduce (increase) the Kullback-Leibler divergence of the cluster probabilities between similar (dissimilar) samples. Zhang *et al.* proposed a framework for constrained clustering that optimizes neural networks with an objective function that equalizes the cluster probabilities of constrained samples [41]. Moreover, Fogel *et al.* proposed CPAC, a framework for constrained clustering using a neural network [42]. CPAC simultaneously optimizes the ordinary objective function of autoencoders and the objective function that reduces the distance of the constrained samples in the latent space obtained from autoencoders. Although these methods improve clustering performance, they assume hard-constrained clustering and are not suitable for medical image clustering tasks where constraints are imposed between distant samples.

The proposed method was formulated as mixed-integer linear programming to perform soft-constrained clustering. Therefore, it directly obtains a clustering result that satisfies the constraint and is suitable for medical image clustering tasks.

2.2.2 Group-Based Labeling

Group-based labeling is an efficient strategy to collect a sufficient number of labeled images for training machine-learning-based methods, and several group-based labeling methods have been reported [9, 10, 12–18]. Here, we can accelerate the labeling process if it is possible to gather images with similar appearances into a cluster based on some criteria. For example, Wigness *et al.* proposed hierarchical cluster guided labeling (HCGL) [9, 10]. In this method, unlabeled samples are first clustered using a hierarchical clustering method; then, groups of samples are repeatedly selected for labeling. Mousavi *et al.* proposed a system for speeding up the labeling of large collections of unlabeled images [14]. In this system, unlabeled samples are mapped to numerical feature embeddings and clustered. The clustered samples are then labeled by a domain expert, and the labeled samples are used to train the classifier, which predicts labels for new unlabeled samples. Galleguillos *et al.* proposed a framework for speeding up object labeling of unlabeled images [13]. In this framework, images are partitioned into multiple segments and then clustered for labeling. In addition, Biswas *et al.* proposed an algorithm to efficiently select sample pairs for constraints in constrained clustering [12].

These methods assume that clustering is performed first in order to obtain a cluster with high purity. Therefore, the idea of obtaining high-purity clusters by imposing constraints such as in the proposed method is useful in any of the above cases.

2.2.3 Bio-Medical Data Annotation with Crowdsourcing Services

Several studies have been reported that validated the effectiveness of constructing annotated medical data datasets using crowdsourcing services [43–45]. In [43], they investigated the effectiveness of dataset construction for endoscopic image registration tasks using crowdsourcing services. They reported that the model trained with the dataset constructed using the crowdsourcing service achieved performance close to the performance when trained with the dataset constructed by medical experts. Kim *et al.* re-

ported the effectiveness of crowdsourcing services for dataset construction for surgical tool detection tasks in cataract surgery videos [44]. Moreover, in [45], they reported the validity of airway region annotation in the chest in the computed tomography (CT) images by crowdsourcing services. They showed that some annotated samples of the datasets acquired by crowdsourcing services were incorrect annotations. However, when incorrect annotation samples were excluded, it was shown that there was a moderate correlation with the annotation results by medical experts.

From these studies, it is shown that the crowdsourcing service is effective for annotation that does not require specialized knowledge such as image registration and surgical tool detection. In contrast, if it involves expert knowledge such as airway area annotation, it may not be possible to construct an accurate data set with a crowdsourcing service. In contrast, when using crowdsourcing services for annotations that require expert knowledge, such as airway region annotations, the obtained dataset may not be valid. The purpose of this thesis is to improve the efficiency of annotation that involves such expert knowledge, and I propose constrained clustering for group-based labeling that can reduce the effort by annotators.

2.3 Soft-Constrained Clustering

To explain the problem setting of soft-constrained clustering, I will start with a description of the standard K-Means clustering with a binary optimization approach. Then, I will describe the constrained clustering task and extend the standard K-means clustering method to the soft-constrained clustering method. In addition, I explain the actual optimization for the proposed soft-constrained clustering method with ordinary constraints.

2.3.1 Problem Setting

Given a set of samples $\mathcal{X} = \{\boldsymbol{x}_j \in \mathbb{R}^D\}_{j=1}^N$, where N and D are the number of the samples and the dimension of the feature vector, respectively, the aim of K-Means is to find K cluster centroids and to assign each sample $\boldsymbol{x}_j$ to a cluster. This clustering algorithm minimizes the within-cluster sum of squares (WCSS). The objective function is formulated as:

$$\min_{\mathbf{S},\mathbf{C}} |\mathbf{X} - \mathbf{CS}|_F^2, \tag{1}$$
$$\text{s.t.} \quad s_{i,j} \in \{0,1\}, \quad \forall i, j,$$
$$\sum_{i=1}^{K} s_{i,j} = 1, \quad \forall j = 1, \ldots, N,$$

where $\mathbf{X} \in \mathbb{R}^{D\times N}$ is a matrix whose j-th column is $\boldsymbol{x}_j \in \mathbb{R}^D$. $\mathbf{C} \in \mathbb{R}^{D\times K}$ is a matrix whose i-th column corresponds to the cluster centroid $\boldsymbol{c}_i \in \mathbb{R}^D$, and $\mathbf{S} \in \mathbb{R}^{K\times N}$ is a cluster assignment matrix, where $\boldsymbol{x}_j$ is assigned to the i-th cluster when the value of the (i, j)-th element $s_{i,j}$ is 1, and 0 otherwise. The first constraint ensures that each cluster assignment $s_{i,j}$ has 1 or 0, and the second constraint ensures that each sample is assigned to only one cluster. The j-th column of $\mathbf{CS}$ is the centroid of the cluster to which the j-th sample $\boldsymbol{x}_j$ is assigned. The operation $|\cdot|_F$ denotes the Frobenius norm.

In the constrained clustering task, we obtain a clustering result that satisfies constraints called must-links and cannot-links. These constraints are attached to a sample pair and define the relationships between the samples. Sample pairs with a must-link constraint should belong to the same cluster. In contrast, the pairs with a cannot-link constraint cannot belong to the same cluster. In general, these constraints are derived from a small number of samples that have been correctly labeled by experts. Denoting $\mathcal{L}_i$ as the set of labeled samples for class i, $\boldsymbol{x}_j \in \mathcal{L}_i$ indicates the j-th sample belongs to the i-th class. Using the labeled samples, we register a sample pair in the must-link set $\mathcal{M}$ if its pair belongs to the same class, and we register it in the cannot-link set $\mathcal{D}$ if the samples of the pair belong to different classes.

The proposed soft-constrained clustering method also acquires the cluster centroids and cluster assignments that minimize the WCSS distortion. The difference from the general constrained clustering is that the must-link set $\mathcal{M}$ behaves as penalties. In contrast, the cannot-link set $\mathcal{D}$ is still employed as hard constraints, because satisfying cannot-link constraints always has a positive effect on the clustering results.

The objective function for the proposed method is formulated as:

$$\min_{\mathbf{S},\mathbf{C}} |\mathbf{X} - \mathbf{CS}|_F^2 + \omega \sum_{(\boldsymbol{x}_p,\boldsymbol{x}_q)\in\mathcal{M}} \sum_{i=1}^{K} |s_{i,p} - s_{i,q}|, \tag{2}$$
$$\text{s.t.} \quad s_{i,j} \in \{0,1\}, \quad \forall i, j,$$
$$\sum_{i=1}^{K} s_{i,j} = 1, \quad \forall j = 1, \ldots, N,$$
$$s_{i,p} + s_{i,q} \le 1, \quad \forall i = 1, \ldots, K, \quad \forall (\boldsymbol{x}_p, \boldsymbol{x}_q) \in \mathcal{D}, \tag{3}$$

where the second term of (2) represents the soft-constraints provided by the must-links. When two samples $\boldsymbol{x}_p$ and $\boldsymbol{x}_q$ registered in the must-link set $\mathcal{M}$ are assigned to the different clusters, it becomes $|s_{i,p} - s_{i,q}| = 1$; thus, it penalizes the objective function with a constant value $\omega \in \mathbb{R}_+$. The inequality constraints (3) represent the cannot-link constraints. When two samples $\boldsymbol{x}_p$ and $\boldsymbol{x}_q$ registered in the cannot-link set $\mathcal{D}$ are assigned to the same cluster i, the solution is not allowed because $s_{i,p} + s_{i,q} = 1 + 1 \not\le 1$, which violates the constraint. The proposed method obtains the optimal clustering result under such constraints and penalties.

2.3.2 Optimization for Soft-Constrained Clustering

To minimize the objective function (2), I take an optimization approach that alternately updates the cluster centroid matrix $\mathbf{C}$ and the assignment matrix $\mathbf{S}$ until convergence. The approach is similar to the original K-means optimization approach (EM algorithm).

