Correlation Clustering Research on SwAV for Single-organoid Images

Xuetong Fu
TU Dresden

Abstract

Single-organoid microscopy images exhibit substantial heterogeneity in morphology, size, and internal organization, which makes automated visual analysis challenging, particularly in the absence of clearly defined class labels. Given the success of visual representation learning methods on large-scale natural image datasets, an important question is whether such methods can be transferred to single-organoid image analysis.

In this project, I investigate the applicability of SwAV[2], a clustering-based self-supervised representation learning method originally developed for natural images. I adapt SwAV to the organoid image domain and evaluate the learned representations through comparison with a supervised correlation-based method, as well as through clustering performance on labeled and unlabeled datasets. The results indicate that SwAV can learn informative representations for single-organoid images, supporting its feasibility for organoid image clustering.

1 Introduction

Recent advances in microscopy and image processing enable large-scale collections of high-resolution organoid images and reliable extraction of single-organoid crops, making computational analysis of organoid morphology increasingly feasible.

Organoid images exhibit large variability in morphology and internal organization, and class boundaries are often unclear. This motivates clustering based on visual similarity rather than supervised classification.

A common strategy for addressing image clustering problems is to first learn meaningful feature representations and then perform clustering in the learned embedding space. In recent years, self-supervised representation learning methods developed in the natural image domain have demonstrated strong performance without requiring manual annotations. Representative approaches such as SimCLR[3], SwAV, MSN[1], and MAE[4] explore different learning principles, including contrastive objectives, clustering-based assignments, masked reconstruction, and teacher–student frameworks. Together, these methods define a broad design space for unsupervised visual representation learning and provide potential candidates for organoid image analysis.

In this project, I review several self-supervised learning paradigms with respect to their suitability for single-organoid images and downstream clustering tasks. Based on this analysis, SwAV is selected as the primary method, as its clustering-based learning objective naturally aligns with the goal of organoid image clustering. The focus of this work is not on improving the SwAV algorithm itself, but on evaluating its applicability to the organoid domain. The learned representations are assessed through downstream clustering using a correlation clustering framework, with quantitative evaluations on labeled test sets and qualitative analysis on unlabeled organoid images.

2 Related Work

This section reviews representative works related to this research and provides background for understanding the methodological context of the study. It focuses on common conceptual distinctions and research paradigms in visual representation learning and clustering, and discusses these directions in the context of organoid microscopy images.

2.1 Self-Supervised Visual Representation Learning

Self-supervised visual representation learning aims to learn transferable features without manual annotations by constructing training signals from the data itself. Representative approaches include contrastive methods (e.g., SimCLR), clustering-based methods (e.g.,

SwAV), teacher–student methods (e.g., MSN), and reconstruction-based methods (e.g., MAE), which differ in how they define and enforce consistency objectives in the embedding space.

In this study, I focus on SwAV and evaluate whether its learned representations transfer to clustering of organoid microscopy images.

2.2 Representation Learning for Clustering

Among existing clustering methods, many algorithms rely on certain prior assumptions, such as pre-setting the number of clusters, assuming the existence of cluster centers, or partitioning samples based on fixed distance metrics. Although these methods have been well validated on natural image datasets, they may not be directly transferable or reproducible at a similar level of performance when applied to organoid images.

One important reason for this lies in the significant differences between the two data types. Organoid images exhibit substantial variation in morphology and structure, making it difficult to assign them to clearly defined semantic categories. As a result, it is challenging to identify clear semantic anchors for clustering. In addition, organoid groupings are often sensitive to subtle morphological changes and cellular-level microscopic structures, and visually similar samples may still belong to different clusters. This places higher demands on the quality of the learned feature representations. Furthermore, both the image content and simple statistical characteristics indicate that organoid images tend to have darker overall tones and often contain large regions with similar or nearly uniform colors. Under such conditions, clustering methods based on fixed distance metrics may struggle to accurately capture sample similarity.

In contrast, correlation clustering methods infer clustering results directly from pairwise similarity relationships between samples. They do not require prior assumptions about the number of clusters or the specific cluster structure, nor do they rely on strong prior knowledge of the data's semantic organization. As a result, they offer greater flexibility in scenarios where category structures are complex or boundaries are ambiguous. Meanwhile, the prototype assignment and soft clustering mechanisms introduced by SwAV during training are also based on modeling similarities between samples, and are therefore conceptually aligned with correlation clustering. Based on this observation, combining SwAV-learned embeddings with correlation clustering provides a feasible and natural approach for the clustering analysis of organoid microscopy images.

3 Methods

Organoid images exhibit substantial morphological variability, and class labels are often difficult to obtain, making unsupervised representation learning a natural choice for this task.

Specifically, I adopt SwAV, a clustering-based self-supervised learning method that enforces consistency of prototype assignments across different augmented views of the same image (swapped prediction). This produces embeddings that are well suited for similarity modeling and downstream clustering.

After learning visual representations in a self-supervised manner, I evaluate their suitability for organoid clustering using correlation clustering, which groups samples based on pairwise join/cut preferences without assuming a predefined number of clusters.

3.1 Self-Supervised Representation Learning with SwAV

This section follows the formulation of SwAV [2] and summarizes the key components and equations used in my implementation.

