ONLINE UNSUPERVISED LEARNING OF VISUAL REPRESENTATIONS AND CATEGORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Real world learning scenarios involve a nonstationary distribution of classes with sequential dependencies among the samples, in contrast to the standard machine learning formulation of drawing samples independently from a fixed, typically uniform distribution. Furthermore, real world interactions demand learning on-the-fly from few or no class labels. In this work, we propose an unsupervised model that simultaneously performs online visual representation learning and few-shot learning of new categories without relying on any class labels. Our model is a prototype-based memory network with a control component that determines when to form a new class prototype. We formulate it as an online Gaussian mixture model, where components are created online with only a single new example, and assignments do not have to be balanced, which permits an approximation to natural imbalanced distributions from uncurated raw data. Learning includes a contrastive loss that encourages different views of the same image to be assigned to the same prototype. The result is a mechanism that forms categorical representations of objects in nonstationary environments. Experiments show that our method can learn from an online stream of visual input data and is significantly better at category recognition compared to state-of-the-art self-supervised learning methods.

1 INTRODUCTION

Human and artificial agents learn from large amounts of unlabeled data through a continuous stream of experience that has strong temporal correlations and hierarchical structure. The stream of experience is determined by the fact that individuals operate in distinct physical environments, such as home and office. Transitions between environments occur on a coarse time scale relative to the rate of encounters with objects in a given environment. The distribution over objects is strongly conditioned on the environment: frozen pizza and milk are in the supermarket, computer monitors and desks at the office. The distribution may also be nonstationary; for example, a visitor may be infrequent, but when they are around, they're encountered frequently. And finally, the distribution can be highly unbalanced: an individual may interact with their co-workers daily but the boss only occasionally.

The goal of our research is to tackle the challenging problem of online unsupervised representation learning in the setting of environments with naturalistic structure. We desire a learning algorithm that will facilitate the categorization of objects encountered in the environment, with few or zero category labels. In representation learning, methods often evaluate their ability to classify from the representation using either supervised linear readout or unsupervised clustering over the full dataset, both of which are typically done in a separate post-hoc evaluation phase. An important aim of our work is to produce object-category predictions throughout training and evaluation, and to allow these predictions to guide subsequent categorization.

Unsurprisingly, the structure of natural environments contrasts dramatically with the standard scenario typically assumed by many machine learning algorithms: mini-batches of independent and identically distributed (iid) samples from a well-curated dataset. In unsupervised visual representation learning, the most successful methods rely on iid samples. Contrastive-based objectives (Chen et al., 2020a; He et al., 2020) typically assume that each instance in the mini-batch forms its own instance class, throwing away the potential similarity between instances. Clustering-based learning frameworks (Caron et al., 2018; Asano et al., 2020; Caron et al., 2020) often assume that the set of cluster centroids remain relatively stable and that the clusters are balanced in size. Unfortunately, none of these assumptions necessarily hold true in a naturalistic online streaming setting. The performance of

methods will suffer as a consequence. Contrastive approaches could eventually fail if examples of the same class are pushed apart; clustering approaches will behave erratically with nonstationary and imbalanced class distributions, two key facets of the natural online non-iid learning we focus on here.

To make progress on the challenge of unsupervised visual representation learning and categorization in a naturalistic setting, we propose the *online unsupervised prototypical network*, which performs learning of visual representations and object categories simultaneously in a single-stage process. Class prototypes are created via an online clustering procedure, and a contrastive loss (van den Oord et al., 2018) is used to encourage different views of the same image to be assigned to the same cluster. Notably, our online clustering procedure is more flexible relative to other clustering-based representation learning algorithms, such as DeepCluster (Caron et al., 2018) and SwAV (Caron et al., 2020): our model performs learning and inference as an online Gaussian mixture model, where clusters can be created online with only a single new example, and cluster assignments do not have to be balanced, which permits an approximation to natural imbalanced distributions from uncurated raw data.