In the update step for the cluster centroid matrix $\mathbf{C}$, the updated $\mathbf{C}$ is computed from $\mathbf{X}$ and a fixed $\mathbf{S}$. Therefore, we can ignore the constraints and the penalty term of the objective function. We update $\mathbf{C}$ by solving a regularized least squares problem to avoid numerical issues with large-scale problems [46]:

$$\min_{\mathbf{C}} |\mathbf{X} - \mathbf{CS}|_F^2 + \lambda |\mathbf{C}|_F^2, \tag{4}$$

where λ is the regularization parameter. In the experiments described later, I set λ to 10^{-4}. The problem is a convex quadratic optimization and can be solved in a closed form:

$$\mathbf{C} = \mathbf{X}\mathbf{S}^{\mathrm{T}}(\mathbf{S}\mathbf{S}^{\mathrm{T}} + \lambda\mathbf{I})^{-1}, \tag{5}$$

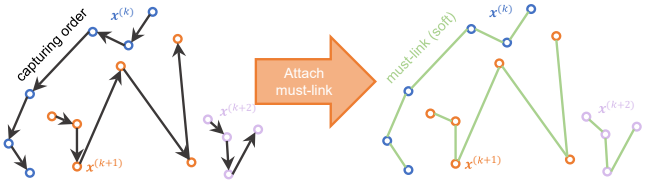


Fig. 5: Outline when attaching must-link to temporally adjacent sample pairs. The circle represents the sample in the sequence, and the circle color indicates each sequence. The arrows indicate the capturing order of the sequence, and the sample pair connected by the green line is the sample pair with self-constraint.

where $\mathbf{I}$ is a $K \times K$ identity matrix, and $(\mathbf{S}\mathbf{S}^{\mathrm{T}} + \lambda\mathbf{I})^{-1}$ is guaranteed to be full-rank for $\lambda > 0$.

Next, we update the cluster assignment matrix $\mathbf{S}$ with $\mathbf{C}$ fixed. Let $\mathbf{Y} \in \mathbb{R}^{K\times N}$ be a matrix whose $(i, j)-$th element y_{ij} is the squared distance between $\boldsymbol{x}_j$ and $\boldsymbol{c}_i$, namely, $y_{i,j} = |\boldsymbol{c}_i - \boldsymbol{x}_j|_2^2$. By using $\mathbf{Y}$, the first term of (2) can be rewritten as $\langle \mathbf{Y}, \mathbf{S}\rangle_F$, which represents the Frobenius inner product of $\mathbf{Y}$ and $\mathbf{S}$. In addition, the second term of (2) is rewritten with a set $\boldsymbol{\gamma} = \{\gamma_{i,(p,q)}\}$ of $K\times M$ variables, which is defined for each must-link pair $(\boldsymbol{x}_p, \boldsymbol{x}_q) \in \mathcal{M}$, where M is the number of must-link pairs. Consequently, the updating problem for $\mathbf{S}$ is:

$$\min_{\mathbf{S},\gamma} \langle \mathbf{Y}, \mathbf{S}\rangle_F + \omega \sum_{(\boldsymbol{x}_p,\boldsymbol{x}_q)\in\mathcal{M}} \sum_{i=1}^{K} \gamma_{i,(p,q)}, \tag{6}$$

$$\text{s.t.} \quad s_{i,j} \in \{0, 1\}, \quad \forall i, j,$$

$$\sum_{i=1}^{K} s_{i,j} = 1, \quad \forall j = 1, \ldots, N,$$

$$s_{i,p} + s_{i,q} \le 1, \quad \forall i = 1, \ldots, K, \quad \forall (\boldsymbol{x}_p, \boldsymbol{x}_q) \in \mathcal{D},$$

$$s_{i,p} - s_{i,q} \le \gamma_{i,(p,q)}, \quad \forall i, \quad \forall (\boldsymbol{x}_p, \boldsymbol{x}_q) \in \mathcal{M},$$

$$-s_{i,p} + s_{i,q} \le \gamma_{i,(p,q)}, \quad \forall i, \quad \forall (\boldsymbol{x}_p, \boldsymbol{x}_q) \in \mathcal{M},$$

where $\boldsymbol{\gamma} = \{\gamma_{i,(p,q)}\}$. This problem is a mixed-integer linear programming problem, i.e., a combination of binary programming for $\mathbf{S}$ and linear programming for $\boldsymbol{\gamma}$. The optimal solution is obtained by using the branch-and-bound optimization technique. The above steps for updating $\mathbf{C}$ and $\mathbf{S}$ are performed alternately until convergence. In the experiments, the initial $\mathbf{C}$ and $\mathbf{S}$ for the proposed method were set based on the standard K-Means clustering for faster convergence.

2.4 Self and Soft-Constrained Clustering

As mentioned in Section 2.1, in soft-constrained clustering of sequence images, such as endoscopic images, soft constraints can be automatically obtained by using prior knowledge, which is that temporally successive images tend to belong to the same class. Specifically, I generate soft constraints between pairs of successive images based on this knowledge.

Given a set of N_s sequences $\mathcal{Z} = \{\mathbf{X}^{(k)} \in \mathbb{R}^{D\times n_k}\}_{k=1}^{N_s}$, where $\mathbf{X}^{(k)}$ is the k-th sequence and n_k is the number of samples in the k-th sequence, I generate a must-link constraint set $\mathcal{T}$ based on the temporal ordering information. Each matrix $\mathbf{X}^{(k)}$ represents a sequence of samples whose j-th column is the j-th sample in a sequence $\boldsymbol{x}_j^{(k)} \in \mathbb{R}^D$. The temporally adjacent sample pairs (i.e., $(\boldsymbol{x}_j^{(k)}, \boldsymbol{x}_{j+1}^{(k)})$) are registered in the must-link constraint set $\mathcal{T}$ as shown in Figure 5. Specifically, when $\mathcal{Z}$ is given, the number of the constraints is $\sum_{k=1}^{N_s}(n_k - 1)$.

In the self and soft-constrained clustering method, the objective function is a formula (2) where the must-link set $\mathcal{M}$ is replaced by $\mathcal{T}$. The optimization manner is the same as that of the proposed soft-constrained clustering method.

2.5 Clustering Experiments

I performed two experiments on the proposed methods described in Sections 2.3 and 2.4. They are overviewed in Table 1. Since the proposed methods have different purposes and constraints, the experiments used different datasets and metrics. The datasets for the experiments were collected from Kyoto Second Red Cross Hospital. The patients were told the aim of the study and provided written informed consent before participating in the trial. The experiments were approved by the Kyoto Second Red Cross Hospital Ethics Committee.

2.6 Evaluation of Soft-Constrained Clustering

2.6.1 Dataset

To evaluate the proposed soft-constrained clustering method, in which a small set of labeled images was given, I conducted experiments using a large-scale endoscopic image dataset. This dataset is comprised of 11,599 stomach images collected from Kyoto Second Red Cross Hospital. For the performance evaluation, the true class label was attached to each image by expert endoscopists. The number of classes were 20, as listed in Table 2. The classes were defined on the basis of the traditional anatomical classification used by endoscopists and by camera angle (look-down and look-up). To make the feature vector, DenseNet169 [47] pre-trained by ImageNet [48] was used to extract 1,664-dimensional feature values from the original RGB image (224×224 pixels). The reason for adopting DenseNet is that this model is widely known as one of the high-performance models and a model pre-trained by ImageNet is distributed within the world.

2.6.2 Experimental Conditions

In group-based labeling, it is important to collect enough labels not only for specific classes, but also for every class. In an image set that consists of imbalanced numbers of class samples, if we simply label the samples belonging to a prominent class for a cluster and discard the others, some classes may not be labeled (e.g., samples in a large class tend to be a prominent class and those in a small class tend to be discarded). To collect labels for every class, we first find the prominent cluster for each class. A prominent cluster contains the most labeled samples of a specific class among all of the clusters, and the prominent cluster is shown to an expert. The expert attaches the same class label to them at once if the cluster only contains samples from the same class. Even if the cluster contains several samples from different classes, the expert can do the same after discarding those samples. This process is repeated for every class. The remaining samples (the discarded samples and the samples in other clusters) are fed to the clustering in the next round. The overall process is repeated until enough samples have been labeled. One way to verify whether a sufficient number of labeled samples has been

Table 1: Outline of the experiments. The target method, evaluation index, and data set of each experiment are shown.

Experiment	Purpose	Method	Metric	Dataset
Section 2.6	Obtain high-purity cluster of each class	Soft-constrained	Prominent-purity Prominent-recall Purity	Stomach dataset (Class listed in Table 2)
Section 2.7	Improve purity of all clusters	Self and soft-constrained	Purity	EGD dataset Colon dataset (Class listed in Table 3)

Table 2: Twenty classes defined for stomach endoscopic images. BD, UP, MD, LO, LD, and LU stand for (stomach) body, upper, middle, lower, (camera) look-down, and look-up, respectively.

