

STR-DC: Spatio-Temporal Reliability Learning for Robust Deep Clustering of Japanese Handwritten Characters

Yang Liu

Harbin Normal University

Ziyang Li

ZL04738520@gmail.com

Northeast Agricultural University

Article

Keywords: deep clustering, Japanese handwritten characters, pseudo-label learning, confidence calibration, spatio-temporal reliability, feature stability

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11022648/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

STR-DC: Spatio-Temporal Reliability Learning for Robust Deep Clustering of Japanese Handwritten Characters

Yang Liu¹ and Ziyang Li^{2*}

¹College of Eastern Language and Culture, Harbin Normal University,
Harbin, 150025, Heilongjiang, China.

^{2*}College of Art, Northeast Agricultural University, Harbin, 150030,
Heilongjiang, China.

*Corresponding author(s). E-mail(s): ZL0473852@gmail.com;
Contributing authors: liuy53327@gmail.com;

Abstract

Clustering Japanese handwritten characters can organize unlabeled collections for subsequent annotation and retrieval, but visually similar categories, diverse writing styles, incomplete strokes, and degraded images can produce unreliable pseudo-labels during self-training. Existing calibrated deep clustering methods reduce overconfidence through calibrated confidence; however, this signal reflects only the current prediction space. We propose STR-DC, a Spatio-Temporal Reliability-Guided Deep Clustering framework that combines calibrated semantic confidence with cross-epoch feature stability. STR-DC first performs class-wise calibrated selection to obtain semantic candidates and then measures temporal reliability by comparing each current weak-view feature with its predecessor from the previous epoch. A strict fusion rule retains only samples satisfying both criteria, reducing the influence of confident but dynamically unstable pseudo-labels. On K49, ETL8-Hira75, and ETL9-Hira71, STR-DC achieves ACC scores of 57.68%, 78.46%, and 82.11%, respectively, exceeding the calibrated-confidence baseline by 5.52%, 9.50%, and 1.83%. Experiments on four general image benchmarks further demonstrate the generalization ability of STR-DC. Ablation and visualization results indicate that calibrated semantic confidence and temporal feature stability provide complementary evidence for reliable pseudo-label selection.

Keywords: deep clustering, Japanese handwritten characters, pseudo-label learning, confidence calibration, spatio-temporal reliability, feature stability

1 Introduction

Organizing Japanese handwritten character collections by character identity is useful for browsing, retrieving similar examples, and preparing data for subsequent annotation. When instance-level labels are unavailable, manually assigning each image to a character category can require substantial effort, particularly for unfamiliar or ambiguous handwritten forms. Unsupervised clustering offers a way to group these samples before manual labeling, providing a basis for inspecting related character images together. Clustering has been widely studied in image analysis and pattern recognition [1–4], but Japanese handwriting poses a particular difficulty: samples of the same character can differ across writers, while different characters may share local stroke structures. Faint or incomplete strokes can further obscure the distinctions between categories. The objective is therefore to recover groups corresponding to character identity despite variations in appearance. Recent deep clustering methods combine representation learning, pseudo-label generation, and clustering optimization [5–7]. A common pipeline first learns visual representations through self-supervised pre-training, then generates pseudo-labels from current predictions, and finally uses these pseudo-labels to refine the model. In this setting, however, early confusion between similar character shapes can produce incorrect pseudo-labels that are repeatedly reinforced, leading to confirmation bias and incorrect cluster boundaries [8–13].

Calibrated Deep Clustering Network (CDC) addresses this issue by introducing confidence calibration into deep clustering. It uses a calibration branch and a dynamic sample-capacity strategy to select high-confidence pseudo-labels [14]. The key insight of CDC is that clustering models should not only predict cluster assignments, but also estimate whether the predictions are trustworthy. However, high confidence from a single forward pass is still insufficient. In Japanese handwritten character images, samples with incomplete or faint strokes, unusual writing styles, or structures shared by visually similar characters may still receive deceptively high probabilities. Such samples appear certain in the current prediction snapshot but may not provide reliable semantic evidence for cluster formation. Thus, CDC mitigates miscalibration but can still suffer from softmax-induced and overfitting-induced high-confidence false positives [15–18].

From another perspective, instead of relying solely on confidence estimated from the current prediction snapshot, recent studies have explored sample reliability from the viewpoint of training dynamics. Along this direction, Li et al. proposed the LFSS framework, which provides a complementary temporal cue for reliable sample selection. LFSS shows that the representation stability of the same sample across consecutive epochs is strongly correlated with clustering correctness; therefore, cross-epoch feature similarity can be used to identify unstable samples [19]. This idea is consistent with studies on memorization and training dynamics, which show that typical, long-tail, noisy, and hard examples often follow different learning trajectories [20, 21]. Nevertheless, LFSS alone is also imperfect. High stability does not necessarily imply semantic value. In handwritten character clustering, a faint or incomplete character may yield a stable representation that lacks the strokes needed to distinguish its identity. Similarly, a character can remain consistently assigned to an incorrect

cluster whose members share a similar shape. These cases motivate checking semantic confidence together with temporal stability before treating a sample as reliable pseudo-label supervision.

These limitations motivate us to characterize pseudo-label reliability for Japanese handwritten character clustering from both semantic and temporal perspectives. CDC-style calibrated confidence provides a semantic reliability signal in the current prediction space, but it may still include high-confidence false positives. In contrast, LFSS-style feature stability provides a temporal reliability signal along the training trajectory, but it may retain samples that are stable yet semantically weak. Motivated by this consideration, we propose STR-DC, a Spatio-Temporal Reliability-Guided Deep Clustering framework for Japanese handwritten characters. STR-DC first uses class-wise calibrated selection to obtain semantically confident candidates within each predicted cluster. It then introduces cross-epoch feature stability to further examine whether these candidates remain stable during training by measuring the cosine similarity between the current weak-view features and their predecessor features from the previous epoch. Finally, a strict reliability fusion rule retains only samples that satisfy both calibrated semantic confidence and temporal stability for pseudo-label-based self-training. This design aims to limit the reinforcement of ambiguous character assignments by requiring both category-level evidence and consistent feature evolution.

2 Related Work

2.1 Self-Supervised Representation Learning

Self-supervised representation learning provides the visual foundation for modern deep clustering. MoCo-v2 improves instance discrimination through contrastive learning [22]. BYOL learns representations by training an online network to predict the target-network representation of another augmented view, without using negative pairs [23]. Barlow Twins reduces redundancy between embedding dimensions, while VICReg combines invariance, variance, and covariance regularization to prevent representation collapse [24, 25]. These methods learn strong generic features, but they usually do not decide which pseudo-labeled samples should be trusted during clustering self-training. For Japanese handwritten characters, the learned features must support grouping by character identity despite differences in writing style. Reliability-aware sample selection is therefore needed to prevent ambiguous character assignments from dominating self-training.

2.2 Deep Image Clustering

Deep image clustering couples representation learning with cluster assignment. DeepCluster and IIC establish early paradigms based on deep embedding optimization, alternating pseudo-label training, and mutual-information maximization [6, 7]. SCAN, Contrastive Clustering (CC), Graph Contrastive Clustering (GCC), IDFD, ProPos, and CoNR improve clustering structures through neighborhood consistency, semantic pseudo labels, instance- and cluster-level contrast, feature decorrelation, prototype

constraints, and contextual neighborhood refinement [8, 9]. Recent studies further investigate supervision sources and hard samples: TAC uses external semantic guidance, IDC mines valuable boundary samples through limited interaction, ConMix addresses long-tailed clustering by representation-level mixup, and ULDC models clustering from an uncertainty perspective [10–13]. Within these advances, pseudo-label reliability and sample stability are especially relevant. CDC shows that confident pseudo labels should be calibrated, but its probability signal may still include high-confidence false positives [14]. LFSS shows that sample stability reflects training dynamics, but stability alone may retain low-information samples without semantic value [19]. Our work addresses these limitations in Japanese handwritten character clustering, where similar stroke patterns can cause both confident errors and stable but incorrect assignments, by combining spatial confidence and temporal stability.