We train and evaluate our algorithm on a recently proposed naturalistic learning dataset, RoamingRooms (Ren et al., 2021), which uses imagery collected from a virtual agent walking through different rooms. Unlike the experiments in Ren et al. (2021), our training is done without using any labeled data, and hence it is online and unsupervised. We compare to state-of-the-art unsupervised methods SimCLR (Chen et al., 2020a) and SwAV (Caron et al., 2020). Because they rely on sampling a large batch from an offline dataset, their performance drops significantly using smaller non-iid episodes. In contrast, our method can directly learn with online streaming data from small episodes, without requiring an example buffer, and surprisingly, we even outperform these strong methods trained offline with large batches of iid data. We also used RoamingOmniglot (Ren et al., 2021) as a benchmark to investigate the effect of imbalanced classes and we find that our method is very robust to an imbalanced distribution of classes. For a version of ImageNet with non-iid structure, RoamingImageNet, we again outperform SimCLR and SwAV when using the same batch size. Finally, qualitative visualizations confirm that our online clustering procedure can automatically discover new concepts and categories by grouping a few similar instances together in the learned embedding space.

2 Related Work

Self-supervised learning. Self-supervised learning methods discover rich and informative visual representations without using class labels. *Instance-based approaches* aim to learn invariant representations of each image under different transformations (van den Oord et al., 2018; Misra & van der Maaten, 2020; Tian et al., 2020; He et al., 2020; Chen et al., 2020a;b; Grill et al., 2020; Chen & He, 2021). They typically work well under iid learning with large batch sizes, which contrasts with realistic learning scenarios. Our method is also related to *clustering-based approaches*, which computes clusters on top of the learned embedding, and uses the cluster assignment to self-supervise the embedding network (Caron et al., 2018; Zhan et al., 2020; Asano et al., 2020; Caron et al., 2020; Li et al., 2021). The self-supervision loss is found to be more powerful than the traditional reconstruction loss (Rao et al., 2019; Greff et al., 2017; Joo et al., 2020). To compute the cluster assignment, DeepCluster (Caron et al., 2018; Zhan et al., 2020) and PCL (Li et al., 2021) use the k-means algorithm whereas SeLa (Asano et al., 2020) and SwAV (Caron et al., 2020) uses the Sinkhorn-Knopp algorithm (Cuturi, 2013). However, they typically assume a fixed number of clusters, and Sinkhorn-Knopp further assumes a balanced assignment as an explicit constraint. In contrast, our online clustering procedure is more flexible, as it can create new clusters on-the-fly with only a single new example and does not assume balanced cluster assignments.

Representation learning from video. There has also been a surge of interest in leveraging video data to learn visual representations (Wang & Gupta, 2015; Orhan et al., 2020; Pathak et al., 2017; Zhu et al., 2020; Xiong et al., 2021). Wang & Gupta (2015) proposed to sample positive example pairs from a video track of the same object; Orhan et al. (2020) divided the video into equal chunks and trains a network to predict the chunk index. Pathak et al. (2017) learned visual representations using low-level motion-based grouping cues, and more recently, Xiong et al. (2021) used optical flow to encourage regions that move together to have consistent representations across frames. These approaches all sample video subsequences uniformly over the entire dataset, whereas our model directly learns from an online stream of data. Our model also does not have the assumption that inputs must be adjacent frames in the video.

Online and incremental representation learning. Our work is also related to online and continual representation learning (Rebuffi et al., 2017; Castro et al., 2018; Rao et al., 2019; Jerfel et al., 2019; Javed & White, 2019; Hayes et al., 2020). Incremental learning (Rebuffi et al., 2017; Castro et al., 2018) studies learning a set of classes incrementally. Continual mixture models (Rao et al., 2019; Jerfel et al., 2019) designate a categorical latent variable that can be dynamically allocated for a new environment. Our model has a similar mixture latent variable setup but one major difference is that we operate on example-level rather than task-level. OML (Javed & White, 2019) is a meta-learning framework that learns the representation to support online learning at test time. Streaming learning (Hayes et al., 2020; 2019) aims to perform representation learning online. Most of the works here except Rao et al. (2019) assume that the training data is fully supervised. Our prototype memory also resembles the replay buffer for experience replay methods (Buzzega et al., 2020; Kim et al., 2020), but we store the hidden feature prototypes instead of the input examples.