Fundus	Fundus on UP BD/ LU	UP BD/ LU
UP BD/ LD	UP-MD BD/LU	UP-MD BD/ LD
MD BD/ LU	MD BD/ LD	MD-LO BD/ LU
MD-LO BD/ LD	LO BD/LD	Angular Incisure LO BD/ LD
AntralZone on LO BD/ LD	Angular Incisure	Angular Incisure-Antral Zone
Antral Zone	Pyloric Antral	Pyloric Zone
Pylorus	Junction	

Table 3: List of classes for the datasets used in Section 2.7.

Dataset	Class
EGD dataset	Esophagus
	Stomach
	Duodenum
Colon dataset	Right colon
	Rectum
	Left colon

obtained is to train a classifier with the labeled samples and investigate the test performance of the trained classifier.

In light of the above methodology, I decided to focus on the prominent clusters in evaluation of the proposed method. To evaluate whether the clustering method has obtained excellent prominent clusters, I designed new performance metrics prominent-purity and prominent-recall. Prominent-purity shows the percentage of samples that can be given the correct label in prominent-cluster by one labeling operation and is the ratio of the number of true positives to the number of samples in the prominent cluster, which is the cluster that contains the most labeled samples of a certain class. Prominent-recall shows the percentage of samples that can be given the correct label in the entire data by one labeling operation and is the ratio of the number of true positives in the prominent cluster to the total number of samples of a certain class in the dataset. Therefore, these metrics can be used to confirm the effectiveness of the clustering method for the efficiency of group-based labeling. For a fair comparison with unsupervised clustering (K-means), these metrics were calculated, excluding the samples used for constraints. The (traditional) purity is the sum of the number of samples of the majority class in each cluster divided by the total number of samples.

The proposed method was compared with several conventional clustering methods, such as K-means (KM), the binary optimization approach for constrained K-means (BOCK) [34], metric-based pairwise constrained K-means (MPCK) [39], and linear-time constrained vector quantization error (LCVQE) [36]. KM is an unconstrained clustering, and BOCK is a hard-constrained clustering. MPCK and LCVQE are state-of-the-art soft-constrained clustering methods.

The evaluation examined the change in the purity of the cluster while varying the number of labeled samples. Since the samples in the experiment had been labeled by medical experts, I used the labels for R percent of the samples to give the constraints and used the remaining samples to compute the purity of the clustering results. Specifically, first, R percent of samples from the dataset were randomly picked. Then, to generate the must-link set, one anchor sample was randomly picked for each class, and the must-link constraints were put between the anchor sample and the remaining samples for each class. In addition, to generate the cannot-link set, one sample was randomly picked for each class, and the cannot-link constraints were attached to all combinations of the picked samples.

For an ablation study, I performed the proposed method without the must-link constraints. In this experiment, the parameter ω of the penalty terms was set to 50. To confirm the robustness for the ratio of the labeled samples to the total samples, R, and the number of clusters K, I conducted the clustering while varying K and R ($K = 50,\ 100$ and $R = 1\%,\ 3\%,\ 5\%$). The proposed method was implemented in MATLAB, and the experiment was run on a computer with the Intel Xeon E5-2620 CPU and 256GB of memory.

2.6.3 Clustering Results

Table 4 shows the results of the quantitative performance evaluation. The proposed method achieved the best performance under all conditions. Except for the case of $K = 50$, $R = 1\%$, it achieved over 0.7 in the prominent-purity. This result indicated that 70% of the samples in the prominent-clusters obtained by the proposed method could be given the correct label. Therefore, in the actual labeling operation for the prominent-cluster obtained from the proposed method, the annotator can complete one labeling operation by removing 30% of the samples in the cluster and labeling the remaining samples once. The removal operation is relatively easy because these samples are a minority in the cluster and are easy to find. From this result, it can be confirmed that the group-based labeling by the proposed method is more efficient than the general labeling process in which each sample is labeled individually, and it is also more efficient than the group-based labeling by the comparative method. This level of prominent-purity is high enough to reduce the time required for the labeling task. In particular, when $K = 100$, and $R = 1\%$, the prominent-purity of

Table 4: Quantitative performance evaluation for constrained clustering with prominent-purity and prominent-recall.

Metric	Conditions	KM	BOCK	MPCK	LCVQE	Proposed w/o must link	Proposed
Prominent-purity	K=100, R= 1%	0.602	0.627	0.421	0.573	0.658	**0.715**
	K=100, R= 3%	0.726	0.673	0.544	0.660	0.712	**0.789**
	K=100, R= 5%	0.681	0.716	0.494	0.587	**0.747**	**0.747**
	K=50, R= 1%	0.623	0.549	0.369	0.616	0.517	**0.660**
	K=50, R= 3%	0.668	0.648	0.521	0.636	0.651	**0.727**
	K=50, R= 5%	0.637	0.694	0.498	0.643	0.662	**0.702**
Prominent-recall	K=100, R= 1%	0.132	0.138	0.109	0.122	0.147	**0.159**
	K=100, R= 3%	0.156	0.167	0.147	0.155	0.158	**0.179**
	K=100, R= 5%	0.162	0.169	0.170	0.159	**0.174**	**0.174**
	K=50, R= 1%	0.254	0.230	0.170	0.230	0.218	**0.270**
	K=50, R= 3%	0.287	0.281	0.263	0.269	0.273	**0.310**
	K=50, R= 5%	0.262	0.296	0.258	0.279	0.295	**0.297**

the proposed method shows an 8% improvement over the second-best, BOCK. Moreover, it achieved the best prominent-recall under all conditions. In the situation of labeling the stomach dataset with a group-based labeling strategy, when $K = 100$, and $R = 1\%$, the labeling utilizing the proposed method can obtain about 240 more labeled images than the second-best method, BOCK. This difference affects the efficiency of the actual labeling operation. Table 5 shows the results of the purity evaluation. Since the proposed method was superior to KM under all conditions, its soft-constrained clustering worked effectively to improve the purity of the clusters. The proposed method was the only method that improved KM under all conditions. It showed relatively good performance although several conventional methods outperformed it under some conditions. However, in the evaluation focusing on prominent clusters, the proposed method achieved the best performance under all conditions, so it does not become a problem in the labeling process considered in this work. I confirmed significant differences between the proposed method and comparative methods in most cases according to paired McNemar's test that was corrected based on the Holm method, performed at a significance level of 0.05. When $K = 100$, significant differences were not confirmed in some cases, but as mentioned above, the proposed method is effective because the number of labeled samples increases in the assumed labeling process that gives a label to a prominent cluster.

The average computational time for clustering over 10,000 images was $14s$ for KM, $1{,}119s$ for BOCK, $10s$ for MPCK, $1{,}824s$ for LCVQE, and $130s$ for the proposed method. The proposed method was much faster than the state-of-the-art of the constrained clustering method and soft-constrained method. The computational cost of the proposed method is thus low enough for group-based labeling.

Figure 6 shows the clustering results when K and R were set to 50 and 0.03, respectively. Figs. 6(b) and (c) show the prominent Fundus clusters, which contain the most Fundus images obtained by the proposed method and BOCK. In the scatter plot, green and magenta dots indicate true positives and false-positives for Fundus. The plot confirms that the samples in the cluster obtained from the proposed method were closer to each other than the samples in the cluster obtained from BOCK because the proposed method allows the violation of the must-link constraints. The images in Figure 6(c) show the false-positive images in the prominent Fundus cluster obtained from BOCK. The prominent-cluster contained these false-positive images that were very similar to Fundus images because BOCK does not assume that the sample distribution for a class is multimodal. In contrast, the proposed method assumes that the sample distribution of a class is multimodal and can reduce the effect of the inappropriate constraints. Therefore, the proposed method obtained the prominent Fundus cluster with fewer false-positive images than the BOCK result, as shown in Figure 6(b).

2.7 Evaluation of Self and Soft-Constrained Clustering

2.7.1 Dataset

I conducted experiments using two large-scale endoscopic image datasets to evaluate the self and soft-constrained clustering method. The first dataset (hereinafter, referred to as the esophagogastroduodenoscopy (EGD) dataset) is comprised of 500 EGD image sequences collected from Kyoto Second Red Cross Hospital, and the total number of images is 15,394. The second dataset (hereinafter, referred to as the colon dataset) is comprised of 388 colonoscopy image sequences collected from Kyoto Second Red Cross Hospital, and the total number of samples is 10,265. Each image sequence was taken in a single examination. Therefore, the EGD dataset contains 500 examination results, and the colon dataset contains 333 examination results.