2.3 Confidence Calibration

Confidence calibration studies show that neural network probabilities can be misaligned with true correctness. Guo et al. systematically demonstrated that modern neural networks can be overconfident on wrong predictions [15], and later work revisited calibration under recent architectures and training recipes [16]. Logit normalization, polynomial alternatives to softmax, and gradient-decay analysis in large-margin softmax further explore ways to reduce overconfidence and improve predictive conservativeness [17]. Prediction-error-based classification also suggests that non-softmax reliability signals can reduce the dependence on inter-class relative probability competition [18]. Calibration is more difficult in deep clustering because no validation labels are available to correct probability estimates. CDC is an important step toward calibrated unsupervised clustering [14]. However, calibrated confidence still describes certainty at the current prediction snapshot. For Japanese handwriting, shared stroke patterns or missing discriminative strokes may still lead to confident but incorrect character assignments. We therefore combine confidence calibration with temporal training evidence.

2.4 Advantages of This Method

The main idea of our method is to use the complementarity between CDC and LFSS to improve pseudo-label reliability in Japanese handwritten character clustering. The calibrated confidence component selects candidates with strong evidence for a predicted character cluster, but it may still admit errors caused by similar character shapes or incomplete strokes. The feature stability component checks whether each candidate’s representation remains consistent across epochs, but stability alone cannot establish character identity. STR-DC retains candidates satisfying both conditions. This joint selection aims to reduce the influence of ambiguous character samples and provide more reliable supervision for organizing different handwritten forms into character-level groups.

3 Preliminaries

3.1 Problem formulation

Given an unlabeled training dataset $\mathcal{D}_u = \{x_i\}_{i=1}^N$ of Japanese handwritten character images, the goal is to assign each sample x_i to one of C predefined clusters corresponding to character categories without using ground-truth labels during training. The desired grouping places different handwritten forms of the same character together while separating characters with similar stroke structures. The proposed framework takes $\mathcal{D}_u$ as input and learns a feature extractor, a clustering head, and a calibration head. During training, calibrated semantic confidence and cross-epoch feature stability are jointly used to select reliable pseudo-labeled samples. After training, the model outputs the final cluster assignment y_i^{pred} for each sample. Cluster indices express the learned partition rather than named character transcriptions. The same formulation is also applied to general image datasets in the supplementary generalization evaluation.

3.2 Calibrated Deep Clustering

Calibrated Deep Clustering (CDC) [14] is a pseudo-label-based deep clustering method proposed to mitigate the overconfidence problem in self-training. Instead of directly trusting the raw confidence produced by the clustering head, CDC introduces a dual-head architecture, consisting of a clustering head and a calibration head. The calibration head estimates more reliable confidence scores for pseudo-label selection, while the clustering head is optimized using the selected pseudo-labels.

Given an unlabeled dataset $\mathcal{D}_u$, CDC first applies weak and strong augmentations to each sample, denoted as $\mathcal{W}(x_i)$ and $\mathcal{A}(x_i)$, respectively. A shared feature extractor $f(\Theta; \cdot)$ maps them into latent representations:

$$z_i^w = f(\Theta; \mathcal{W}(x_i)), \quad z_i^s = f(\Theta; \mathcal{A}(x_i)). \quad (1)$$

The weakly augmented view is used to estimate calibrated confidence and generate pseudo-labels, whereas the strongly augmented view is used to train the clustering head. The clustering head $g_{clu}(\theta_{clu}; \cdot)$ predicts the cluster distribution of the strongly augmented sample:

$$p_i^{s,clu} = \sigma(g_{clu}(\theta_{clu}; z_i^s)), \quad (2)$$

where $\sigma(\cdot)$ denotes the softmax function. Meanwhile, the calibration head $g_{cal}(\theta_{cal}; \cdot)$ produces the calibrated prediction for the weakly augmented sample:

$$p_i^{w,cal} = \sigma(g_{cal}(\theta_{cal}; z_i^w)). \quad (3)$$

Based on the calibrated prediction, CDC selects reliable pseudo-labeled samples in a class-wise manner. For each cluster c , samples are ranked according to $p_{i,c}^{w,cal}$, and the top $\lfloor B/C \rfloor$ samples form $\text{TOP}(c)$. The number of selected samples for cluster c is determined by

$$M(c) = \left| \sum_{x_i \in \text{TOP}(c)} p_{i,c}^{w,cal} \right|, \quad c = 1, 2, \dots, C, \quad (4)$$

where B is the batch size and C is the number of clusters. CDC then selects the $M(c)$ highest-ranked samples from each predicted cluster c to construct the pseudo-labeled set $\mathcal{S}$. The pseudo-label of each selected sample is assigned as

$$\hat{y}_i = \arg \max_c p_{i,c}^{w,cal}. \quad (5)$$

Given the selected pseudo-labeled set $\mathcal{S}$, the clustering head is optimized by the cross-entropy loss:

$$\mathcal{L}_{clu} = -\frac{1}{|\mathcal{S}|} \sum_{x_i \in \mathcal{S}} \log p_{i,\hat{y}_i}^{s,clu}. \quad (6)$$

The calibration head is trained by exploiting local consistency in the feature space. Specifically, CDC partitions the batch features into K mini-clusters $\{Q_k\}_{k=1}^K$ using K-means. For each mini-cluster Q_k , the average prediction of the clustering head is used as a locally smoothed target:

$$\hat{q}_k = \frac{1}{|Q_k|} \sum_{x_i \in Q_k} \sigma(g_{clu}(\theta_{clu}; f(\Theta; x_i))). \quad (7)$$

The calibration loss is then defined as

$$\mathcal{L}_{cal} = -\frac{1}{B} \sum_k \sum_{x_i \in Q_k} \hat{q}_k^\top \log p_i^{cal}, \quad (8)$$

where

$$p_i^{cal} = \sigma(g_{cal}(\theta_{cal}; f(\Theta; x_i))). \quad (9)$$

To prevent trivial cluster collapse, CDC further adds an entropy regularization term:

$$\mathcal{L}_{en} = \frac{1}{C} \sum_{j=1}^C \bar{p}_j^{cal} \log \bar{p}_j^{cal}, \quad \bar{p}_j^{cal} = \frac{1}{B} \sum_{i=1}^B p_{i,j}^{cal}. \quad (10)$$

Thus, the calibration objective is

$$\mathcal{L}_{cali} = \mathcal{L}_{cal} + w_{en} \mathcal{L}_{en}. \quad (11)$$

In summary, CDC uses the calibration head to provide more reliable confidence estimates for pseudo-label selection, and then uses the selected samples to train the clustering head. This calibrated pseudo-labeling mechanism provides an important basis for our subsequent reliability-aware sample selection strategy.

4 Proposed STR-DC Framework

4.1 Overview of the proposed framework

Although CDC provides calibrated pseudo-label selection, applying it to Japanese handwritten characters still requires distinguishing category evidence from visual

ambiguity. Characters sharing local stroke structures, as well as samples with faint or incomplete strokes, may receive high confidence under the current model. Since this confidence reflects a single prediction snapshot, it does not show whether the corresponding representation is consistently learned across epochs. Selecting such ambiguous samples can reinforce incorrect character assignments during self-training.

To address this limitation, we propose a *Spatio-Temporal Reliability-Guided Deep Clustering* framework for Japanese handwritten character clustering, termed STR-DC. The overall framework is shown in Fig. 1. Given an unlabeled character image x_i , STR-DC constructs weakly and strongly augmented views and uses them for reliability estimation and clustering optimization, respectively. The main idea is to select character images using two complementary components: the calibrated confidence component evaluates evidence for the predicted character cluster, and the feature stability component verifies consistency across training epochs.

As illustrated in Fig. 1, STR-DC estimates reliability from both spatial and temporal perspectives. Specifically, the spatial semantic reliability is estimated by the calibration head using the weakly augmented view $\mathcal{W}(x_i)$, which provides calibrated evidence for the predicted character cluster. Meanwhile, the temporal reliability is measured by comparing the current weak-view feature with its predecessor feature from the previous epoch, so as to evaluate whether the sample representation remains stable during training. These two reliability signals are then combined through the reliability fusion component, and only samples satisfying both semantic confidence and temporal stability are retained for self-training.

