SPICE: Semantic Pseudo-Labeling for Image Clustering

- instance-level similarity
- cluster-level discrepancy

$\mathcal{X} = \{x_i\}_{i=1}^{N}$ N images $\xrightarrow{\text{cluster}}$ K classes

two parts:

① a feature model: images → feature vectors

$f_i = F(x_i; \theta_F)$ measure the similarity among samples

② a clustering head: feature vectors → probabilities

$p_i = C(f_i; \theta_C)$ identify the discrepancy between clusters

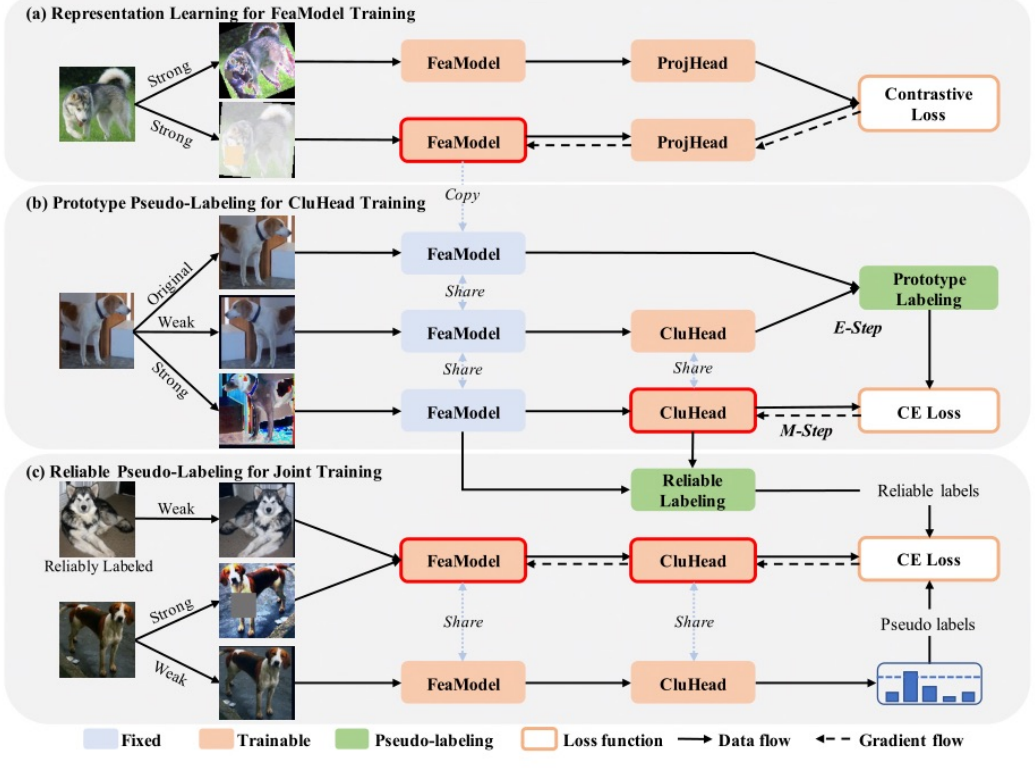
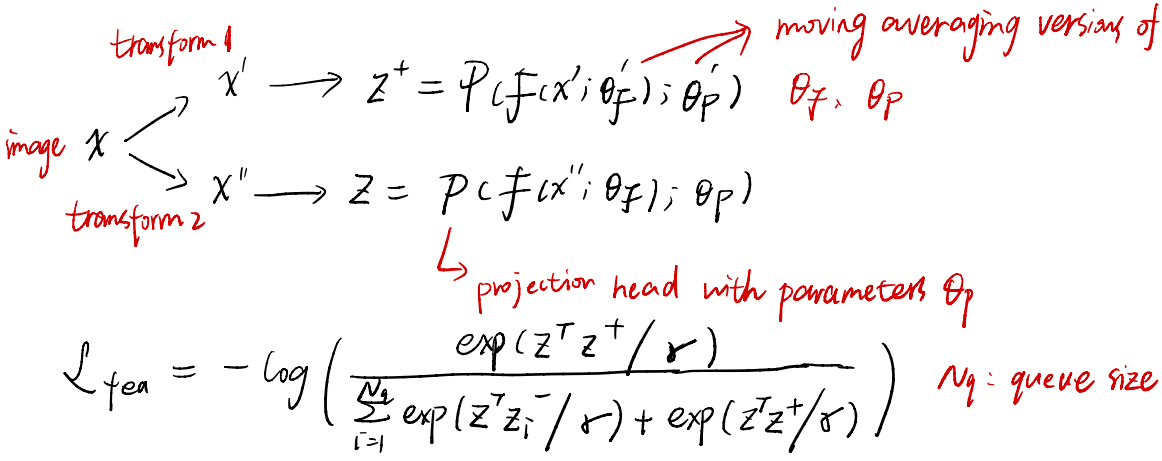