SwAV jointly optimizes three components during training: a backbone network, which is responsible for extracting high-dimensional and highly transferable feature representations from the preprocessed images; a projection head, which maps these backbone features into a low-dimensional embedding space where constraints based on clustering assignments are explicitly enforced; and a set of learnable prototype vectors, which serve as candidate centers of potential clusters in the dataset and shape the organization of embeddings in the projection space. Through the joint learning of these components, SwAV learns low-dimensional embeddings that exhibit clear clustering structures, providing a suitable representation basis for subsequent similarity modeling and clustering tasks.

3.1.1 Network Architecture

Each augmented view x is processed by the backbone network f_θ (ResNet-50 in the experiments) to extract a high-dimensional feature representation

$$h = f_\theta(x) \in R^{d_b}. \tag{1}$$

This representation is designed to capture general and transferable characteristics of the input image. These features are then passed to a projection head g_ϕ, which maps them into a lower-dimensional embedding space,

$$z = g_\phi(h) \in R^{d}, \tag{2}$$

where the embedding z is ℓ_2-normalized. Hence, the resulting low-dimensional embeddings are particularly well-suited for organoid clustering and are directly used in the subsequent stages of this project.

3.1.2 Prototype Assignment

To enable clustering-guided representation learning, SwAV introduces a set of learnable prototype vectors in the projection space:

$$\{c_1, \ldots, c_K\}, \quad c_k \in R^d. \tag{3}$$

These prototypes act as cluster centers that evolve during training and provide assignment-based pseudo-supervisory signals to guide representation learning.

Given a projected embedding z, its relationship to the current clustering structure is measured by computing the similarity to each prototype using the dot product

$$\begin{aligned} s_k &= z^\top c_k, \\ S &= C^\top Z \in R^{K\times B}, \end{aligned} \tag{4}$$

which defines a soft, probabilistic association with all prototypes, rather than a one-hot assignment to a single prototype. Meanwhile, at the level of a full mini-batch, SwAV computes a soft assignment by solving an entropy-regularized optimal transport problem. Specifically, for a set of embeddings (typically obtained from the large-crop views as defined in SwAV), this process produces an assignment matrix

$$Q = [q_{ij}] \in R^{K\times B}, \tag{5}$$

where q_{ij} denotes the assignment probability of sample j to prototype i. The assignment matrix Q is obtained by solving the following optimization problem:

$$\max_{Q\in\mathscr{Q}} \operatorname{Tr}(Q^\top S) + \varepsilon H(Q), \tag{6}$$

where $S$is the similarity matrix defined above, $H(Q)$ denotes the entropy of the assignment matrix, and the constraint set $\mathscr{Q}$ enforces approximately balanced assignments across prototypes:

$$\mathscr{Q} = \left\{ Q \in R_+^{K\times B} \;\middle|\; Q\mathbf{1}_B = \frac{1}{K}\mathbf{1}_K,\; Q^\top \mathbf{1}_K = \frac{1}{B}\mathbf{1}_B \right\}, \tag{7}$$

In this project, the optimization problem is solved approximately using the Sinkhorn-Knopp algorithm.

3.1.3 Multi-Crop Strategy

To ensure consistency across data augmentations, SwAV adopts a multi-crop strategy, which enables the model to learn both global and local features at different spatial scales. This is particularly important for clustering organoid images, as they exhibit substantial morphological heterogeneity: samples within the same biological category may differ in appearance, while samples from different categories can appear visually similar. Consequently, representations extracted at a single scale are often insufficient to capture all relevant visual cues.

By generating multiple views at different resolutions, the multi-crop strategy encourages SwAV to learn scale-consistent representations from both global and local crops. High-resolution crops capture the overall shape and coarse structural patterns, whereas low-resolution crops emphasize local and fine-grained features. Together, this design helps the model focus on biologically relevant characteristics while reducing sensitivity to irrelevant variations.

3.1.4 Swapped Prediction Loss

For each image x, I generate multiple augmented views $\{x^{(1)}, \ldots, x^{(V)}\}$ using a random transformation pipeline. Each view is independently processed by the network illustrated in section 3.1.1 to produce a normalized embedding.

Given an embedding z_t from view t and the prototype assignment q_s computed from a different view $s \neq t$, SwAV defines a swapped prediction loss to encourage consistency between different views of the same image. Specifically, the loss for predicting the assignment q_s from the embedding z_t is defined as

$$\ell(z_t, q_s) = -\sum_{k=1}^{K} q_s^{(k)} \log p_t^{(k)}, \tag{8}$$

where

$$p_t^{(k)} = \frac{\exp\left(\frac{1}{\tau} z_t^\top c_k\right)}{\sum_{k'=1}^{K} \exp\left(\frac{1}{\tau} z_t^\top c_{k'}\right)} \tag{9}$$

denotes the softmax-normalized similarity between the embedding z_t and the prototype c_k, and τ is a temperature parameter.

Taking this loss over all views leads to the following SwAV loss function:

$$\mathscr{L}_{\text{SwAV}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{\substack{s,t\in\mathscr{T}\\ s\neq t}} \left[\frac{1}{\tau} z_{nt}^\top C q_{ns} - \log \sum_{k=1}^{K} \exp\left(\frac{z_{nt}^\top c_k}{\tau}\right) \right], \tag{10}$$

where N is the number of images in the batch, $\mathscr{T}$ is the set of augmented views, C is the matrix of prototype vectors, and q_{ns} is the assignment vector computed from the view s of an image n.