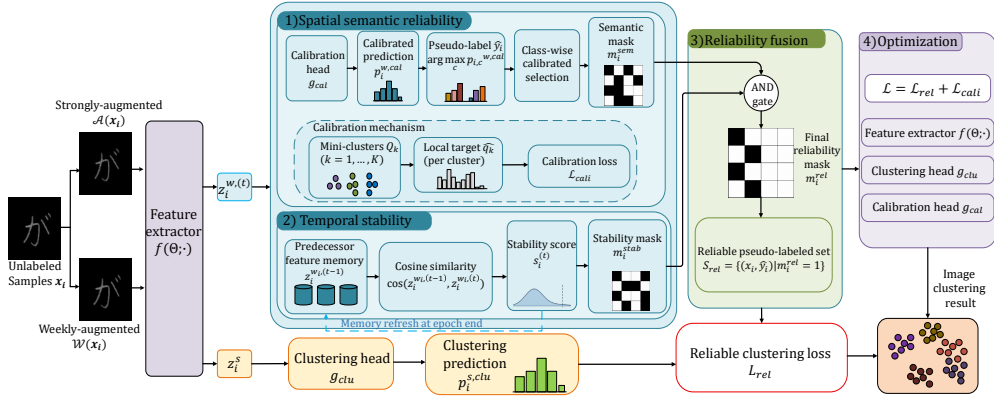


Fig. 1 Overview of STR-DC for Japanese handwritten character clustering. Given an unlabeled character image x_i , the calibrated confidence component and the feature stability component jointly select reliable pseudo-labeled samples for clustering optimization.

The strongly augmented view $\mathcal{A}(x_i)$ is used for clustering optimization. The reliable pseudo-labeled set obtained from the spatio-temporal reliability selection is used to supervise the clustering head, while the calibration objective is maintained to support reliable confidence estimation. In this way, STR-DC does not directly trust all calibrated high-confidence samples. Instead, it further filters them through cross-epoch

stability verification, thereby reducing the risk of learning from high-confidence but unstable pseudo-labels.

4.2 Spatial semantic reliability from calibrated confidence

As described in Section 3.2, CDC estimates the calibrated prediction of each weakly augmented sample through the calibration head, as shown in Eq. (3), and assigns the corresponding pseudo-label according to Eq. (5). Based on these calibrated probabilities, CDC further performs class-wise pseudo-label selection using Eq. (4).

In STR-DC, we retain the calibrated semantic selection process introduced in the preliminary section, and use its selected pseudo-labeled character images as the initial spatial semantic candidates. Here, semantic reliability refers to evidence for character-category membership, rather than the linguistic meaning of a character. Let $\mathcal{S}_{sem}$ denote the sample set retained by the calibrated class-wise selection:

$$\mathcal{S}_{sem} = \{(x_i, \hat{y}_i) \mid x_i \text{ is selected by Eq. (4)}\}. \quad (12)$$

For subsequent reliability filtering, we define the spatial semantic reliability mask as

$$m_i^{sem} = \mathbb{I}((x_i, \hat{y}_i) \in \mathcal{S}_{sem}), \quad (13)$$

where $\mathbb{I}(\cdot)$ is the indicator function. The set $\mathcal{S}_{sem}$ represents samples that are reliable from the perspective of calibrated semantic confidence. However, this reliability is still estimated from the current prediction space and therefore reflects only an instantaneous semantic judgment. It does not indicate whether the feature representation of the same sample remains stable during training. To address this limitation, STR-DC further introduces temporal feature stability as an additional reliability criterion to refine $\mathcal{S}_{sem}$ before using it for self-training.

4.3 Temporal reliability from cross-epoch feature stability

To refine the spatial semantic candidate set $\mathcal{S}_{sem}$, the feature stability component evaluates whether each candidate character image is consistently represented during training. A handwritten sample may be confidently assigned based on a shared stroke pattern even when its character identity remains ambiguous. Comparing its features across consecutive epochs supplies an additional check on this instantaneous judgment. A large representation change suggests that the current high-confidence assignment may not yet be reliable.

According to the weak-view feature definition in Eq. (1), the feature representation of x_i at epoch t is denoted as

$$z_i^{w,(t)} = f(\Theta^{(t)}; \mathcal{W}(x_i)), \quad (14)$$

where $\Theta^{(t)}$ denotes the feature extractor parameters at epoch t . At the end of each epoch, STR-DC stores the weak-view feature representation of each sample as a predecessor feature memory. Therefore, the feature representation of the same sample from the previous epoch is denoted as $z_i^{w,(t-1)}$.

The temporal stability score of x_i is measured by the cosine similarity between its current and predecessor representations:

$$s_i^{(t)} = \frac{\left(z_i^{w,(t)}\right)^\top z_i^{w,(t-1)}}{\left\|z_i^{w,(t)}\right\| \left\|z_i^{w,(t-1)}\right\|}. \quad (15)$$

A larger $s_i^{(t)}$ indicates that the representation of x_i changes only slightly between two consecutive epochs, suggesting that the sample has been learned in a more stable manner. In contrast, a smaller value indicates stronger feature drift, which implies that the sample may still be ambiguous or not yet reliably organized in the current feature space.

Instead of using a fixed stability threshold, STR-DC adopts an adaptive percentile-based threshold. For each mini-batch, the temporal stability threshold is computed as

$$\theta_{stab}^{(t)} = \text{Percentile} \left(\{s_i^{(t)}\}_{i=1}^B, \tau_{pct} \right), \quad (16)$$

where τ_{pct} is a percentile hyperparameter controlling the proportion of samples with low temporal stability to be filtered out. The temporal stability mask is then defined as

$$m_i^{stab} = \mathbb{I} \left(s_i^{(t)} \geq \theta_{stab}^{(t)} \right). \quad (17)$$

This adaptive design allows the stability criterion to change with the training stage and the feature drift distribution of the current mini-batch, rather than relying on a manually fixed similarity threshold. This is useful for handwritten character collections whose feature variation may differ with writing style and stroke visibility.

4.4 Spatio-temporal reliability-guided sample selection

After obtaining the spatial semantic reliability mask m_i^{sem} from the calibrated class-wise selection in Eq. (4) and the temporal stability mask m_i^{stab} from the cross-epoch feature comparison, STR-DC selects the final reliable pseudo-labeled samples through a strict conjunctive rule:

$$m_i^{rel} = m_i^{sem} \wedge m_i^{stab}. \quad (18)$$

The final reliable pseudo-labeled set is then defined as

$$\mathcal{S}_{rel} = \{(x_i, \hat{y}_i) \mid m_i^{rel} = 1\}. \quad (19)$$

This spatio-temporal selection rule is the key mechanism of STR-DC. The calibrated confidence component checks whether a candidate has sufficient evidence for its predicted character cluster, while the feature stability component checks whether its representation remains consistent during training. A visually ambiguous character with high confidence but unstable features is excluded, as is a stable sample lacking calibrated category evidence. Their intersection therefore provides character images with both semantic and temporal support for self-training.

4.5 Optimization objective

For each sample retained in $\mathcal{S}_{rel}$, the pseudo-label $\hat{y}_i$ is generated from the weakly augmented view through Eq. (5). The strongly augmented view $\mathcal{A}(x_i)$ is then used to optimize the clustering head. According to Eq. (1), the corresponding strong-view feature is $z_i^s = f(\Theta; \mathcal{A}(x_i))$, and the clustering prediction $p_i^{s,clu}$ is obtained from Eq. (2). The following loss is evaluated for mini-batches with $|\mathcal{S}_{rel}| > 0$.

Based on the original CDC clustering loss in Eq. (6), STR-DC replaces the CDC-selected set with the spatio-temporally reliable set $\mathcal{S}_{rel}$. The reliable clustering loss is therefore defined as

$$\mathcal{L}_{rel} = -\frac{1}{|\mathcal{S}_{rel}|} \sum_{(x_i, \hat{y}_i) \in \mathcal{S}_{rel}} \log p_{i, \hat{y}_i}^{s,clu}. \quad (20)$$

Thus, $\mathcal{L}_{rel}$ is not a new supervision form, but a reliability-guided refinement of Eq. (6). The clustering head is updated only using samples that have passed both calibrated semantic confidence selection and temporal stability verification.