Fig. 3. Illustration of the SPICE framework. (a) Train the feature model with the contrastive learning based unsupervised representation learning. (b) Train the clustering head via the prototype pseudo-labeling algorithm in an EM framework. (c) Jointly train the feature model and the clustering head through the reliable pseudo-labeling algorithm.

three stages:

A. Feature Model Training with constrastive learning



Finally, the optimized feature model parameters are denoted as θ_f^s.

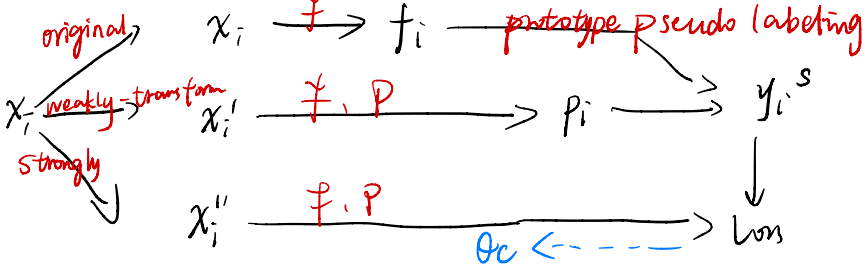
B. Clustering Head Training with Prototype Pseudo-Labeling

- the parameters of clustering head C, θ_C — $p_i = C(f_i; \theta_C)$
- the cluster labels $\{y_i^s\}$ of x over K clusters. where $f_i = f(x_i; \theta_f^s)$

EM framework:

Expectation (E) step → solve $\{y_i^s\}$ given θ_C

Maximization (M) step → solve θ_C given $\{y_i^s\}$



B-1 : Prototype Pseudo-Labeling (E-step)

top branch : $F = [f_1, f_2, \cdots, f_M]^T \in \mathbb{R}^{M \times D}$ for $\mathcal{X}_b$ Given

middle branch : $P = [p_1, p_2, \cdots, p_M]^T \in \mathbb{R}^{M \times K}$ for $\alpha(\mathcal{X}_b)$

Select the top confident samples for k-th cluster :

$$\bar{\mathcal{S}}_k = \{ f_i \mid i \in \text{argtopk}(P_{:,k}, \frac{M}{K}), \forall i = 1, 2, \cdots, M \} \quad (2)$$

$P_{:,k}$: k-th column of P

$\text{argtopk}(P_{:,k}, \frac{M}{K})$ returns the top $\frac{M}{K}$ confident sample indices

eg. $P = \begin{bmatrix} 0.7 & 0.6 & 0.2 & 0.9 \\ 0.3 & 0.4 & 0.8 & 0.1 \end{bmatrix}^T$ $M = 4$, $K = 2$ $\bar{\mathcal{S}}_1 = \{ f_1, f_4 \}$ $\bar{\mathcal{S}}_2 = \{ f_2, f_3 \}$