This swapped prediction loss enables view-invariant prototype assignments without requiring explicit negative samples. It also jointly updates the backbone network, projection head, and prototype vectors.

3.1.5 Training Objective and Output Representation.

The overall training objective is obtained by averaging the swapped prediction loss across all images and view pairs in each batch. In this work, for each organoid image x, I use the projection head output as the final representation,

$$z = \text{normalize}\big(g_{\hat{\phi}}(f_{\hat{\theta}}(x))\big), \tag{11}$$

since both the prototype assignments and the SwAV training objective are defined in the projection space. The resulting low-dimensional embeddings are therefore particularly well suited for similarity modeling and clustering.

These embeddings serve as inputs to subsequent stages of the pipeline, including cosine-similarity-based pairwise modeling and the construction of edge costs for correlation clustering.

3.2 Embedding Extraction

After the self-supervised training phase, the parameters of the encoder are frozen to obtain a fixed representation space for downstream clustering. For each image, a forward pass through the frozen network is performed to extract its visual representation.

Specifically, embeddings are taken from the output of the projection head, resulting in a fixed 128-dimensional feature vector for each image. During embedding extraction, a single standard-resolution view is used per image, without applying multi-crop augmentation.

These fixed embeddings are then used as input for the downstream clustering methods described in the following section.

3.3 Linear Pairwise Head

To further assess the representation quality learned by SwAV, I implement a linear pairwise classification head on top of the frozen embeddings. Using labeled data, sample pairs are constructed and assigned binary labels indicating whether the two samples belong to the same category. A linear classifier is then trained to predict join/cut decisions based on the paired embeddings. The model is trained on Train-100 and evaluated on the corresponding test splits.

This pairwise head is used solely as an evaluation tool to measure the linear separability of the learned representations at the pairwise level and is not involved in the subsequent correlation clustering pipeline. The evaluation metrics reported for this model are aligned with the pairwise metrics used in correlation clustering, enabling direct comparison across evaluation settings.

3.4 Correlation Clustering on Learned Embeddings

Given image representations extracted from the frozen SwAV encoder, I perform downstream clustering using correlation clustering. This method does not require setting the number of clusters in advance. Instead, it formulates clustering as a data-partitioning problem that aims to minimize disagreement with a given set of pairwise join/cut preferences (or costs). I select a margin m for pairwise cosine similarity to determine join or cut decisions by maximizing the accuracy of pairwise decisions on the labeled data.

3.4.1 Problem Formulation

Let $\{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$ denote the ℓ_2-normalized embeddings of N images. In this work, I define pairwise preferences using the cosine similarity between image embeddings from the frozen SwAV encoder. For example, consider two image samples p and q. For ℓ_2-normalized embedding representations, the pairwise similarity is defined as

$$s_{pq} = \mathbf{z}_p^\top \mathbf{z}_q. \tag{12}$$

By introducing a margin parameter m, I convert the pairwise similarities into costs for correlation clustering:

$$c_{pq} = m - s_{pq}. \tag{13}$$

A negative cost $c_{pq} < 0$ indicates that the pair (p,q) is encouraged to be assigned to the same cluster, while a positive cost $c_{pq} > 0$ favors separating the pair into different clusters.

3.4.2 Exact and Approximate Solvers

Since the exact optimization problem associated with correlation clustering is NP-hard, exact solutions are computationally feasible only for small datasets. Therefore, I adopt two different solution strategies depending on the dataset size, using exact optimization for small splits and approximate algorithms for larger ones.

Exact solver: ILP-based correlation clustering. For small-scale datasets, I solve the problem as an integer linear programming (ILP) problem. Specifically, for each pair of samples (p,q), I introduce a binary decision variable

$$x_{pq} \in \{0,1\}, \tag{14}$$

where $x_{pq} = 1$ means that samples p and q are assigned to different clusters (cut), while $x_{pq} = 0$ means that they are assigned to the same cluster (join).

To ensure that cluster assignments are consistent, I restrict the solution space by adding triangle inequality constraints. For any triplet (p,q,r), the following constraint is imposed:

$$x_{pq} \leq x_{pr} + x_{rq}, \tag{15}$$

and the analogous constraints obtained by permuting (p,q,r). Among all feasible solutions, the objective is to select the one that best agrees with the pairwise join/cut preferences induced by the embedding similarities.

For implementation convenience, I introduce non-negative penalty weights that quantify the disagreement between a binary decision x_{pq} and the corresponding pairwise preference. This matches my code. Here, the signed pairwise preference is defined as

$$w_{pq} = -c_{pq} = s_{pq} - m, \tag{16}$$

where positive values indicate a preference for joining and negative values indicate a preference for cutting. Specifically, a penalty is produced when a pair that is preferred to be joined is cut, or when a pair that is preferred to be separated is joined. Decisions that agree with the pairwise preference do not result in any penalty:

$$w^+_{pq} = \max(w_{pq}, 0), \quad w^-_{pq} = \max(-w_{pq}, 0), \tag{17}$$

where cutting a preferred-join pair incurs cost w^+_{pq}, while joining a preferred-cut pair incurs cost w^-_{pq}. Under these constraints, the goal is to minimize the following total disagreement with the pairwise join/cut preferences:

$$\sum_{p<q} \left(w^+_{pq} x_{pq} + w^-_{pq} (1 - x_{pq}) \right), \tag{18}$$

By solving this ILP, an exact optimal solution to the correlation clustering problem can be obtained.