The calibration head is optimized using the calibration objective defined in Eq. (11). Therefore, the overall objective of STR-DC is formulated as

$$\mathcal{L} = \mathcal{L}_{rel} + \lambda_{cali} \mathcal{L}_{cali}, \quad (21)$$

where λ_{cali} is a trade-off parameter. The first term encourages the clustering head to learn from spatio-temporally purified pseudo-labels under strong augmentation, while the second term preserves the calibration ability of the calibration head. Through this joint objective, STR-DC learns to separate handwritten character categories using the selected pseudo-labels while maintaining calibrated confidence estimation.

4.6 Training procedure

The training procedure of STR-DC is designed to progressively refine the pseudo-labeled set of Japanese handwritten character images by integrating calibrated semantic confidence and cross-epoch feature stability. Specifically, STR-DC first constructs a spatial semantic candidate set $\mathcal{S}_{sem}$ based on calibrated confidence, and then further verifies the temporal reliability of these candidate samples before using them for self-training. In this way, only samples that are both semantically confident and temporally stable are retained for optimizing the clustering head.

Before iterative training, the feature extractor $f(\Theta; \cdot)$, the clustering head $g_{clu}(\theta_{clu}; \cdot)$, and the calibration head $g_{cal}(\theta_{cal}; \cdot)$ are initialized. Then, STR-DC performs a full forward pass over the unlabeled training set $\mathcal{D}_u$ and stores the weak-view feature representation $z_i^{w,(0)}$ of each sample as the initial predecessor feature memory. This memory provides the temporal reference for measuring feature stability in the subsequent epoch.

At epoch t , for each mini-batch, the weakly augmented view $\mathcal{W}(x_i)$ is first used to obtain the calibrated prediction $p_i^{w,cal}$ and the pseudo-label $\hat{y}_i$. Based on the calibrated class-wise selection strategy, STR-DC constructs the spatial semantic candidate set $\mathcal{S}_{sem}$ and the corresponding semantic reliability mask m_i^{sem} . Meanwhile, the current weak-view feature $z_i^{w,(t)}$ is compared with its predecessor feature $z_i^{w,(t-1)}$ to compute

the temporal stability score $s_i^{(t)}$. The adaptive threshold $\theta_{stab}^{(t)}$ is then used to obtain the temporal stability mask m_i^{stab} . Finally, STR-DC combines m_i^{sem} and m_i^{stab} to obtain the final reliability mask m_i^{rel} , and only samples satisfying $m_i^{rel} = 1$ are retained in the reliable pseudo-labeled set $\mathcal{S}_{rel}$.

For model optimization, the strongly augmented view $\mathcal{A}(x_i)$ of each sample in $\mathcal{S}_{rel}$ is fed into the feature extractor and clustering head to produce $p_i^{s,clu}$. The reliable clustering loss $\mathcal{L}_{rel}$ is computed only on $\mathcal{S}_{rel}$, while the calibration head is optimized by the calibration objective $\mathcal{L}_{cali}$. The whole model is updated by minimizing the overall objective $\mathcal{L}$. After all mini-batches in the current epoch are processed, the predecessor feature memory is refreshed using the latest weak-view features, so that temporal stability in the next epoch can be evaluated with updated historical representations. The detailed procedure is summarized in Algorithm 1.

After training, the final cluster assignment of each sample is obtained from the clustering head:

$$y_i^{pred} = \arg \max_c \sigma(g_{clu}(\theta_{clu}; f(\Theta; x_i)))_c. \quad (22)$$

4.7 Time complexity analysis

We analyze the time complexity of STR-DC according to the training procedure in Algorithm 1. Let N denote the number of training samples, B the mini-batch size, C the number of clusters, and d the dimension of the latent feature representation. Let K denote the number of mini-clusters used by the calibration objective, and let I_K denote the number of K-means iterations. We use $\mathcal{C}_{net}$ to represent the per-sample cost of the network forward and backward computations during mini-batch training, and $\mathcal{C}_{fwd}$ to represent the per-sample cost of the additional forward pass used to refresh the predecessor feature memory.

For each mini-batch, the forward and backward computations of the feature extractor, clustering head, and calibration head require $\mathcal{O}(B\mathcal{C}_{net})$. Computing calibrated predictions, pseudo-labels, clustering predictions, and probability-based losses over C clusters costs $\mathcal{O}(BC)$. The calibrated class-wise selection ranks the samples for each cluster and therefore requires $\mathcal{O}(CB \log B)$ under a sorting-based implementation. In addition, the K-means procedure used by the calibration objective has a complexity of $\mathcal{O}(I_K B K d)$.

The additional computation introduced by STR-DC mainly arises from temporal reliability estimation. Computing the cosine similarity between the current and predecessor features requires $\mathcal{O}(Bd)$ per mini-batch. Computing the percentile-based stability threshold costs $\mathcal{O}(B \log B)$ using sorting, whereas constructing the semantic, stability, and final reliability masks requires only $\mathcal{O}(B)$. Therefore, after absorbing the lower-order terms $\mathcal{O}(BC)$, $\mathcal{O}(B \log B)$, and $\mathcal{O}(B)$ into the dominant terms, the time complexity of STR-DC for one mini-batch can be written as $\mathcal{O}(B(\mathcal{C}_{net} + I_K K d + C \log B + d))$. Since each epoch contains approximately $\lceil N/B \rceil$ mini-batches, the mini-batch training complexity per epoch is $\mathcal{O}(N(\mathcal{C}_{net} + I_K K d + C \log B + d))$. At the end of each epoch, STR-DC refreshes the predecessor feature memory by performing an additional forward pass over all training samples. This operation requires $\mathcal{O}(N\mathcal{C}_{fwd})$. Consequently, the overall time complexity

Algorithm 1: Training procedure of STR-DC

Input: Unlabeled dataset $\mathcal{D}_u = \{x_i\}_{i=1}^N$; number of clusters C ; maximum epoch T ; batch size B ; percentile parameter τ_{pct} ; trade-off parameter λ_{cal} .

Output: Final cluster assignments $\{y_i^{pred}\}_{i=1}^N$.

Initialize $f(\Theta; \cdot)$, $g_{clu}(\theta_{clu}; \cdot)$, and $g_{cal}(\theta_{cal}; \cdot)$;

Initialize the predecessor feature memory by storing $z_i^{w,(0)}$ for all $x_i \in \mathcal{D}_u$;

for $t = 1$ **to** T **do**

for each mini-batch sampled from $\mathcal{D}_u$ **do**

 Generate weakly augmented views $\mathcal{W}(x_i)$ and strongly augmented views $\mathcal{A}(x_i)$;

 Compute calibrated predictions $p_i^{w,cal}$ using the calibration head;

 Assign pseudo-labels $\hat{y}_i$ from $p_i^{w,cal}$;

 Construct the spatial semantic candidate set $\mathcal{S}_{sem}$ by calibrated class-wise selection;

 Obtain the semantic reliability mask m_i^{sem} ;

 Compute the current weak-view feature $z_i^{w,(t)}$;

 Retrieve the predecessor feature $z_i^{w,(t-1)}$ from the feature memory;

 Compute the temporal stability score $s_i^{(t)}$;

 Compute the adaptive stability threshold $\theta_{stab}^{(t)}$ using τ_{pct} ;

 Obtain the temporal stability mask m_i^{stab} ;

 Generate the final reliability mask $m_i^{rel} = m_i^{sem} \wedge m_i^{stab}$;

 Construct the reliable pseudo-labeled set $\mathcal{S}_{rel}$;

 Compute strong-view clustering predictions $p_i^{s,clu}$ for samples in $\mathcal{S}_{rel}$;

 Compute the reliable clustering loss $\mathcal{L}_{rel}$;

 Compute the calibration objective $\mathcal{L}_{cal}$;

 Update Θ , θ_{clu} , and θ_{cal} by minimizing $\mathcal{L}$;

 Refresh the predecessor feature memory by storing the latest weak-view features $z_i^{w,(t)}$ for all $x_i \in \mathcal{D}_u$;

Obtain final cluster assignments $\{y_i^{pred}\}_{i=1}^N$ from the clustering head;

for T training epochs is $\mathcal{O}(TN(\mathcal{C}_{net} + \mathcal{C}_{fwd} + I_K Kd + C \log B + d))$. The initialization of the predecessor feature memory before the first epoch incurs a one-time cost of $\mathcal{O}(N\mathcal{C}_{fwd})$, which does not change the asymptotic complexity over T epochs. Relative to the CDC training pipeline, STR-DC additionally performs cross-epoch feature comparison, percentile-based thresholding, and predecessor feature memory refresh. Therefore, its additional time complexity over CDC is $\mathcal{O}(TN(\mathcal{C}_{fwd} + d + \log B))$.