For the performance evaluation, I had expert endoscopists label the correct class of each image. The classes of these datasets are listed in Table 3. To make the feature vector, DenseNet201 [47] pre-trained by ImageNet was used to extract 1,920-dimensional feature values from the original RGB image (224×224 pixels). The stomach dataset used in Section 2.6 was not used in this experiment because it does not have temporal ordering information.

2.7.2 Experimental Conditions

I evaluated the purity of the clusters obtained by the proposed method. Since the dataset in this experiment has temporal ordering information, I imposed a sequence-based constraint between all temporally adjacent samples. Unlike the experiment in Section 2.6, I did not use true labels to provide the constraints; I used all of the samples to compute the purity of the clusters.

In contrast to soft-constrained clustering in which a small number of class labels were given, the self and soft-constrained clustering did not have any ground truth for the class labels, because the constraints were generated from the temporal ordering information without any labeling. Thus, I could not estimate the prominent cluster for each class in the group-based labeling. In

Table 5: Quantitative performance evaluation using purity to confirm the validity of the constraints. If the value of the comparative method is marked with an asterisk, there is significant difference between the proposed method and that method.

Metric	Conditions	KM	BOCK	MPCK	LCVQE	Proposed w/o must link	Proposed
Purity	K=100, R= 1%	0.674*	**0.723***	0.663*	0.698	0.708	0.707
	K=100, R= 3%	0.712*	0.709*	0.647*	0.698*	0.717	**0.724**
	K=100, R= 5%	0.683*	0.724	0.661*	0.715	0.705*	**0.727**
	K=50, R= 1%	0.716*	0.686*	0.606*	**0.732***	0.708*	0.718
	K=50, R= 3%	0.738*	0.684*	0.613*	0.712*	0.715*	**0.755**
	K=50, R= 5%	0.696*	**0.732***	0.603*	0.729*	0.704*	0.718

*: $p < 0.05$

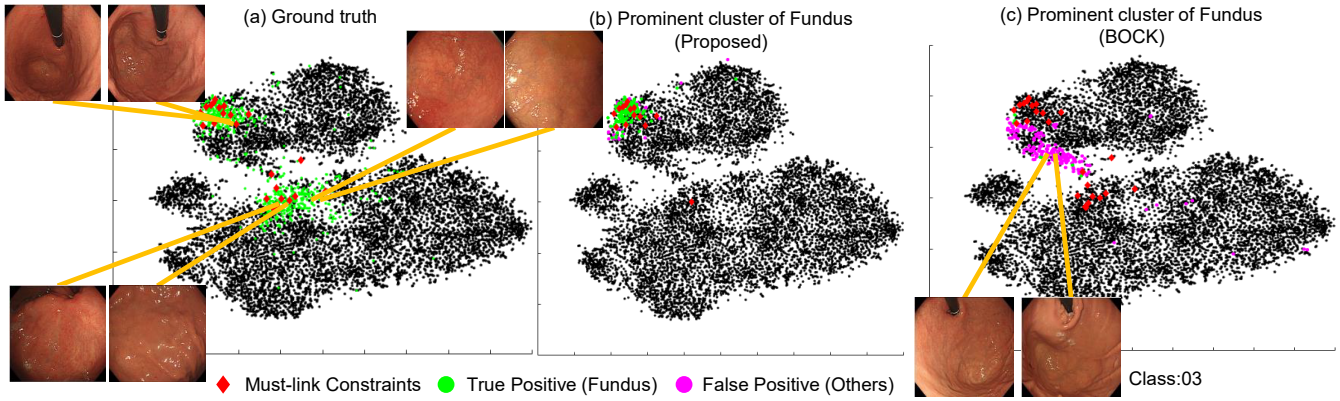


Fig. 6: Clustering result for $K = 50$ and $R = 0.03$. I used t-SNE [49] for these two-dimensional visualizations. (a) Distribution of samples whose true label is "Fundus". Green dots indicate unlabeled samples, and red ones indicate labeled samples that were used as must-link constraints. (b) Prominent cluster of Fundus generated by the proposed method and (c) prominent cluster of Fundus generated by BOCK [34]. In (b) and (c), magenta dots indicate samples from other classes. The images in (a), and (c) are Fundus image example, and non-Fundus image example, respectively.

this case, I consider that it is better if the purity of every cluster is high. Therefore, I evaluated the (traditional) purity instead of prominent-purity.

I compared the proposed method with the same ones described in Section 2.6 except for BOCK (hard-constraints) since if I simply generated hard constraints from the temporal ordering information, all images in a sequence would belong to the same cluster, and this is obviously wrong. In addition, to investigate the influence of the parameter ω, I varied ω from 1 to 30. The proposed method was implemented in MATLAB and, the experiment was run on a computer with the Intel Core i9-9980XE and 64GB of memory.

2.7.3 Clustering Results

Table 6 shows the purity for each method for two value of K. To show the robustness of the hyper-parameter ω, the table lists the performance of the worst and best cases when I changed ω. The proposed method using the best ω outperformed the other methods; even when I used the worst ω, the purity of the proposed method was slightly better than those of the other methods. In contrast, the conventional soft-constrained clustering methods were worse than KM. These conventional methods adversely affected purity because they assume that some improper constraints are eliminated, and the samples are clustered. In contrast, the proposed method worked well under these constraints because it assumed that the data distribution is often multimodal and the distant samples are given the constraint. In the situation of labeling the EGD dataset with a group-based labeling strategy, when $K = 30$, the labeling utilizing the proposed method can obtain about 266 more labeled images than the second-best, KM. This result shows that by using the proposed method, the correct labels can be given to about 2% of the images in the dataset in one labeling operation without additional annotation. Therefore, the proposed method is more effective in accelerating the labeling process than the comparison method. Needless to say, the cost of this labeling process is low compared to the general labeling process. In addition, I confirmed significant differences between the proposed method and the comparative methods according to paired McNemar's test, which was corrected based on the Holm method, performed at a significance level of 0.05.

Figure 7 shows the cluster assignment results of KM and the proposed method with $\omega = 10, 20, 30$, together with the ground-truth sequence. In this figure, the vertical partitions indicate the timing when the assigned cluster changes. For example, in the results of the proposed method with $\omega = 10$, the initial two images assigned to the same cluster are separated by a black line from the next two images, which are assigned to a different cluster. It can be seen that the number of adjacent images belonging to the same cluster increases with ω, and the proposed method preferentially assign samples with similar appearances, such as the third and fourth samples, to the same cluster. As a result, the purity of each sequence cluster improved.

Figure 8 is a quantitative evaluation showing the effect of changing the parameter ω. Here, the proposed method outperformed KM under all conditions, even though the constraints were generated in an unsupervised manner from the temporal ordering information. In particular, in Figure 8(a), the proposed method with $K = 15$, $5 \leq \omega$ outperformed KM with $K = 30$. Moreover, in Figure 8(b), the proposed method with

Table 6: Quantitative performance evaluation of self and soft-constrained clustering utilizing sequence-based constraints, as measured by purity. If the value of the comparative method is marked with an asterisk, there is a significant difference between the proposed method and that method.

Dataset	Conditions	KM	MPCK	LCVQE	Proposed w/o worst ω	Proposed best ω
EGD dataset	K=30	0.8645*	0.6399*	0.8484*	0.8660*	**0.8818**
	K=15	0.8583*	0.6399*	0.8303*	0.8597*	**0.8770**
Colon dataset	K=30	0.6967*	0.4334*	0.6728*	0.6996*	**0.7164**
	K=15	0.6819*	0.4341*	0.6760*	0.6831*	**0.6940**

*: $p < 0.05$

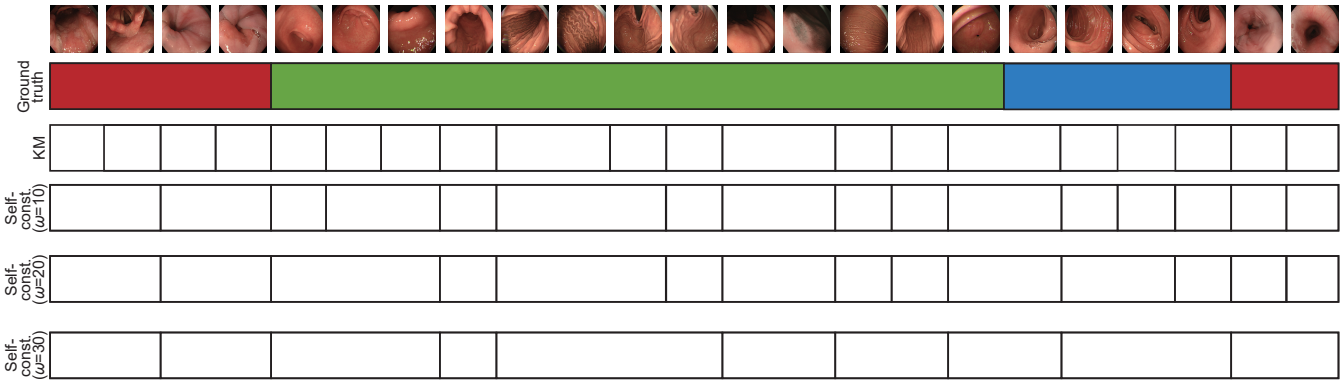
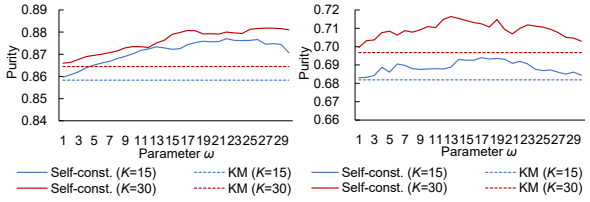


Fig. 7: Example of the cluster switching point determined by the proposed method when the parameter ω changed in the experiment on the EGD dataset. The color of the class label bar indicates the type of class label. The vertical black lines in the bar of each method indicate cluster label switching points.



