Introduction

• In this paper, we propose a novel contrastive learning method which explores the cross-video relation by using cycle-consistency for general image representation learning.
• We validate our method by transferring our image representation to multiple downstream tasks including visual object tracking, image classification, and action recognition.

Cross-video Cycle Consistent Learning

Given the encoded representation $q_i$ for an image,

• Step 1. Cycle Forward. Find the soft nearest neighbor $\hat{q}_i$ in the candidate neighbor set $U$

$$\hat{q}_i = \sum_{u \in U} \text{sim}(q_i, u)u,$$

where $\text{sim}(\cdot)$ calculates the normalized cosine similarity with temperature $\tau$.

• Step 2. Cycle Backward. $\hat{q}_i$ should be close to $q_i$ or invariant set of $q_i$. Thus we use $k_j$, which is the representation of image within the same video to form the positive pair $(\hat{q}_i, k_j)$. The loss is

$$L_{\text{cycle}} = - \log \frac{\exp(\text{sim}(\hat{q}_i, k_j) / \tau)}{\sum_{u \in \{U, k_j\}} \exp(\text{sim}(\hat{q}_i, u) / \tau)}.$$