Approximate solver: Pivot For larger datasets, I use the Pivot algorithm (KwikCluster) as an approximate solution for correlation clustering. The algorithm runs through simple iterations. It repeatedly selects an unassigned sample at random as a starting point. Based on a pre-computed k-nearest-neighbor graph, all first-level neighbors connected to this sample are grouped into the same cluster. The samples in this cluster are then removed from the unassigned set. The procedure above is repeated until all samples are assigned.

3.4.3 Sparse Graph Construction via Mutual KNN

In the approximate solution, the most important part is the construction of a sparse graph based on mutual k-nearest neighbors. To achieve high efficiency, the graph stores only positive edges, that is, edges corresponding to k-nearest-neighbor relations. All other edges are treated as negative edges and are not stored explicitly.

Note that although the pairwise similarity is symmetric by definition, the k-nearest-neighbor graph construction introduces directionality due to the local top-k selection performed independently for each sample. As a result, asymmetric neighbor relations can arise when one sample selects another as a nearest neighbor, but not vice versa. To address this issue, a mutual constraint is applied, such that an edge is kept only when two samples select each other as nearest neighbors. This removes asymmetric edges. In addition, the neighbor iteration depth is limited to one level.

In practice, the k-nearest neighbors are identified using cosine distance in the embedding space. For ℓ_2-normalized embeddings, the cosine distance between two samples is given by

$$d_{pq} = 1 - \mathbf{z}_p^\top \mathbf{z}_q, \tag{19}$$

which is converted back to cosine similarity as

$$s_{pq} = 1 - d_{pq} = \mathbf{z}_p^\top \mathbf{z}_q. \tag{20}$$

Rather than being used as a continuous optimization weight, the cosine similarity serves as a binary criterion for sparse graph construction. Specifically, a neighboring sample pair (p,q) is retained as a positive edge only if its cosine similarity exceeds the same predefined margin threshold m, used in ILP. Pairs that do not satisfy this condition are discarded and treated as negative edges.

The resulting sparse graph therefore consists only of sample pairs that are mutual k-nearest neighbors and have sufficiently high cosine similarity. This graph is subsequently used as the input to the pivot-based correlation clustering algorithm to obtain an approximate clustering result.

4 Experiments

The goal of the experimental design is to evaluate whether visual representations learned in the SwAV model are suitable for clustering organoid images, in particular under the evaluation protocol defined in the project description. Correlation clustering has previously been explored in the context of organoid image analysis, providing a suitable evaluation framework for this domain [5].

4.1 Datasets and Splits

I use the organoid image dataset provided in the project description [6]:

- Unlabeled-80k is used exclusively for self-supervised representation learning with SwAV. No labels are used at this stage.

- Train-100 is used only for selecting hyperparameters and thresholds in downstream evaluation, such as similarity margins or clustering costs.

- Test-100 contains organoid classes seen during training and is used to evaluate generalization to unseen samples of known classes.

- Test-30 contains organoid classes not present in Train-100 and Test-100 and is used to evaluate generalization to entirely unseen classes.

- An additional unlabeled test set is used for qualitative evaluation of clustering results.

This split ensures a clear separation between representation learning, model fine-tuning, and evaluation. Hence, it allows us to analyze both qualitative and quantitative.

4.2 Implementation Details

4.2.1 Representation Learning

The SwAV model for representation learning was trained on four GPUs using distributed data parallel (DDP). Training was performed on the Unlabeled-80k dataset. A ResNet-50 backbone was used together with a projection head producing 128-dimensional embeddings. In total, 100 prototype vectors were used during training, and the batch size was set to 64.

During the early stages of training, unstable training behavior was observed. The possible causes were investigated, and several parameters of the original SwAV configuration were adjusted accordingly. The final hyperparameter setting used in this experiment is summarized as follows. A multi-crop strategy is applied, consisting of two large crops of size 224×224 and four small crops of size 96×96. The scale range is set to $[0.3, 1.0]$ for large crops and $[0.1, 0.3]$ for small crops. The temperature parameter is set to 0.2, and two Sinkhorn iterations are used to compute prototype assignments.

The learning rate follows a warm-up and cosine decay schedule. The initial learning rate is 0.025 and the final learning rate is 2×10^{-4}. The warm-up stage lasts for 10 epochs. During the first 8,000 iterations, prototype updates are frozen. A queue of length 4096 is enabled after 30 epochs. The weight decay is set to 10^{-4}, and mixed-precision (FP16) training is used.

4.2.2 Correlation Clustering

The downstream work of correlation clustering is based on cosine similarity in the SwAV embedding space. For labeled datasets, an ILP solver is used to exactly minimize pairwise disagreement. For unlabeled datasets, a k-nearest-neighbor graph is constructed to build a sparse positive-edge graph, with $k = 30$. During evaluation, the batch size is set to 64, the input image size is 224, and a fixed random seed of 0 is used to ensure reproducibility.

Margin Selection. The similarity margin m is selected exclusively on the Train-100 dataset. A grid sweep is performed over the interval $m \in [0.10, 0.90]$ with 81 values, and the accuracy (ACC) of pairwise join/cut decisions is used as the selection criterion. The selected margin m is then fixed and used in all subsequent evaluations.