(a) EGD dataset result **(b)** Colon dataset result

Fig. 8: Purity of the proposed method versus ω.

$K = 15$, $13 \leq \omega \leq 20$ approached the performance of KM when $K = 30$. These results show that setting the parameter ω properly can significantly improve clustering performance.

3. Semi-Supervised Learning Using Prior Knowledge of Bio-Medical Data

3.1 Background

In the classification of ulcerative colitis (UC) using deep neural networks, where endoscopic images are classified into the lesion and normal classes, it is difficult to collect a sufficient number of labeled images because the annotation requires significant effort by medical experts. UC is an inflammatory bowel disease that causes inflammation and ulcers in the colon. Specialist knowledge is required to annotate UC because texture features, such as bleeding, visible vascular patterns, and ulcers, should be captured among the image appearances that drastically vary depending on location in the colon to detect UC.

Semi-supervised learning methods [11, 19, 20, 50] have been used to train classifiers based on a limited number of labeled images, involving the use of both labeled and unlabeled images. If a classifier with a moderate classification performance is obtained with few labeled data, the performance of a classifier can be further improved by applying these semi-supervised learning methods. However, existing semi-supervised learning methods do not show satisfactory performance for UC classification because they implicitly assume that the major appearance of images is determined by the classification target class, whereas the major appearance of UC images is determined by the location in the colon, not by the disease condition.

Incorporating domain-dependent knowledge can also compensate for the lack of labeled data. In endoscopic images, we can utilize two types of prior knowledge: location information and temporal ordering information, that is, the order in which the endoscopic images were captured. Location information can be obtained easily by tracking the movement of the endoscope during the examination [51, 52], with the rough appearance of endoscopic images characterized by their location. Endoscopic images are acquired in sequence while the endoscope is moved through the colon. Therefore, the temporal ordering information is readily available, and temporally adjacent images tend to belong to the same UC label. If the above information can be incorporated into semi-supervised learning, more accurate and reliable networks for UC classification can be developed.

In this study, I propose a semi-supervised learning method for UC classification that utilizes location and temporal ordering information obtained from endoscopic images. Figure 9 shows the underlying concept for the proposed method. In the proposed method, a UC classifier is trained with incomplete UC labels, whereas the location and ordering information are available. By utilizing the location information, I aim to improve UC classification performance by simultaneously extracting the UC and location features from endoscopic images. I introduce disentangled representation learning [23, 24] to effectively embed the UC and location features into the feature space separately. To compensate

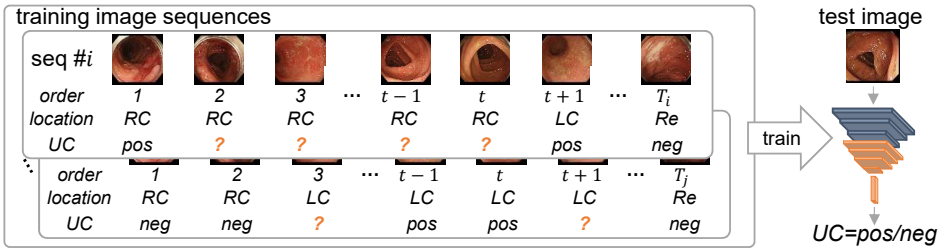


Fig. 9: Underlying concept for the proposed method. The objective of the study is to train an ulcerative colitis (UC) classifier with incomplete UC labels. The order and location are used as the guiding information (RC: right colon. LC: left colon. Re: rectum).

for the lack of UC-labeled data using temporal ordering information, I formulated the ordinal loss, which is an objective function that brings temporally adjacent images closer in the feature space.

The contributions of this study are as follows:

- I propose a semi-supervised learning method that utilizes the location and temporal ordering information for UC classification. The proposed method introduces disentangled representation learning using location information to extract UC classification features that are separated from the location features.
- I formulate an objective function for order-guided learning to utilize temporal ordering information of endoscopic images. Order-guided learning can obtain the effective feature for classifying UC from unlabeled images by considering the relationship between the temporally adjacent images.

3.2 Related Work

Semi-supervised learning methods that utilize unlabeled samples efficiently have been reported in the training of classifiers when limited labeled data are available [11, 19, 20, 50]. Lee [19] proposed a method called Pseudo-Label, which uses the class predicted by the trained classifier as the ground-truth for unlabeled samples. Despite its simplicity, this method improves the classification performance in situations where labeled images are limited. Sohn *et al.* [20] proposed FixMatch, which improves the classification performance by making the predictions for weakly and strongly augmented unlabeled images closer during training. These semi-supervised learning methods work well when a classifier with a moderate classification performance has already been obtained using limited labels. However, in UC classification, which requires the learning of texture features from endoscopic images whose appearance varies depending on imaging location, it is difficult to obtain a classifier with a moderate classification performance using limited labeled endoscopic images, and applying these methods to UC classifications may not improve classification performance. Therefore, I propose a semi-supervised learning method that does not directly use the prediction results returned by a classifier trained by limited-labeled data, but utilizes two additional features: the location and the temporal ordering.

Several methods that utilize the temporal ordering information of images have been reported [21, 22, 53]. For example, Cao *et el.* proposed Temporal-Cycle Consistency (TCC), which is a self-supervised learning method that utilizes temporal alignment between sequences [21]. The TCC yields good image feature representation by maximizing the number of points where the temporal alignment matches. Dwibedi *et al.* [22] proposed a few-shot video classification method that utilizes temporal alignment between labeled and unlabeled video, then improved the video classification accuracy by minimizing the distance between temporally aligned frames. Moreover, a method for segmenting endoscopic image sequences has been proposed [53]. By utilizing the prior knowledge that temporally adjacent images tend to belong to the same class, this method segments an image sequence without requiring additional annotation. However, the methods proposed [21, 22] are not suitable for the task of this work, where involves a sequence with indefinite class transitions, because they assume that the class transitions in the sequence are the same. Furthermore, the method proposed in [53], which assumes segmentation of normal organ image sequences, is not suitable for the task of this work where the target image sequence consists of images of both normal and inflamed organs. In the proposed method, temporal ordering information is used to implement order-guided learning, which brings together temporal adjacency images that tend to belong to the same UC class, thus obtaining a good feature representation for detecting UC in the feature space.

3.3 Order-Guided Disentangled Representation Learning

The classification of UC using deep neural networks trained by general learning methods is difficult for two reasons. First, the appearances of the endoscopic images vary dynamically depending on location in the colon, whereas UC is characterized by the texture of the colon surface. Second, the number of UC-labeled images is limited because annotating UC labels to a large number of images requires significant effort by medical experts.

To overcome these difficulties, the proposed method introduces *disentangled representation learning* and *order-guided learning*. Figure 10 shows the overview of the proposed method. In disentangled representation learning using location information, I disentangle the image features into features for UC-dependent and location-dependent to mitigate the worse effect from the various appearance depending on the location. Order-guided learning utilizes the characteristics of an endoscopic image sequence in which temporally adjacent images tend to belong to the same class. I formulated an objective function that represents this characteristic and employed it during learning to address the limitation of the UC-labeled images.