If the predecessor features are updated online during mini-batch training, the term $\mathcal{O}(TNC_{fwd})$ can be omitted. However, the cached features within the same epoch may then correspond to different network parameter states. The epoch-level memory refresh adopted in STR-DC provides a more consistent temporal reference at the cost of one additional forward pass per epoch.

5 Experiments and Analyses

5.1 Experimental settings

5.1.1 Evaluation metrics

Following common practice in deep clustering, we evaluate how closely the predicted character groups match the reference character categories using three widely used metrics: clustering accuracy (ACC), normalized mutual information (NMI), and adjusted

Rand index (ARI). All results are reported as percentages (%), and higher values indicate better clustering performance.

5.1.2 Datasets

We conduct the primary experiments on three Japanese handwritten character datasets, namely Kuzushiji-49 (K49) [29], ETL8-Hira75, and ETL9-Hira71 [30]. K49 is an imbalanced dataset containing 232,365 training images and 38,547 test images from 49 character classes. ETL8-Hira75 and ETL9-Hira71 comprise 75 and 71 Hiragana categories derived from ETL8 and ETL9, respectively. ETL8-Hira75 contains 9,600 training images and 2,400 test images, while ETL9-Hira71 contains 11,360 training images and 2,840 test images. All three datasets are resized to 64×64 pixels. Ground-truth labels are used only for evaluation. To evaluate whether STR-DC generalizes beyond Japanese handwritten characters, we additionally retain four commonly used general image clustering benchmarks: CIFAR-10 [26], CIFAR-20 [26], ImageNet-10 [28], and ImageNet-Dogs [28]. The dataset statistics are summarized in Table 1.

Table 1 Summary of Japanese handwritten character datasets and general image clustering benchmarks.

Dataset	Samples	Classes	Image size
K49	270,912	49	64×64
ETL8-Hira75	12,000	75	64×64
ETL9-Hira71	14,200	71	64×64
CIFAR-10	60,000	10	32×32
CIFAR-20	60,000	20	32×32
ImageNet-10	13,000	10	224×224
ImageNet-Dogs	19,500	15	224×224

5.1.3 Comparison methods

For the three Japanese handwritten character datasets, we compare STR-DC with MoCo-v2 [22], BYOL [23], LFSS [19], and CDC [14]. These baselines respectively represent contrastive representation learning, self-distillation, stability-guided clustering, and calibrated pseudo-label selection. All methods in the Japanese handwritten character experiments are trained for 100 epochs under the same evaluation protocol. For the generalization experiments, we retain the broader comparison methods used on the four general image benchmarks, including ProPos [8], CoNR [9], DMICC [35], BYOL, MoCo-v2, and LFSS. LFSS denotes the published baseline, whereas “Feature Stability Only” in Table 3 is an internal STR-DC variant that uses only the feature stability component for reliable sample selection.

Table 2 Clustering performance on three Japanese handwritten character datasets. All methods are trained for 100 epochs. The best results are highlighted in bold, and the second-best results are underlined.

Dataset	Metric	MoCo-v2	BYOL	LFSS	CDC	STR-DC
K49	ACC	44.34	46.04	35.98	<u>52.16</u>	57.68
	NMI	53.27	54.13	42.14	<u>67.46</u>	69.39
	ARI	29.94	32.17	21.08	<u>39.46</u>	44.92
ETL8-Hira75	ACC	64.33	18.21	39.96	<u>68.96</u>	78.46
	NMI	79.76	42.06	65.11	<u>90.51</u>	92.54
	ARI	53.23	6.67	28.55	<u>69.66</u>	76.49
ETL9-Hira71	ACC	65.60	18.27	46.80	<u>80.28</u>	82.11
	NMI	78.29	40.25	65.37	<u>91.97</u>	93.33
	ARI	53.55	7.02	33.60	<u>75.74</u>	79.47

5.2 Clustering performance

Table 2 reports the clustering results on K49, ETL8-Hira75, and ETL9-Hira71. STR-DC achieves the best ACC, NMI, and ARI on all three datasets, while CDC consistently ranks second. The consistent advantage across datasets with different numbers of character categories demonstrates the effectiveness of the proposed method for Japanese handwritten character clustering.

The improvements are particularly evident on K49 and ETL8-Hira75. On K49, STR-DC obtains 57.68% ACC, 69.39% NMI, and 44.92% ARI, outperforming CDC by 5.52, 1.93, and 5.46 percentage points, respectively. On ETL8-Hira75, STR-DC improves over CDC by 9.50 percentage points in ACC and 6.83 percentage points in ARI. STR-DC also increases the ARI on ETL9-Hira71 from 75.74% to 79.47%. These improvements, especially in ARI, indicate that STR-DC produces more consistent sample-to-cluster assignments. In contrast, the competing methods do not explicitly combine semantic confidence with temporal stability when selecting pseudo-labels and therefore show less consistent performance across the three datasets.

Overall, the results suggest that calibrated confidence and cross-epoch feature stability provide complementary evidence. The calibrated confidence component supplies semantic candidates, while the feature stability component removes dynamically unstable samples. Their joint use improves clustering of handwritten forms by character identity despite visual ambiguity.

5.3 Ablation study

Table 3 evaluates the calibrated confidence component and the feature stability component. Each single-component variant uses only the named component for reliable sample selection. The complete STR-DC achieves the best ACC, NMI, and ARI on all three datasets, showing that both components contribute to the final performance. The calibrated confidence component provides category-level evidence, whereas the

Table 3 Ablation study on three Japanese handwritten character datasets. The best results are highlighted in bold.

Dataset	Variant	ACC	NMI	ARI
K49	Calibrated Confidence Only	52.16	67.46	39.46
	Feature Stability Only	50.59	67.00	39.00
	STR-DC	57.68	69.39	44.92
ETL8-Hira75	Calibrated Confidence Only	68.96	90.51	69.66
	Feature Stability Only	31.00	66.91	22.07
	STR-DC	78.46	92.54	76.49
ETL9-Hira71	Calibrated Confidence Only	80.28	91.97	75.74
	Feature Stability Only	76.44	91.30	74.59
	STR-DC	82.11	93.33	79.47

feature stability component verifies whether the representation remains consistent across epochs.

On K49, the complete model improves upon the calibrated-confidence-only variant by 5.52, 1.93, and 5.46 percentage points in ACC, NMI, and ARI, respectively. The distinction between the two components is most evident on ETL8-Hira75: the feature-stability-only variant obtains 31.00% ACC, whereas the calibrated-confidence-only variant reaches 68.96%. Thus, feature stability is more effective as a verification component applied after semantic candidate selection than as an independent selection criterion.

Similar improvements are observed on ETL9-Hira71. Overall, the results support the AND fusion in Eq. (18): the calibrated confidence component selects semantically plausible candidates, and the feature stability component filters candidates with inconsistent representations.

5.4 Visualization analysis

Figs. 2–4 analyze the relationship between cross-epoch feature stability and clustering correctness from distributional and geometric perspectives. In all three stability-distribution plots, the curve for correctly clustered samples is shifted toward higher cosine-similarity values, while the incorrectly clustered samples occupy a wider low-stability tail. This common pattern supports the central assumption of STR-DC: samples whose representations change substantially between consecutive epochs are more likely to carry unreliable pseudo-labels. At the same time, the two distributions overlap over a considerable interval. Stability therefore provides useful reliability evidence but cannot by itself determine semantic correctness, which is consistent with the ablation results in Table 3.