Quantitative Evaluation on Labeled Data. With the margin m fixed, correlation clustering is evaluated on Test-100, Test-30, and their combined set (Test-130). Test-100 and Test-30 are used to assess generalization to seen and completely unseen classes, respectively, while Test-130 provides an overall evaluation across both test splits. For these labeled datasets, both pairwise decision metrics and clustering-level evaluation metrics are reported. For smaller splits, the ILP solver is used to obtain exact correlation clustering solutions.

Qualitative Evaluation on Unlabeled Data. For the unlabeled dataset, the same fixed margin m and k-nearest-neighbor setting are used as in the labeled evaluation. Since ground-truth labels are not available, margin selection and quantitative evaluation are not possible. Therefore, only qualitative analysis is performed. Clustering results are visualized using cluster montages and two-row summary visualizations for manual inspection.

4.3 Quantitative Evaluation-Metric Settings

In the original paper of SwAV, the linear classification after freezing the representation is usually adopted as the evaluation method, which is used to measure the linear separability of the learned representations on known categories. However, in the dataset setting of organoids used in this experiment, this evaluation method has certain limitations. Specifically, Train-100 and Test-100 come from the same distribution of organoid categories. They are highly similar in morphology and statistical characteristics. When conducting linear classification evaluation under this setting, it reflects more the ability to distinguish known categories. However, Test-30 and

the unlabeled dataset contain clusters that have never appeared in Train-100. The linear classifier trained based on 10 classes cannot be meaningfully tested on these data. Therefore, the linear classification accuracy cannot be used as a unified evaluation metric.

Based on these reasons, this experiment did not adopt the standard linear classification evaluation method. Instead, a linear pairwise evaluation task based on frozen representations was introduced. This task uses "whether the sample pairs belong to the same category" as the supervisory signal to evaluate the discriminative ability of the representations at the pairwise relationship level. This evaluation method does not rely on a fixed set of categories and can maintain a consistent evaluation form across different data partitions. The evaluation metrics used are also consistent with the pairwise metrics used in subsequent related clustering, facilitating comparisons and analyses across different methods.

4.3.1 Correlation Clustering Evaluation

To evaluate the correlation clustering results, and to ensure comparability with related work, I chose a set of evaluation metrics for the experimental setup. After completing the SwAV-based self-supervised correlation clustering pipeline, I report metrics at two different levels: **pairwise metrics** and **cluster-level metrics**.

(1) Pairwise metrics (Join / Cut). Ground-truth pairwise relations come from class labels. Pairs with $y_i = y_j$ are labeled as *join*, while the rest are labeled as *cut*. And predicted pairwise relations are obtained from the cosine similarity matrix using a margin m, where a pair is classified as *join* if its cosine similarity exceeds the margin. Based on these definitions, I compute the following pairwise decision metrics over all unordered sample pairs (i, j):

- **Pairwise accuracy (ACC)**:

$$ACC = \frac{TP + TN}{\binom{N}{2}}, \tag{21}$$

where TP denotes the number of pairs that are correctly predicted as *join*, and TN denotes the number of pairs that are correctly predicted as *cut*.

- **Precision and recall for Join (PJ, RJ)** and **for Cut (PC, RC)**:

$$PJ = \frac{TP_J}{TP_J + FP_J}, \quad RJ = \frac{TP_J}{TP_J + FN_J}, \tag{22}$$

$$PC = \frac{TP_C}{TP_C + FP_C}, \quad RC = \frac{TP_C}{TP_C + FN_C}, \tag{23}$$

where TP_J denotes the number of pairs that are correctly predicted as *join*, FP_J denotes pairs that are incorrectly predicted as *join* (false joins), and FN_J denotes pairs that should be predicted as *join* but are incorrectly predicted as *cut*. Correspondingly, TP_C, FP_C, and FN_C denote true positives, false positives, and false negatives for *cut* decisions, respectively.

- **F1 scores for Join and Cut (F1J, F1C)**:

$$F1J = \frac{2 \cdot PJ \cdot RJ}{PJ + RJ}, \quad F1C = \frac{2 \cdot PC \cdot RC}{PC + RC}. \tag{24}$$

(2) Cluster-level metrics. After obtaining the final clustering assignment $\hat{c}_i$ for each sample, I further assess the agreement between the predicted clusters and the ground-truth labels. When labels are available, I report the following metrics:

- **Rand Index (RI)**:

$$RI = \frac{TP + TN}{\binom{N}{2}}, \tag{25}$$

where TP counts sample pairs that belong to the same ground-truth class and are assigned to the same cluster, and TN counts pairs that belong to different ground-truth classes and are assigned to different clusters.

- **Variation of Information (VI)**:

$$VI(U,V) = H(U) + H(V) - 2I(U;V), \tag{26}$$

where U denotes the ground-truth partition and V denotes the predicted clustering. Lower VI values indicate higher agreement between the two partitions. In addition, I report Purity, Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI) as supplementary clustering metrics, to improve interpretability.