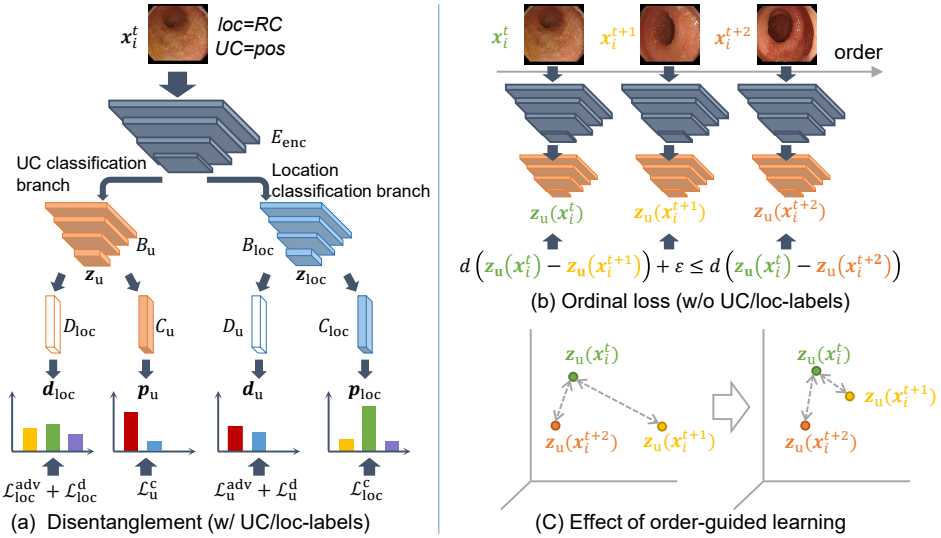


Fig. 10: Overview of the proposed method. (a) Disentanglement into the UC feature $\boldsymbol{z}_{\mathrm{u}}$ and the location feature $\boldsymbol{z}_{\mathrm{loc}}$. (b) Ordinal loss for order-guided learning. (c) Effect of order-guided learning.

3.4 Disentangled Representation Learning Using Location Information

Disentangled representation learning for the proposed method aims to separate the image features into UC and location-dependent features. These features are obtained via multi-task learning of UC and location classification. Along with the training of classifiers for UC and location classification tasks, the feature for one task is learned to fool the classifier for the other task; that is, the UC-dependent feature is learned to be non-discriminative with respect to location classification, and vice versa.

The network structure for learning disentangled representations is shown in Figure 10(a). This network has a hierarchical structure in which a feature extraction module branches into two task-specific modules, each of which further branches into two classification modules. The feature extraction module E_{enc} extracts a common feature vector for UC and location classification from the input image. The task-specific modules B_{u} and B_{loc} extract the UC feature $\boldsymbol{z}_{\mathrm{u}}$ and the location feature $\boldsymbol{z}_{\mathrm{loc}}$, which are disentangled features for UC and location classification. Out of four classification modules, the modules C_{u} and C_{loc} are used for UC and location classification, respectively, whereas D_{u} and D_{loc} are used to learn the disentangled representations.

In the left branch of Figure 10(a), the network obtains the prediction results for UC classes, $\boldsymbol{p}_{\mathrm{u}}$, as the posterior probabilities, based on the disentangled UC feature $\boldsymbol{z}_{\mathrm{u}}$ through learning. Hereinafter, I explain only the training of the left branch in detail because that of the right branch can be formulated by simply swapping the subscripts "loc" and "u" in the symbols for the left branch.

Given a set of N image sequences and corresponding location class labels $\{\boldsymbol{x}_i^{(1:T_i)}, \boldsymbol{l}_i^{(1:T_i)}\}_{i=1}^{N}$ and a set of limited UC class labels $\{\boldsymbol{u}_j^k \mid (j,k) \in \mathcal{U}\}$, where T_i is the number of images in the i-th image sequence and $\boldsymbol{u}_j^k$ is the UC class label corresponding to the j-th image in the k-th sequence, the training is performed based on three losses: classification loss $\mathcal{L}_{\mathrm{u}}^{\mathrm{c}}$, discriminative loss $\mathcal{L}_{\mathrm{loc}}^{\mathrm{d}}$, and adversarial loss $\mathcal{L}_{\mathrm{loc}}^{\mathrm{adv}}$. To learn the UC classification, I minimize the classification loss $\mathcal{L}_{\mathrm{u}}^{\mathrm{c}}$, which is computed by taking the cross-entropy between the UC class label $\boldsymbol{u}_i^t$ and the UC class prediction $\boldsymbol{p}_{\mathrm{u}}(\boldsymbol{x}_i^t)$ that is output from C_{u}. The discriminative loss $\mathcal{L}_{\mathrm{loc}}^{\mathrm{d}}$ and adversarial loss $\mathcal{L}_{\mathrm{loc}}^{\mathrm{adv}}$ are used to learn the disentangled representation, and are formulated as follows:

$$\mathcal{L}_{\mathrm{loc}}^{\mathrm{d}}(\boldsymbol{x}_i^t) = -\sum_{j=1}^{K_{\mathrm{loc}}} l_i^t \log d_{\mathrm{loc}}^j(\boldsymbol{x}_i^t),\ \mathcal{L}_{\mathrm{loc}}^{\mathrm{adv}}(\boldsymbol{x}_i^t) = \sum_{j=1}^{K_{\mathrm{loc}}} \log d_{\mathrm{loc}}^j(\boldsymbol{x}_i^t), \tag{7}$$

where $\boldsymbol{d}_{\mathrm{loc}}(\boldsymbol{x}_i^t)$ is the location class prediction estimated by D_{loc}. By minimizing the discriminative loss $\mathcal{L}_{\mathrm{loc}}^{\mathrm{d}}$, the classification module D_{loc} is trained to classify the location. In contrast, the minimization of the adversarial loss $\mathcal{L}_{\mathrm{loc}}^{\mathrm{adv}}$ results in the UC feature $\boldsymbol{z}_{\mathrm{u}}$ that is non-discriminative with respect to the location. Note that $\mathcal{L}_{\mathrm{loc}}^{\mathrm{d}}$ is back-propagated only to D_{loc}, whereas the parameters of D_{loc} are frozen during the back-propagation of $\mathcal{L}_{\mathrm{loc}}^{\mathrm{adv}}$. As mentioned above, some images are not labeled for UC classification in this problem. Therefore, the classification loss $\mathcal{L}_{\mathrm{u}}^{\mathrm{c}}$ and the disentangle losses $\mathcal{L}_{\mathrm{u}}^{\mathrm{adv}}$ and $\mathcal{L}_{\mathrm{u}}^{\mathrm{d}}$ are ignored for UC-unlabeled images.

3.5 Order-Guided Learning

Order-guided learning considers the relationship between temporally adjacent images, as shown in Figure 10(b). Since an endoscopic image is more likely to belong to the same UC class as its temporally adjacent images than the UC class of temporally distant images, the UC-dependent features of temporally adjacent images should be close to each other. To incorporate this assumption into learning of the network, the ordinal loss for order-guided learning is formulated as:

$$\mathcal{L}_{\mathrm{seq}}(\boldsymbol{x}_i^t, \boldsymbol{x}_i^{t+1}, \boldsymbol{x}_i^{t+2}) = \left[\|\boldsymbol{z}_{\mathrm{u}}(\boldsymbol{x}_i^t) - \boldsymbol{z}_{\mathrm{u}}(\boldsymbol{x}_i^{t+1})\|_2^2 - \|\boldsymbol{z}_{\mathrm{u}}(\boldsymbol{x}_i^t) - \boldsymbol{z}_{\mathrm{u}}(\boldsymbol{x}_i^{t+2})\|_2^2 + \varepsilon\right]_+, \tag{8}$$

where $\boldsymbol{z}_{\mathrm{u}}(\boldsymbol{x}_i^t)$ is a UC feature vector for the sample $\boldsymbol{x}_i^t$ and is extracted via E_{enc} and B_{u}, $[\cdot]_+$ is a function that returns zero for

a negative input and outputs the input directly otherwise, and ε is a margin that controls the degree of discrepancy between two temporally separated samples.

The UC features of temporally adjacent samples get closer by updating the network with the order-guided learning, as shown in Figure 10(c). This warping in the UC feature space functions as a regularization that allows the network to make more correct predictions because the temporally adjacent images tend to belong to the same UC class. The order-guided learning can be applied without the UC label, and therefore it is also effective for the UC-unlabeled images.

3.6 Experiments

I conducted the UC classification experiment to evaluate the validity of the proposed method. In the experiment, I used an endoscopic image dataset collected from the Kyoto Second Red Cross Hospital. Participating patients were informed of the aim of the study and provided written informed consent before participating in the trial. The experiment was approved by the Ethics Committee of the Kyoto Second Red Cross Hospital.