The t-SNE visualizations in subfigure (b) of Figs. 2–4 provide a complementary geometric view. ETL8-Hira75 and ETL9-Hira71 contain many compact and well-separated local groups, while the K49 embedding includes more mixed, elongated, and partially overlapping structures. Low-stability samples frequently appear near apparent cluster boundaries, along transitions between groups, or in sparse regions

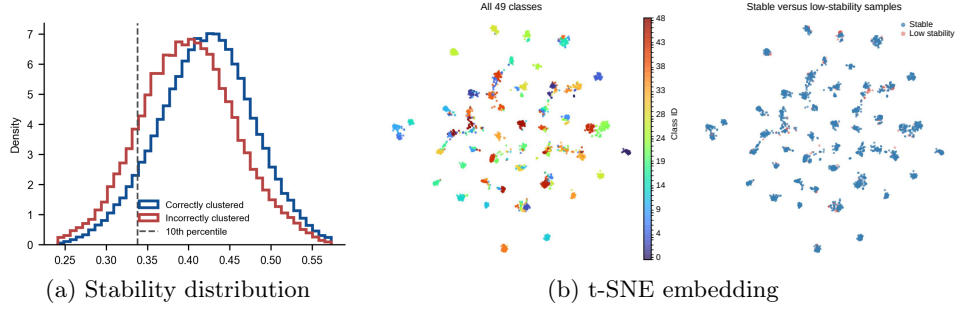


Fig. 2 Visualization results on K49. Subfigure (a) compares the cross-epoch feature-stability distributions of correctly and incorrectly clustered samples; the dashed line denotes the 10th-percentile threshold. Subfigure (b) shows the t-SNE embedding colored by ground-truth classes and by stable versus low-stability samples.

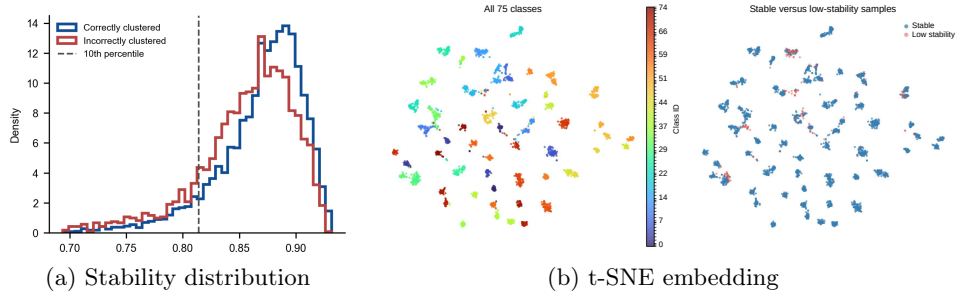


Fig. 3 Visualization results on ETL8-Hira75. The definitions of subfigures (a) and (b) are the same as those in Fig. 2.

surrounding dense cluster cores. These regions may contain samples with ambiguous cluster assignments. Low-stability points also occur inside several apparently compact groups, showing that temporal instability is not equivalent to being a simple geometric outlier. The feature stability component therefore contributes information that cannot be obtained from a single t-SNE snapshot alone.

Fig. 5 further compares samples from the high- and low-stability ends of the ranking. Highly stable samples generally contain visible and coherent stroke structures and maintain a consistent foreground pattern. Low-stability samples more frequently exhibit faint strokes, partial cropping, fragmented components, severe deformation, or ambiguous character shapes. However, a small number of visually clear characters are also assigned low stability, which indicates that visual quality alone does not fully explain the temporal signal. Taken together, the distribution plots, t-SNE embeddings, and sample gallery show that temporal stability identifies both visibly degraded inputs and less obvious representation-level ambiguities. When combined with calibrated confidence, this signal allows STR-DC to focus self-training on samples that are simultaneously semantically plausible and dynamically stable.

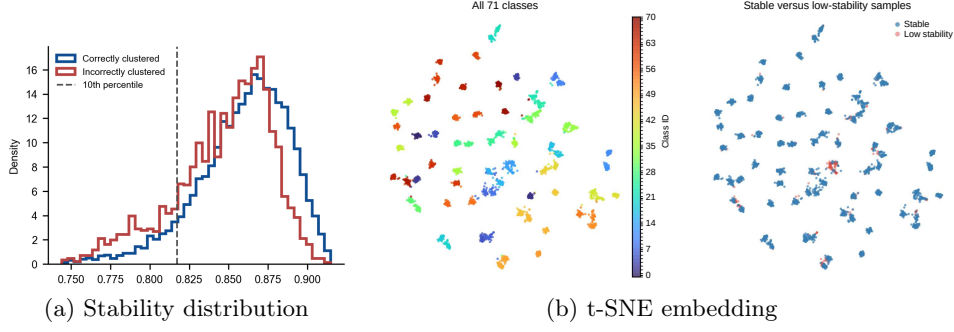


Fig. 4 Visualization results on ETL9-Hira71. The definitions of subfigures (a) and (b) are the same as those in Fig. 2.

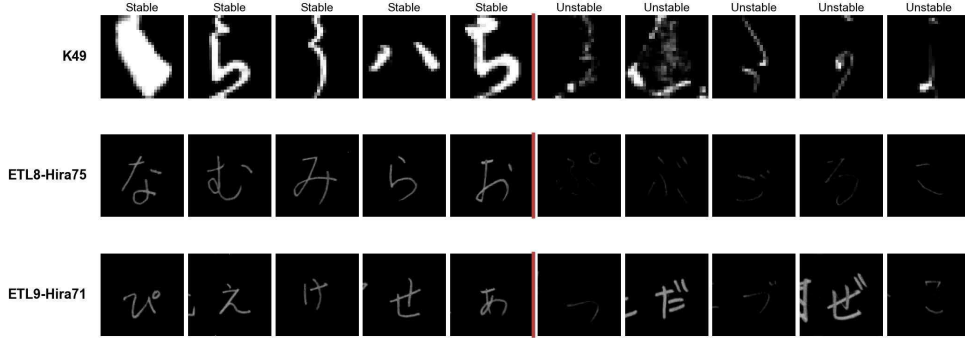


Fig. 5 Representative stable and unstable samples on K49, ETL8-Hira75, and ETL9-Hira71. For each dataset, the five samples on the left have high cross-epoch feature stability, while the five samples on the right have low stability.

5.5 Generalization ability

Following the primary evaluation on Japanese handwritten characters, we examine the generalization ability of the reliability mechanism on four general image clustering benchmarks. These additional experiments cover coarse-grained object categories, fine-grained semantic categories, and different image resolutions, and test whether the same components remain effective outside the target character-clustering task.

5.5.1 Comparison with state-of-the-art methods

To evaluate the generalization ability of STR-DC beyond Japanese handwritten characters, we compare it with a broad range of representative deep clustering methods, including ProPos, CoNR, DMICC, BYOL, MoCo-v2, and LFSS. These comparison methods cover several major paradigms in deep clustering, such as mutual-information-based clustering, neighbor-consistency-based clustering, contrastive representation learning, pseudo-label-based self-training, and stability-guided clustering. Among them, LFSS is a particularly relevant baseline because it also considers

sample stability during clustering. Therefore, comparison with LFSS allows us to examine whether the proposed spatio-temporal reliability-guided selection can further improve clustering performance beyond stability-only sample selection. Table 4 reports the clustering performance of STR-DC and existing state-of-the-art methods on four benchmark datasets. Overall, STR-DC achieves the best performance across all datasets and evaluation metrics.

Table 4 Generalization performance on four general image clustering benchmarks. The metrics are reported in the order of NMI, ACC, and ARI (%). The best results are highlighted in bold, and the second-best results are underlined.

Method	CIFAR-10			CIFAR-20			ImageNet-10			ImageNet-Dogs		
	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
ProPos	85.6	92.0	84.1	54.9	51.8	37.8	81.2	89.0	78.7	54.8	63.9	45.2
CoNR	78.1	85.3	71.1	53.5	49.9	35.8	79.2	87.7	75.0	52.7	62.2	42.1
DMICC	72.2	83.1	67.5	40.0	41.8	24.3	84.3	92.5	85.2	55.1	54.2	39.8
BYOL	72.3	83.3	68.0	51.9	49.3	34.5	73.2	82.1	67.4	51.8	59.7	41.3
MoCo-v2	82.8	90.1	80.6	<u>60.0</u>	<u>59.9</u>	<u>45.6</u>	75.3	77.8	66.1	71.5	70.6	<u>59.4</u>
LFSS	<u>87.2</u>	<u>93.4</u>	<u>86.6</u>	59.9	58.7	43.5	<u>85.6</u>	<u>93.2</u>	<u>85.7</u>	61.7	69.1	53.3
STR-DC (Ours)	87.5	94.2	87.8	60.4	61.1	45.9	93.1	97.2	94.0	77.3	76.1	65.6

On CIFAR-20, a more challenging dataset with fine-grained semantic ambiguity, STR-DC achieves 61.1% ACC, 60.4% NMI, and 45.9% ARI. These scores correspond to relative improvements of 2.0%, 0.7%, and 0.7% over the second-best ACC, NMI, and ARI, respectively. This indicates that the proposed reliability-guided selection remains effective when inter-class semantic boundaries are less distinct. The advantage of STR-DC becomes more evident on ImageNet-10. On ImageNet-10, STR-DC reaches 97.2% ACC, 93.1% NMI, and 94.0% ARI, corresponding to relative improvements of 4.3%, 8.8%, and 9.7%, respectively. These results show that STR-DC can substantially enhance clustering quality when the learned feature space contains clear but still potentially noisy semantic structures. On ImageNet-Dogs, which contains 15 visually similar dog subclasses, STR-DC also achieves strong improvements. This demonstrates that STR-DC is particularly effective in fine-grained clustering scenarios, where unreliable pseudo-labels are more likely to arise from visually similar categories.