- **Purity**:

$$\text{Purity} = \frac{1}{N} \sum_k \max_j |C_k \cap G_j|. \tag{27}$$

- **Normalized Mutual Information (NMI)**:

$$\text{NMI}(U,V) = \frac{2I(U;V)}{H(U) + H(V)}. \tag{28}$$

- **Adjusted Rand Index (ARI)**:

$$\text{ARI} = \frac{\text{RI} - E[\text{RI}]}{\max(\text{RI}) - E[\text{RI}]}. \tag{29}$$

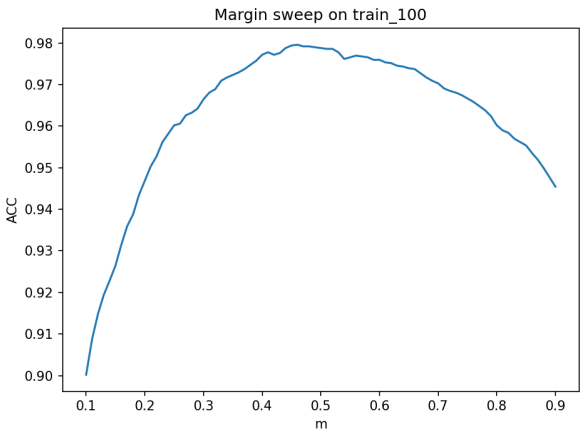


Figure 1: Accuracy on Train-100 as a function of the margin m.

4.4 Qualitative Evaluation-Visualization Settings

To complement the quantitative assessment results, I generated a series of visual figures to provide a more intuitive understanding of the key choices and results presentation throughout the training process.

First, for all labeled pairwise analysis sections, I plotted a histogram of cosine similarities for all unordered sample pairs to visualize the distribution. This also indirectly verifies the effectiveness of the learning in SWAV.

Secondly, I presented the relationship between the selected pairwise evaluation metrics (such as accuracy) and the parameter m by drawing line graphs. These line graphs demonstrated the sensitivity of the similarity threshold for pairwise decisions and provided support for the parameter selection process.

Thirdly, for the clustering results on the labeled dataset, I generated an image mosaic of the clusters, arranging multiple images from the same predicted cluster in a grid. These visualization effects help to check the consistency within the clusters and common error situations. Additionally, I created two-row comparison charts to align the original image arrangements with the predicted cluster assignments, enabling a direct comparison between the true labels and the clustering results.

Finally, for the unlabeled dataset, I also generated a mask map by clustering to qualitatively evaluate the structure of the discovered clusters in the absence of true labels.

5 Results

5.1 Margin Selection on Train-100

To select the similarity margin m used for correlation clustering, I perform a parameter sweep on the Train-100 dataset. Specifically, I evaluate multiple values of m in the range $[0.10, 0.90]$. For each value, join / cut decisions are calculated from cosine similarity, and the corresponding pairwise accuracy (ACC) is measured.

According to the figure 1, as m increases, the pairwise accuracy first increases rapidly, reaches a maximum around $m \approx 0.45$, and then gradually decreases. The curve shows a clear single peak. Based on this result, I fix the similarity margin $m = 0.46$ in all subsequent experiments.

5.2 Quantitative Results on Labeled Data

Table 1 presents the quantitative results of pairwise classification and related clustering based on the SwAV embedding representation. The similarity threshold is set to $m = 0.46$, selected from Train-100 and uniformly applied to Test-100, Test-30, and the mixed dataset Test-100/30.

The experimental results show that when combined with related clustering methods, the learned embedding representations can meaningfully partition the collection of organoid images. On Test-100, SwAV+ILP achieves strong agreement with the ground truth (RI 99.1, VI 0.13), comparable to the best supervised baseline (TNIa). When this method was applied to the organoid categories that were not observed during training (Test-30), clustering performance declined slightly but remained stable overall. The results surpassed all the supervised learning reference methods in multiple indicators. This indicates that the representations learned by SwAV are robust to previously unseen morphological changes. The results on the comprehensive dataset (Test-100/30) also reflect this trend accordingly. Overall, these observations suggest that the embedding space of SwAV can capture the key visual structures in tissue culture images and provide a reasonable and effective foundation for subsequent similarity modeling and clustering.

5.3 Visualization Results

In the labeled splits, the arrangement of clusters follows the original read order of the dataset. Specifically, clusters are first aggregated in sequence according to the data reading order; subsequently, in each cluster, the category with the largest proportion and the most prominent position in the original order is taken as the representative category, and the sorting order of the clusters is determined accordingly. Ultimately, each cluster is filled into

Test data	Model	Linear (pairwise)			Correlation clustering (A)				
		ACC	F1C	F1J	num_clusters	RI	VI	F1C	F1J
100	**SwAV + ILP**	97.2	98.3	84.1	11	99.1	0.13	98.9	89.2
	PQAP	–	–	–	–	97.2	0.68	98.5	83.7
	TNIa	–	–	–	–	99.4	0.14	99.7	96.9
	TNI	–	–	–	–	98.2	0.38	99.0	90.4
	TNH	–	–	–	–	93.0	1.58	96.2	60.9
	dH	–	–	–	–	91.6	1.76	95.4	52.1
30	**SwAV + ILP**	91.3	93.5	85.2	4	92.4	0.38	97.2	93.7
	PQAP	–	–	–	–	77.2	0.79	80.9	73.2
	TNIa	–	–	–	–	80.2	1.16	85.6	68.6
	TNI	–	–	–	–	72.9	1.20	78.9	62.2
	TNH	–	–	–	–	77.5	1.49	84.2	60.5
	dH	–	–	–	–	78.6	1.31	84.5	65.4
100/30	**SwAV + ILP**	97.4	98.5	78.4	16	97.7	0.39	98.5	82.3
	PQAP	–	–	–	–	94.8	1.24	97.2	65.8
	TNIa	–	–	–	–	95.3	0.88	97.4	72.6
	TNI	–	–	–	–	94.7	1.09	97.1	68.5
	TNH	–	–	–	–	92.0	2.13	95.7	48.0
	dH	–	–	–	–	90.3	2.62	94.8	37.0