3.6.1 Dataset

The dataset consists of 388 endoscopic image sequences, each of which contains a different number of images, comprising 10,262 images in total. UC and location labels were attached to each image based on annotations by medical experts. Out of 10,262 images, 6,678 were labeled as UC (positive), and the remaining 3,584 were normal (negative). There were three classes for the location label: right colon, left colon, and rectum. In the experiments, the dataset was randomly split into image sequence units, and 7,183, 2,052, and 1,027 images were used as training, validation, and test set, respectively. To simulate the limitation of the UC-labeled images, the labeled image ratio R for the training set used by the semi-supervised learning methods was set to 0.1.

3.6.2 Comparative Method

I compared the proposed method with two semi-supervised learning methods. One is the Pseudo-Label [19], which is one of the famous semi-supervised learning methods. The other is FixMatch [20], which is the state-of-the-art semi-supervised learning method for the general image classification task. Since the distribution of data differs greatly between general and endoscopic images, I changed the details of FixMach to maximize its performance for UC classification. Specifically, strong augmentation was changed to weak augmentation, and weak augmentation was changed to rotation-only augmentation for processing unlabeled images. I also compared the proposed method with two classifiers trained with only labeled images in the training set with the labeled image ratio $R = 0.1$ and 1.0.

In addition, I conducted an ablation study to evaluate the effectiveness of the location label, disentangled representation learning, and order-guided learning. The best network parameter for each method was determined based on the accuracy of the validation set. I used precision, recall, F1 score, specificity, and accuracy as the performance measures.

3.6.3 Experimental Results

Table 7 shows the result of the quantitative performance evaluation for each method. Excluding specificity, the proposed method achieved the best performance for all performance measures. Although the specificity of the proposed method was the third-best, it was hardly different from that of the fully supervised classification. Moreover, I confirmed that the proposed method improved all measures of the classifier trained using only UC-labeled images in the training set with $R = 0.1$. In particular, the improvement in recall was confirmed only in the proposed method. Therefore, disentangled representation learning and order-guided learning, which use additional information other than UC information, were effective for improving UC classification performance.

Table 8 shows the results of the ablation study. The results demonstrated that each element of the proposed method was effective for improving the UC classification. The order-guided learning was effective for improving the precision and recall. Since the precision and recall were further improved by using the order-guided learning and disentangled representation learning simultaneously, it was confirmed that feature separation by disentangled representation learning is useful for the effective utilization of temporal ordering information. In contrast, when only location information was used, each score were improved. Since there is a correlation between the transitions between location and UC labels, the performance of the proposed method on UC classification is slightly improved as the proposed method learns to classify location.

To demonstrate the effect of the order-guided learning, the examples of prediction results are shown in Figure 11. In this figure, the prediction results from the proposed method with the order-guided learning for temporally adjacent images tend to belong to the same class. For example, the proposed method predicted the first and second images from the right in Figure 11(b) as the same class, whereas the proposed method without the order-guided learning predicted them as different classes.

The limitations of the proposed method are as follows. First, the proposed method is optimized from several objective functions, and tuning their weights is time-consuming task. In particular, if the weights of the objective functions for the disentangled representation learning are inappropriate, the feature space suitable for UC classification cannot be obtained. Second, since the proposed method is provided a pair of images as input data for training, the model size is larger and the training time is longer than the existing semi-supervised learning methods.

4. Conclusion and Future Work

4.1 Conclusion

This paper proposed two approaches to solve the problem of limited annotation in bio-medical data analysis tasks.

I proposed a new constrained clustering method suitable for bio-medical data clustering. The appearance of a medical image often changes even in the same class due to some factors, such as camera angle and light source intensity. When the must-link is attached to the pair of samples that their appearance differs greatly, satisfying this link results in an unexpectedly large cluster with low purity. Therefore, the proposed method allows a violation of must-links to handle the above issue. In the experiment, the proposed method achieved higher prominent-purity and

Table 7: Quantitative performance evaluation. Labeled image ratio R represents the ratio of the UC-labeled images in the training set.

Method	R	Precision	Recall	F1	Specificity	Accuracy
Supervised learning	1.0	0.805	0.849	0.826	0.902	0.885
	0.1	0.692	0.671	0.681	0.858	0.797
Pseudo-Label [19]	0.1	0.752	0.613	0.676	0.904	0.811
FixMatch [20]	0.1	0.752	0.468	0.577	**0.927**	0.779
Proposed	0.1	**0.776**	**0.731**	**0.753**	0.899	**0.845**

Table 8: Results of the ablation study with order-guided learning (Order), the location label (Location), and disentangled representation learning (Disentangle)

Order	Location	Disentangle	Precision	Recall	F1	Specificity	Accuracy
			0.692	0.671	0.681	0.858	0.797
✓			0.752	0.686	0.717	0.892	0.826
	✓		**0.795**	0.680	0.733	0.917	0.840
✓	✓		0.789	0.622	0.696	**0.921**	0.825
✓	✓	✓	0.776	**0.731**	**0.753**	0.899	**0.845**

higher prominent-recall, compared with several state-of-the-art soft-constrained clustering methods. In addition, I proposed a self and soft-constrained clustering method, where self-constraints are defined as prior knowledge that temporally adjacent endoscopic images tend to belong to the same class. The experimental results showed that the proposed method improved clustering performance compared to the ordinary clustering method and several soft-constrained clustering methods that utilize the self-constraints.

I also proposed a semi-supervised learning method, called the order-guided disentangled representation learning method, for learning ulcerative colitis (UC) classification. The proposed method utilizes the location information and image capturing order of endoscopic images. The proposed method performed disentangled representation learning that separates the UC-dependent and location-dependent features with image capturing order that is effective for learning UC classification. The experiments using an endoscopic image dataset demonstrated that the proposed method outperforms several existing semi-supervised learning methods.

4.2 Future Work

Future work of this study is listed as follows:

- For the proposed constrained clustering method, I will further analyze the relationship between the purity of the clusters and the hyper-parameter that determines the strength of the self-constraint. I will extend the proposed method to other bio-medical data clustering tasks. In addition, I will consider the introduction of the proposed method to representation learning using clustering results for deep learning techniques.
- For the proposed semi-supervised learning method, I will focus on extending the proposed method to other bio-medical data analysis tasks, such as the detection of polyps and cancer. In addition, I will extend the proposed method to a classification method using an image capturing order at the testing time.

References

[1] C. A. Kulikowski, S. M. Weiss: Representation of expert knowledge for consultation: The CASNET and EXPERT projects, *Artificial Intelligence in Medicine*, pp. 21–55 (2019).

[2] I. Hatzilygeroudis, J. Prentzas: Integrating (rules, neural networks) and cases for knowledge representation and reasoning in expert systems, *Expert Systems with Applications*, Vol. 27, No. 1, pp. 63–75 (2004).

[3] Y. Shen *et al.*: Emerging medical informatics with case-based reasoning for aiding clinical decision in multi-agent system, *Journal of Biomedical Informatics*, Vol. 56, pp. 307–317 (2015).

[4] G. Litjens *et al.*: Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis, *Scientific Reports*, Vol. 6, No. 1, p. 26286 (2016).

[5] M. Bakator, D. Radosav: Deep Learning and Medical Diagnosis: A Review of Literature, *Multimodal Technologies and Interaction*, Vol. 2, No. 3 (2018).

[6] W. Sun, B. Zheng, W. Qian: Computer aided lung cancer diagnosis with deep learning algorithms, *Medical imaging 2016: computer-aided diagnosis*, Vol. 9785, International Society for Optics and Photonics, p. 97850Z (2016).

[7] T. M. K. Connor Shorten: A survey on Image Data Augmentation for Deep Learning, *Journal of Big Data* (2019).

[8] B. K. Iwana, S. Uchida: An empirical survey of data augmentation for time series classification with neural networks, *PLOS ONE*, Vol. 16, No. 7, pp. 1–32 (online), DOI: 10.1371/journal.pone.0254841 (2021).

[9] M. Wigness, B. A. Draper, R. J. Beveridge: Efficient label collection for unlabeled image datasets, *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 108, pp. 4594–4602 (2015).

[10] M. Wigness, B. A. Draper, J. R. Beveridge: Efficient Label Collection for Image Datasets via Hierarchical Clustering, *International Journal of Computer Vision*, Vol. 126, pp. 59–85 (2018).

[11] E. Arazo *et al.*: Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning, *Proceedings of the International Joint Conference on Neural Networks* (2020).

[12] A. Biswas, D. W. Jacobs: Active image clustering: Seeking constraints from humans to complement algorithms, *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2152–2159 (2012).

[13] C. Galleguillos, B. McFee, G. R. G. Lanckriet: Iterative Category Discovery via Multiple Kernel Metric Learning, *International Journal of Computer Vision*, Vol. 108, pp. 115–132 (2014).

[14] S. Mousavi *et al.*: Collaborative Learning of Semi-Supervised Clustering and Classification for Labeling Uncurated Data (2020).