These improvements can be attributed to the stricter reliability criterion used in STR-DC. Confidence-based pseudo-labeling can identify samples with strong semantic predictions, but such samples may still be unstable across epochs. Stability-based filtering can identify samples whose representations evolve smoothly, but stability alone does not guarantee semantic correctness. STR-DC combines these two complementary criteria through a spatio-temporal reliability-guided selection mechanism, thereby reducing the risk of using high-confidence but unstable pseudo-labels during self-training.

5.5.2 Visualization analysis

To qualitatively evaluate the generalization behavior of STR-DC, we conduct visualization experiments on CIFAR-10, CIFAR-20, ImageNet-10, and ImageNet-Dogs. First, we compare feature-space similarity relative to the MoCo backbone for correctly and incorrectly clustered samples. Second, we visualize the relationship between ground-truth labels and the stability-based selection mask in the t-SNE feature space. Third, we present examples from the high- and low-stability ends of the sample ranking. Together with the direct cross-epoch analyses on the Japanese handwritten character datasets, these visualizations provide complementary evidence about reliability-guided sample selection across different visual domains.

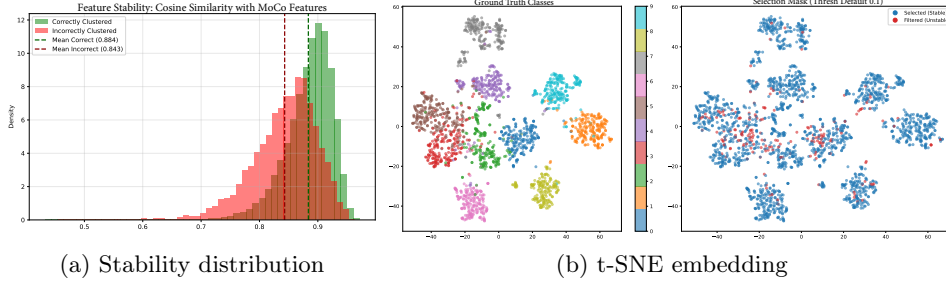


Fig. 6 Generalization visualization on CIFAR-10. Subfigure (a) shows the feature similarity distribution between STR-DC and MoCo for correctly and incorrectly clustered samples. Subfigure (b) compares the t-SNE embeddings colored by ground-truth labels and by the stability-based selection mask.

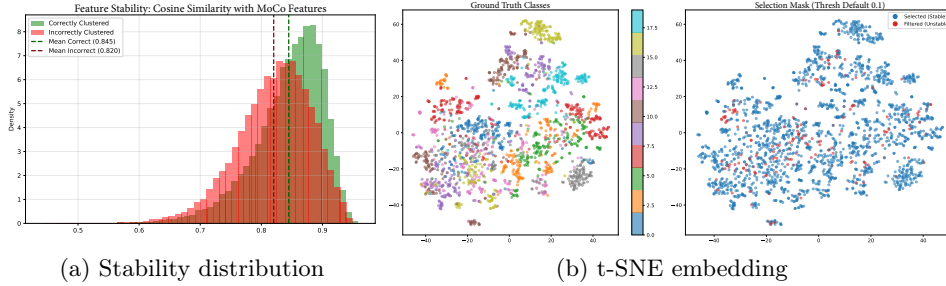


Fig. 7 Visualization results on CIFAR-20. The definitions of subfigures (a) and (b) are the same as those in Fig. 6.

As shown in Figs. 6–9, subfigure (a) reports the cosine similarity between the features learned by STR-DC and those extracted by MoCo. Correctly clustered samples are generally concentrated in the higher-similarity region, whereas incorrectly clustered samples show wider and more left-shifted distributions. Because this comparison uses MoCo features as an external reference rather than consecutive-epoch features, it should be interpreted as a representation-consistency analysis that complements, but does not replace, the direct temporal-stability measurements used by STR-DC.

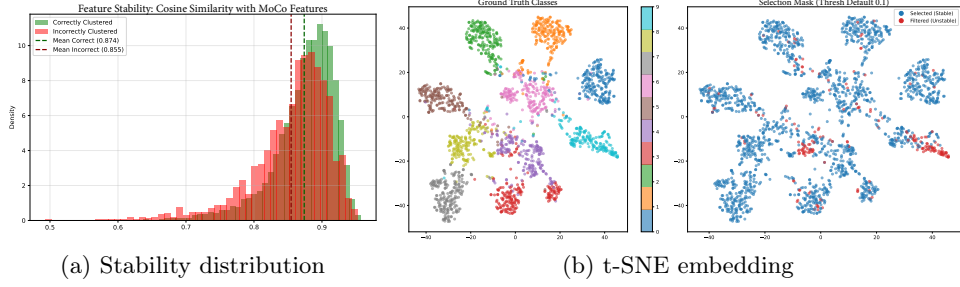


Fig. 8 Visualization results on ImageNet-10. The definitions of subfigures (a) and (b) are the same as those in Fig. 6.

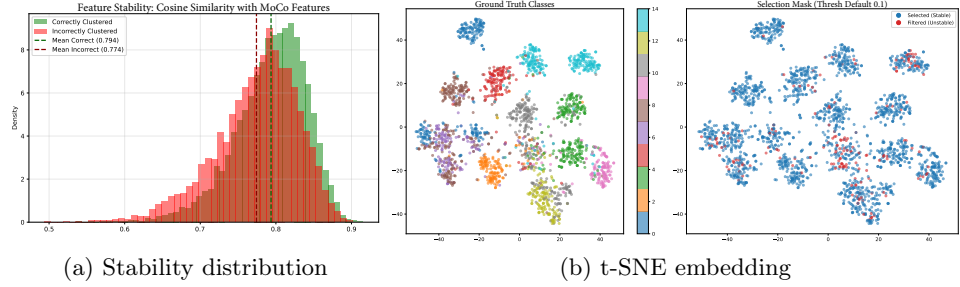


Fig. 9 Visualization results on ImageNet-Dogs. The definitions of subfigures (a) and (b) are the same as those in Fig. 6.

The t-SNE plots in subfigure (b) of Figs. 6–9 further illustrate the geometric distribution of the stability-based selection in the learned feature space. The left panel shows the t-SNE embeddings colored by ground-truth labels, while the right panel distinguishes selected stable samples from filtered low-stability samples. Filtered samples appear more frequently near apparent cluster boundaries, overlapping regions, or sparse outlier areas. This pattern is consistent with the use of temporal stability as an additional verification signal after calibrated semantic selection.

6 Conclusions

In this paper, we proposed STR-DC for unsupervised clustering of Japanese handwritten characters, motivated by the need to organize character images when instance-level annotations are unavailable. The method addresses unreliable pseudo-labels arising from ambiguous character shapes and variation in handwriting by combining two components. The calibrated confidence component selects candidates with category-level evidence, and the feature stability component verifies their representation consistency across consecutive epochs. Only candidates satisfying both conditions are used for self-training, supporting more reliable grouping by character identity.