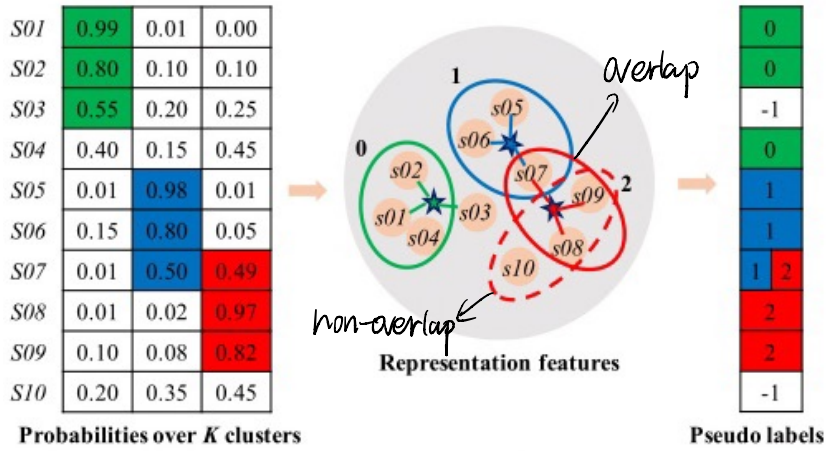
The cluster centers $\{ \gamma_k \}_{k=1}^K$: $\gamma_k = \frac{K}{M} \sum_{f_i \in \bar{\mathcal{S}}_k} f_i$, $\forall k = 1, 2, \cdots, K$ (3)

balanced ?

Compute the cosine similarity between f_i and γ_k, select $\frac{M}{K}$ nearest samples to γ_k, denoted by $\mathcal{X}^K \rightarrow y_i^S = k$

$$\mathcal{X}^S = \{ (x_i, y_i^S) \mid \forall x_i \in \mathcal{X}^K, k = 1, 2, \cdots, K \} \quad (4)$$

Mini-batch of images with semantic pseudo-labels



B-2 : Training Clustering Head (M-step)

Given $\mathcal{X}^s$ → supervised learning

bottom batch : compute $\beta(\mathcal{X}^s)$'s probabilities

C can be optimized by minimizing CE loss :

$$L_{clu} = \frac{1}{M} \sum_{i=1}^{M} L_{ce}(y_i^s, p_i') \qquad (5)$$

where $p_i' = \text{softmax}(p_i)$, $p_i = A(\mathcal{F}(\beta(x_i); \theta_{\mathcal{F}}^s); \theta_A)$

double softmax (p_i already softmax)

Algorithm 1: Training Clustering Head.

Input: Dataset $\mathfrak{X} = \{\boldsymbol{x}_i\}_{i=1}^N$, $\boldsymbol{\theta}_{\mathcal{F}}^s$, K, M, m, T, α, β
Output: Cluster label y_i^s of $x_i \in \mathcal{X}$

Set feature model parameters to $\boldsymbol{\theta}_{\mathcal{F}}^s$, $t = 0$, and initialize $\boldsymbol{\theta}_{\mathcal{C}}$;
while $t < T$ **do**
  for $b = 1, 2, \ldots, \lfloor \frac{N}{M} \rfloor$ **do**
    ***E-step*:**
    Select M samples from $\mathfrak{X}$ as $\mathfrak{X}_b$;
    Compute embedding features $\boldsymbol{F} = \mathcal{F}(\mathcal{X}_b; \boldsymbol{\theta}_{\mathcal{F}}^s)$;
    Predict probabilities $\boldsymbol{P} = \mathcal{C}(\mathcal{F}(\alpha(\mathfrak{X}_b); \boldsymbol{\theta}_{\mathcal{F}}^s); \boldsymbol{\theta}_{\mathcal{C}})$;
    Construct labeled image set $\mathcal{X}^s$ with Eqs. (2), (3), and (4) ;
    ***M-step*:**
    Compute probabilities $\boldsymbol{P} = \mathcal{C}(\mathcal{F}(\beta(\mathcal{X}_b); \boldsymbol{\theta}_{\mathcal{F}}^s); \boldsymbol{\theta}_{\mathcal{C}})$;
    Optimize $\boldsymbol{\theta}_{\mathcal{C}}$ by minimizing Eq. (5) ;
  end
  $t \leftarrow t + 1$
end
Select the best clustering head with the minimum loss as $\boldsymbol{\theta}_{\mathcal{C}}^s$;
foreach $\boldsymbol{x}_i \in \mathfrak{X}$ **do**
  $\boldsymbol{p}_i = \mathcal{C}(\mathcal{F}(\boldsymbol{x}_i; \boldsymbol{\theta}_{\mathcal{F}}^s); \boldsymbol{\theta}_{\mathcal{C}}^s)$;
  $y_i^s = \arg\max_k(\boldsymbol{p}_i)$;
end