[15] Y. J. Lee, K. Grauman: Object-Graphs for Context-Aware Visual Category Discovery, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 34, No. 2, pp. 346–358 (2012).

[16] T. Tuytelaars *et al.*: Unsupervised Object Discovery: A Comparison, *International Journal of Computer Vision*, Vol. 88, No. 2, pp. 284–302 (2010).

[17] D. Dai *et al.*: Ensemble Partitioning for Unsupervised Image Categorization, *Proceeding of the European Conference on Computer Vision*, Springer Berlin Heidelberg, pp. 483–496 (2012).

[18] C. Xiong, D. Johnson, J. Corso: Spectral Active Clustering via Purification of the *k*-Nearest Neighbor Graph, *Proceeding of the European Conference on Data Mining* (2012).

[19] D.-H. Lee: Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks, *Proceedings of the ICML 2013 Workshop: Challenges in Representation Learning* (2013).

[20] K. Sohn *et al.*: FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence, *Proceedings of the Advances in Neural Information Processing Systems*, Vol. 33, pp. 596–608 (2020).

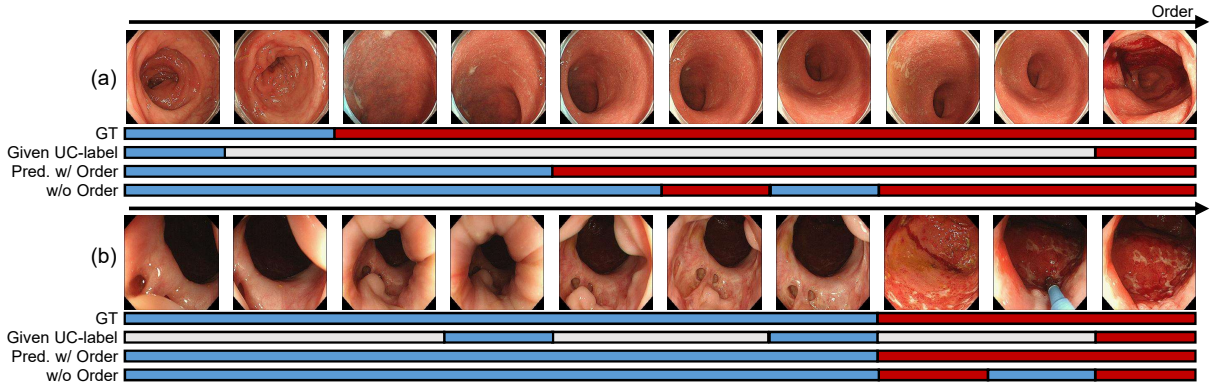


Fig. 11: Examples of the prediction results. Each bar represents the ground-truth labels, labels given during training, prediction results by the proposed method with and without order-guided learning. The red, blue, and gray bars represent UC, normal, and unlabeled images, respectively.

[21] K. Cao *et al.*: Few-Shot Video Classification via Temporal Alignment, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020).
[22] D. Dwibedi *et al.*: Temporal Cycle-Consistency Learning, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019).
[23] A. H. Liu *et al.*: A Unified Feature Disentangler for Multi-Domain Image Translation and Manipulation, *Proceedings of the Advances in Neural Information Processing Systems*, Vol. 31 (2018).
[24] Y. Liu *et al.*: Exploring Disentangled Feature Representation Beyond Face Identification, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018).
[25] C. Bass *et al.*: ICAM: Interpretable classification via disentangled representations and feature attribution mapping, *Proceedings of International Conference on Neural Information Processing Systems* (2020).
[26] Z. Zhang *et al.*: Gait recognition via disentangled representation learning, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4710–4719 (2019).
[27] R. Cai *et al.*: Learning disentangled semantic representation for domain adaptation, *Proceedings of the International Conference on Artificial Intelligence*, Vol. 2019, NIH Public Access, p. 2060 (2019).
[28] V.-H. Tran, C.-C. Huang: Domain adaptation meets disentangled representation learning and style transfer, *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, IEEE, pp. 2998–3005 (2019).
[29] L. Ma *et al.*: Disentangled person image generation, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 99–108 (2018).
[30] K. Wagstaff *et al.*: Constrained K-means Clustering with Background Knowledge, *Proceedings of the International Conference on Machine Learning*, ICML '01, Morgan Kaufmann Publishers Inc., pp. 577—-584 (2001).
[31] N. Shental *et al.*: Computing Gaussian Mixture Models with EM Using Equivalence Constraints, *Proceedings of the International Conference on Neural Information Processing Systems*, NIPS'03, MIT Press, pp. 465—-472 (2003).
[32] Z. Li, J. Liu, X. Tang: Constrained clustering via spectral regularization, *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 421–428 (2009).
[33] Z. Li, J. Liu: Constrained clustering by spectral kernel learning, *Proceeding of IEEE International Conference on Computer Vision*, pp. 421–427 (2009).
[34] H. Le *et al.*: A Binary Optimization Approach for Constrained K-means Clustering, *Proceeding of the Asian Conference on Computer Vision*, pp. 577–584 (2018).
[35] I. Davidson, S. S. Ravi: Clustering with constraints: Feasibility issues and the K-means algorithm, *Proceedings of the SIAM international conference on data mining*, pp. 138–149 (2005).
[36] D. Pelleg, D. Baras: K-means with Large and Noisy Constraint Sets, *Proceeding of the European Conference on Machine Learning*, Springer Berlin Heidelberg, pp. 674–682 (2007).
[37] M. E. Ares, J. Parapar, Á. Barreiro: Avoiding Bias in Text Clustering Using Constrained K-means and May-Not-Links, *Proceedings of the International Conference on the Theory of Information Retrieval*, Springer Berlin Heidelberg, pp. 322–329 (2009).
[38] S. Basu, A. Banerjee, R. J. Mooney: Active Semi-Supervision for Pairwise Constrained Clustering, *Proceedings of the SIAM International Conference on Data Mining*, pp. 333–344 (2004).
[39] M. Bilenko, S. Basu, R. J. Mooney: Integrating Constraints and Metric Learning in Semi-Supervised Clustering, *Proceedings of the International Conference on Machine Learning*, ICML '04, New York, NY, USA, Association for Computing Machinery, pp. 81–88 (2004).
[40] Y.-C. Hsu, Z. Kira: Neural network-based clustering using pairwise constraints (2015).
[41] H. Zhang, S. Basu, I. Davidson: A Framework for Deep Constrained Clustering – Algorithms and Advances (2019).
[42] S. Fogel *et al.*: Clustering-driven Deep Embedding with Pairwise Constraints (2018).
[43] L. Maier-Hein *et al.*: Crowdsourcing for Reference Correspondence Generation in Endoscopic Images, *Medical Image Computing and Computer-Assisted Intervention*, Springer International Publishing, pp. 349–356 (2014).
[44] T. S. Kim *et al.*: Crowdsourcing Annotation of Surgical Instruments in Videos of Cataract Surgery, *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis* (D. Stoyanov *et al.*, eds.), Cham, Springer International Publishing, pp. 121–130 (2018).
[45] V. Cheplygina *et al.*: Early Experiences with Crowdsourcing Airway Annotations in Chest CT, *Deep Learning and Data Labeling for Medical Applications*, Springer International Publishing, pp. 209–218 (2016).
[46] R. M. Rifkin, R. A. Lippert: Notes on regularized least-squares, Technical report, Massachusetts Institute of Technology Computer Science and Artificial Intelligence Laboratory (2007).
[47] G. Huang *et al.*: Densely Connected Convolutional Networks, *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2261–2269 (2017).
[48] J. Deng *et al.*: ImageNet: A Large-Scale Hierarchical Image Database, *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255 (2009).
[49] L. van der Maaten, G. Hinton: Visualizing Data using t-SNE, *Journal of Machine Learning Research*, Vol. 9, No. 86, pp. 2579–2605 (2008).
[50] D. Berthelot *et al.*: MixMatch: A Holistic Approach to Semi-Supervised Learning, *Proceedings of the Advances in Neural Information Processing Systems*, Vol. 32 (2019).
[51] J. Herp *et al.*: Feature Point Tracking-Based Localization of Colon Capsule Endoscope, *Diagnostics*, Vol. 11, No. 2 (2021).
[52] K. Mori *et al.*: A Method for Tracking the Camera Motion of Real Endoscope by Epipolar Geometry Analysis and Virtual Endoscopy System, *Proceedings of the Medical Image Computing and Computer-Assisted Intervention*, pp. 1–8 (2001).
[53] S. Harada *et al.*: Endoscopic Image Clustering with Temporal Ordering Information Based on Dynamic Programming, *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 3681–3684 (2019).