Experiments on K49, ETL8-Hira75, and ETL9-Hira71 demonstrate the effectiveness of STR-DC for Japanese handwritten character clustering. The ablation

results further show that the calibrated confidence component and the feature stability component are complementary, and their combination is more effective than either component alone. Experiments on CIFAR-10, CIFAR-20, ImageNet-10, and ImageNet-Dogs additionally demonstrate that the proposed reliability mechanism generalizes beyond handwritten character images. The visualization results provide complementary evidence for reliable sample selection. These findings support clustering as a step toward organizing unlabeled Japanese handwritten character collections for subsequent inspection and annotation.

Despite its effectiveness, STR-DC still has several limitations. The current temporal reliability estimation mainly relies on short-term feature stability between adjacent epochs, and the stability threshold still requires empirical selection. In addition, maintaining and updating feature memory may introduce extra computational cost for large-scale datasets. In future work, we will explore more adaptive temporal reliability modeling strategies, more efficient memory update mechanisms, and extensions to Japanese handwritten character collections with imbalanced category frequencies or previously unseen character categories.

Declarations

Funding. The authors declare that financial support was received for the research, authorship, and/or publication of this article. This research was funded by the 2021 Fundamental Research Funds for Provincial Universities of the Heilongjiang Provincial Department of Education (Grant No. 2021-KYYWF-0201).

Author contributions. Conceptualization, Y.L. and Z.L.; methodology, Y.L.; software, Y.L.; validation, Y.L. and Z.L.; formal analysis, Y.L.; investigation, Y.L.; resources, Y.L. and Z.L.; data curation, Y.L.; writing—original draft preparation, Y.L.; writing—review and editing, Y.L. and Z.L.; visualization, Y.L.; supervision, Z.L.; project administration, Z.L. All authors have read and agreed to the published version of the manuscript.

Competing interests. The authors declare no competing interests.

Ethics approval and consent to participate. Not applicable.

Consent for publication. Not applicable.

Data availability. The publicly available datasets analyzed in this study can be obtained from the original repositories cited in the manuscript. No new dataset was created.

References

- [1] Yang Q, Li Z, Han G, Gao W, Zhu S, Wu X, et al. An improvement of spectral clustering algorithm based on fast diffusion search for natural neighbor and affinity propagation. *The Journal of Supercomputing*. 2022;78(12):14597-625.

- [2] Yang Q, Gao W, Han G, Li Z, Tian M, Zhu S, et al. HCDC: A novel hierarchical clustering algorithm based on density-distance cores for data sets with varying density. *Information Systems*. 2023;114:102159.
- [3] Yang Q, Han G, Gao W, Yang Z, Zhu S, Deng Y. A Robust Learning Membership Scaling Fuzzy C-Means Algorithm Based on New Belief Peak. *IEEE Transactions on Fuzzy Systems*. 2023;31(12):4486-500.
- [4] Huang J, Deng Y, Zhou Y, Pang S, Yang Q, Min G. DBCGM: A Granular Model for Big Data Classification Based on Data Bisection and Cascade Weighted Clustering. *IEEE Transactions on Knowledge and Data Engineering*. 2025;37(12):7045–7059.
- [5] Gui J, Chen T, Zhang J, Cao Q, Sun Z, Luo H, et al. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2024;46(12):9052-71.
- [6] Caron M, Bojanowski P, Joulin A, Douze M. Deep clustering for unsupervised learning of visual features. In: *Proceedings of the European conference on computer vision (ECCV)*; 2018. p. 132-49.
- [7] Ji X, Henriques JF, Vedaldi A. Invariant information clustering for unsupervised image classification and segmentation. In: *Proceedings of the IEEE/CVF international conference on computer vision*; 2019. p. 9865-74.
- [8] Huang Z, Chen J, Zhang J, Shan H. Learning representation for clustering via prototype scattering and positive sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2022;45(6):7509-24.
- [9] Yu C, Shi Y, Wang J. Contextually affinitive neighborhood refinery for deep clustering. *Advances in Neural Information Processing Systems*. 2023;36:5778-90.
- [10] Li Y, Hu P, Peng D, Lv J, Fan J, Peng X. Image clustering with external guidance. *arXiv preprint arXiv:231011989*. 2023.
- [11] Liu H, Hu P, Zhang C, Li Y, Peng X. Interactive deep clustering via value mining. *Advances in Neural Information Processing Systems*. 2024;37:42369-87.
- [12] Li Z, Jia Y. Conmix: Contrastive mixup at representation level for long-tailed deep clustering. In: *The Thirteenth International Conference on Learning Representations*; 2025. p. 3595-614.
- [13] Chang L, Niu X, Li Z, Zhang Z, Li S, Fournier-Viger P. ULDC: uncertainty-based learning for deep clustering: L. Chang et al. *Applied Intelligence*. 2025;55(3):223.

- [14] Jia Y, Cheng J, Liu H, Hou J. Towards calibrated deep clustering network. In: The Thirteenth International Conference on Learning Representations; 2025.
- [15] Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: International conference on machine learning. PMLR; 2017. p. 1321-30.
- [16] Minderer M, Djolonga J, Romijnders R, Hubis F, Zhai X, Houlsby N, et al. Revisiting the calibration of modern neural networks. *Advances in neural information processing systems*. 2021;34:15682-94.
- [17] Zhang S, Xie L. Advancing neural network calibration: The role of gradient decay in large-margin Softmax optimization. *Neural Networks*. 2024;178:106457.
- [18] Zając M, Tuytelaars T, van de Ven G. Prediction error-based classification for class-incremental learning. In: The Twelfth International Conference on Learning Representations. vol. 2024; 2024. p. 8239-67.
- [19] Li Z, Jia Y, Liu H, Hou J. Learning from sample stability for deep clustering. In: Proceedings of the 42nd International Conference on Machine Learning; 2025. p. 34904–34919.
- [20] Feldman V. Does learning require memorization? a short tale about a long tail. In: Proceedings of the 52nd annual ACM SIGACT symposium on theory of computing; 2020. p. 954-9.
- [21] Maini P, Mozer MC, Sedghi H, Lipton ZC, Kolter JZ, Zhang C. Can neural network memorization be localized? *arXiv preprint arXiv:2307.09542*. 2023.
- [22] Chen X, Fan H, Girshick R, He K. Improved Baselines with Momentum Contrastive Learning. *arXiv preprint arXiv:2003.04297*. 2020. Available from: <https://arxiv.org/abs/2003.04297>.
- [23] Grill JB, Strub F, Altché F, Tallec C, Richemond P, Buchatskaya E, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*. 2020;33:21271-84.
- [24] Zbontar J, Jing L, Misra I, LeCun Y, Deny S. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In: Proceedings of the 38th International Conference on Machine Learning; 2021. p. 12310-20.
- [25] Bardes A, Ponce J, LeCun Y. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. In: The Tenth International Conference on Learning Representations; 2022.
- [26] Krizhevsky A, Hinton G, et al. Learning multiple layers of features from tiny images. Master’s thesis. 2009.

- [27] Coates A, Ng A, Lee H. An Analysis of Single-Layer Networks in Unsupervised Feature Learning. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics; 2011. p. 215-23.
- [28] Chang J, Wang L, Meng G, Xiang S, Pan C. Deep Adaptive Image Clustering. In: 2017 IEEE International Conference on Computer Vision (ICCV); 2017. p. 5880-8.
- [29] Clanuwat T, Bober-Irizar M, Kitamoto A, Lamb A, Yamamoto K, Ha D. Deep learning for classical Japanese literature. arXiv preprint arXiv:1812.01718. 2018.
- [30] National Institute of Advanced Industrial Science and Technology. The ETL Character Database. Available from: <https://etlcdb.db.aist.go.jp/the-etl-character-database/>. Accessed: Sep. 6, 2026.
- [31] Wu J, Long K, Wang F, Qian C, Li C, Lin Z, et al. Deep Comprehensive Correlation Mining for Image Clustering . In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019. p. 8149-58.
- [32] Huang J, Gong S, Zhu X. Deep semantic clustering by partition confidence maximisation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2020. p. 8849-58.
- [33] Dang Z, Deng C, Yang X, Wei K, Huang H. Nearest Neighbor Matching for Deep Clustering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021. p. 13693-702.
- [34] Shen Y, Shen Z, Wang M, Qin J, Torr P, Shao L. You never cluster alone. Advances in Neural Information Processing Systems. 2021:27734-46.
- [35] Li H, Zhang L, Su K. Dual Mutual Information Constraints for Discriminative Clustering. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2023. p. 8571–8579.