C. Joint Training with Reliable Pseudo-Labeling

Reliable Pseudo-Labeling:

Given $\{(x_i, f_i, y_i^s)\}_{i=1}^{N}$ from "B".

For each x_i, select N_s nearest samples, corresponding labels $\mathcal{L}_i$

The semantically consistent ration r_i of x_i:

$$r_i = \frac{1}{N_s} \sum_{y \in \mathcal{L}_i} \mathbb{1}(y = y_i^s)$$

Given a threshold λ, if $r_i > \lambda$, (x_i, y_i^s) is reliably labeled

otherwise, is ignored.

partially labeled

$$\mathcal{X}^r = \{(x_i, y_i^s) \mid r_i > \lambda, \forall i = 1, 2, \cdots, N\}$$

Joint Training: Semi-supervised learning.

we adapt a simple semi-supervised learning method [54]. During training, the subset of reliably labeled samples keep fixed. On the other hand, all training samples should be consistently clustered, *i.e.*, different transformations of the same image are constrained to have the consistent prediction. To this end, as shown in Fig. 3(c), the confidently predicted label of weak transformations is used as the pseudo-label for strong transformations of the same image. Formally, the consistency pseudo label y_j^u of the sample $\boldsymbol{x}_j$ is calculated as in Eq. (8):

$$y_j^u = \begin{cases} \arg\max(\boldsymbol{p}_j) & \text{if} \max(\boldsymbol{p}_j) \geq \eta, \\ -1 & \text{otherwise} \end{cases} \tag{8}$$

where $\boldsymbol{p}_j = \mathcal{C}(\mathcal{F}(\alpha(\boldsymbol{x}_j); \boldsymbol{\theta}_{\mathcal{F}}); \boldsymbol{\theta}_{\mathcal{C}}))$, and η is the confidence threshold.

Then, the whole network parameters $\boldsymbol{\theta}_{\mathcal{F}}$ and $\boldsymbol{\theta}_{\mathcal{C}}$ are optimized with the following loss function:

$$\begin{aligned} \mathcal{L}_{joint} = & \frac{1}{L} \sum_{i=1}^{L} \underbrace{\mathcal{L}_{ce}(y_i^s, \mathcal{C}(\mathcal{F}(\alpha(\boldsymbol{x}_i); \boldsymbol{\theta}_{\mathcal{F}}); \boldsymbol{\theta}_{\mathcal{C}}))}_{\text{partial samples with reliable pseudo-labels}} \\ & + \frac{1}{U} \sum_{j=1}^{U} \underbrace{\mathbb{1}(y_j^u \geq 0) \mathcal{L}_{ce}(y_j^u, \mathcal{C}(\mathcal{F}(\beta(\boldsymbol{x}_j); \boldsymbol{\theta}_{\mathcal{F}}); \boldsymbol{\theta}_{\mathcal{C}}))}_{\text{all samples with consistency pseudo-labels}}, \end{aligned} \tag{9}$$

where the first term is computed with reliably labeled samples $(\boldsymbol{x}_i, y_i^s)$ drawn from $\mathcal{X}^r$, and the second term is computed with pseudo-labeled samples $(\boldsymbol{x}_i, y_i^u)$ drawn from the whole dataset $\mathcal{X}$, which is dynamically labeled by thresholding the confident predictions as in Eq. (8). L and U denote the numbers of labeled and unlabeled images in a mini-batch.