Table 1: Quantitative results. Linear (pairwise) metrics (ACC/F1C/F1J) are reported from the linear head on frozen SwAV embeddings. Correlation clustering metrics (A) report clustering quality (RI/VI), the number of clusters, and edge-level consistency summarized by F1 scores for cut (F1C) and join (F1J).Results of PQAP, TNIa, TNI, TNH and dH are taken from [5]

the visualization grid one by one in this order, without using samples from other categories for padding. The unlabeled split is shown using the same cluster-wise layout (no ground-truth ordering)

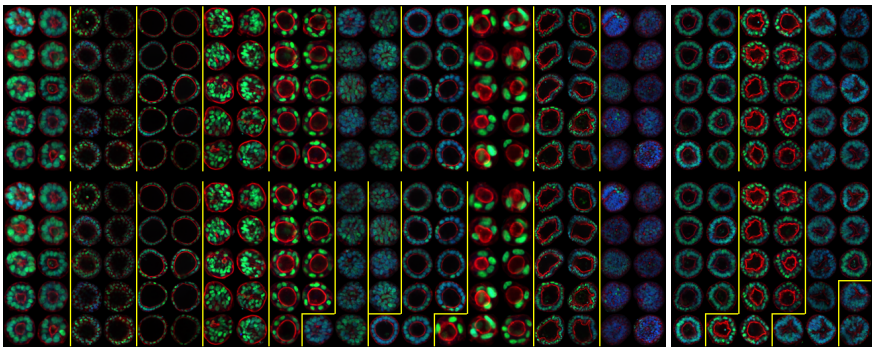
6 Discussion

Summary of Experimental Results. The experimental results show that by combining the learned embedding representations of SwAV with relevant clustering methods, meaningful visual features can be captured in single organoid images. The excellent performance achieved on Test-100 indicates that even in the absence of explicit supervision, the learned representation space still retains discriminative morphological features.

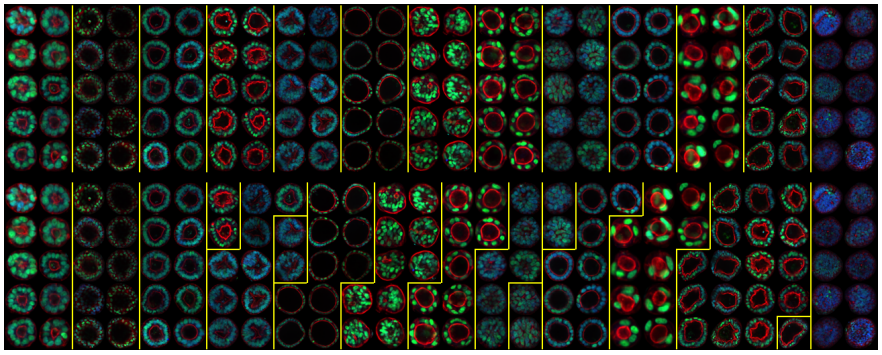
The observed performance decline on Test-30 reflects the increased morphological diversity brought about by the previously unseen organ types. Nevertheless, the clustering results remain generally stable, indicating that the learned embedding representations have a certain degree of robustness to unknown morphological changes. This phenomenon is also in line with intuition, as SwAV tends to capture global visual structures rather than highly relying on category-specific detailed features.

Transferability. Furthermore, this project demonstrates that the self-supervised representation learning method initially proposed on the ImageNet natural image dataset has certain transferability on single-organoid images, providing a new research direction for organoid image analysis.

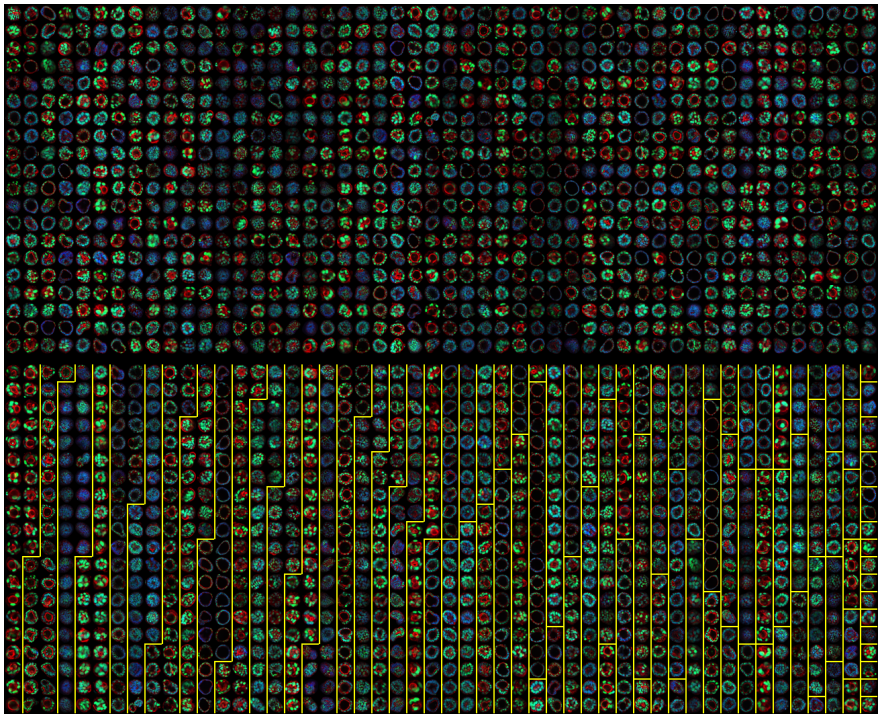
Engineering Challenges. However, during the actual training process, SwAV exhibited significant numerical instability on the organoid dataset. Unlike ImageNet, organoid images often contain large areas of low texture or nearly uniform color regions, which amplifies the distribution shift caused by random cropping and normalization operations, resulting in frequent oscillations or even numerical explosion in the loss during training. To alleviate this problem, multiple adjustments to the training strategy are needed, including reducing the scale range of random cropping, increasing the crop size, delaying the intervention speed of prototypes, and extending the prototype freezing stage. Additionally, in mixed-precision training, although the overall computation uses FP16 to improve training efficiency, FP32 calculation is retained in steps that are numerically sensitive, such as logarithmic operations, normalization, and Sinkhorn normalization, which is particularly crucial for avoiding nu-



(a) Test-100 (b) Test-30



(c) Test-130



(d) Unlabeled (1000 images)

Figure 2: Visualization of correlation clustering results based on SwAV embeddings. (a) Test-100, (b) Test-30, and (c) Test-130 are shown using the same cluster-wise layout and rendered at a consistent height to facilitate comparison. (d) The unlabeled split is displayed using the same layout but rendered at full width for readability.

merical instability. Despite these adjustments, on larger datasets, training still struggles to be completely stable after 150 epochs, and there are significant differences compared to the results of 200 epochs. In contrast, when reproducing the experiment on smaller datasets, numerical explosions and NaN issues are significantly reduced, indicating that this problem intensifies to some extent with increasing data size.

Limitations and Bottlenecks. Another limitation of the current method is that the similarity boundary parameters are fixed and selected based on Train-100. Although this strategy avoids information leakage, its optimality cannot be guaranteed when the organoid images have strong heterogeneity and unclear category boundaries. Moreover, the cosine-similarity-based related clustering method used in this experiment relies on pairwise consistency constraints. Its exact solution is limited by the size of the solution space and computational resources as the dataset size increases, because the number of pairwise variables and triangle-inequality constraints grows rapidly. Therefore, exact ILP solving is only feasible for small splits, while larger datasets require approximate solvers on sparse kNN graphs, which may reduce optimality and introduce sensitivity to graph construction.

7 Conclusion

Overall, this study verified the feasibility of SwAV for the organoid image clustering task. The experimental results showed that the representations learned with SwAV not only adapted well to clustering tasks involving existing categories but also demonstrated some generalization when dealing with new types of organoids.

However, due to time and experimental resource constraints, this experiment did not conduct sufficient hyperparameter tuning for the model. Therefore, it is still impossible to determine whether performance can be further improved with more refined parameter settings. Additionally, in the current method, the correlation between samples is mainly directly calculated based on the similarity in the embedding space. Another possible improvement direction is to introduce a learnable pairwise similarity model, such as by training a linear pairwise classification head to model the relationships between samples, thereby obtaining a more flexible and comprehensive correlation estimation. However, this direction has not been verified in this study.

Future work can further expand in the aforementioned directions, including more systematic parameter exploration and more flexible similarity modeling methods, to better extract and utilize the complex, continuous morphological changes in organ and tissue images.

References

[1] ASSRAN, M., CARON, M., MISRA, I., BOJANOWSKI, P., BORDES, F., VINCENT, P., JOULIN, A., RABBAT, M., AND BALLAS, N. Masked Siamese Networks for Label-Efficient Learning. In *Computer Vision – ECCV 2022* (Cham, 2022), S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., Springer Nature Switzerland, pp. 456–473.

[2] CARON, M., MISRA, I., MAIRAL, J., GOYAL, P., BOJANOWSKI, P., AND JOULIN, A. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems* (2020), H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., pp. 9912–9924.

[3] CHEN, T., KORNBLITH, S., NOROUZI, M., AND HINTON, G. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning* (13–18 Jul 2020), H. D. III and A. Singh, Eds., vol. 119 of *Proceedings of Machine Learning Research*, PMLR, pp. 1597–1607.

[4] HE, K., CHEN, X., XIE, S., LI, Y., DOLLÁR, P., AND GIRSHICK, R. Masked autoencoders are scalable vision learners, 2021.

[5] PRESBERGER, J., KESHARA, R., STEIN, D., KIM, Y. H., GRAPIN-BOTTON, A., AND ANDRES, B. Correlation clustering of organoid images, 2024.

[6] PRESBERGER, J., KESHARA, R., STEIN, D., KIM, Y. H., GRAPIN-BOTTON, A., AND ANDRES, B. Correlation clustering of organoid images: Data. Course project dataset, 2024.