Latent variable modeling on sequential data. Our model also relates to a family of latent variable generative models for sequential data (Johnson et al., 2016; Krishnan et al., 2015; He et al., 2018; Denton & Fergus, 2018; Zhu et al., 2020). Both our model and these approaches aim to infer latent variables with temporal structure. However, instead of using input reconstruction as the main learning objective, our model leverages self-supervised contrastive learning to avoid learning difficulty in the generative decoder.

Online mixture models. Our clustering module is related to the literature on online mixture models, e.g., Hughes & Sudderth (2013); Pinto & Engel (2015); Song & Wang (2005). Typically, these are designed for fast and incremental learning of clusters without having to recompute clustering over the entire dataset. Despite presenting a similar online clustering algorithm, our goal is to jointly learn both online clusters and input representations that facilitate future online clustering episodes.

Few-shot learning. Our model is able to recognize a new class with only one or a few examples, which relates to the literature on few-shot learning (Li et al., 2003; Lake et al., 2015). Our prototype-based memory is also inspired by the Prototypical Network (Snell et al., 2017) and its variants, e.g., IMP (Allen et al., 2019) and Online ProtoNet (Ren et al., 2021). Ren et al. (2018); Huang et al. (2019) demonstrated improved performance by using unlabeled examples. More recent methods can entirely remove the reliance on class labels in the training pipeline, using self-supervised learning (Hsu et al., 2019; Gidaris et al., 2019; Antoniou & Storkey, 2019; Khodadadeh et al., 2019; Medina et al., 2020).

Classical few-shot learning, however, relies on episodes of equal number of training and test examples from a fixed number of new classes. Incremental few-shot learning approaches (Gidaris & Komodakis, 2018) consider both old and new classes, while Meta-Dataset (Triantafillou et al., 2020) considers episodes with varying numbers of examples and classes. Ren et al. (2021) proposed a new setup that incrementally accumulates new classes and re-visits old classes over a sequence of inputs. We evaluate our algorithm on a similar setup; however, unlike that work, our proposed algorithm learns visual representations without relying on any class labels.

Human category learning. Our work is inspired by human learning settings and relevant models from cognitive science. Lake et al. (2009) proposed an online category learning framework using online mixture models on top of fixed representations. SUSTAIN (Love et al., 2004) is another computational model for human category learning that can perform both supervised and unsupervised clustering of categories. Mack et al. (2016) further augmented the SUSTAIN model such that categories can be learned in a goal-oriented manner. In contrast to these human learning models, our model learns the mapping between images to categories in an end-to-end fashion.

3 Online Unsupervised Prototypical Networks

In this section, we introduce our proposed model, *online unsupervised prototypical networks* (*OUPN*). We study the online categorization setting, where the model receives an input $\mathbf{x}_t$ at every time step t, and predicts both a categorical variable $\hat{y}_t$ to indicate the object class and also a binary variable $\hat{u}_t$ to indicate whether this is a known ($u = 0$) or new ($u = 1$) class. OUPN uses a network h to encode the input to obtain embedding $\mathbf{z}_t = h(\mathbf{x}_t; \theta)$, where θ represents the learnable parameters of the encoder network. $\mathbf{z}$ is then processed by a *prototype memory* which predicts $(\hat{y}_t, \hat{u}_t)$.

3.1 Prototype memory

We formulate our prototype memory as a probabilistic mixture model, where each cluster corresponds to a Gaussian distribution $f(\mathbf{z}; \mathbf{p}, \sigma^2)$, with mean $\mathbf{p}$, a constant isotropic variance σ^2 shared across

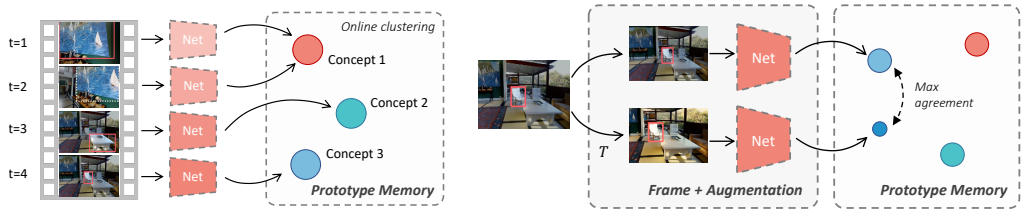

Figure 1: Our proposed online unsupervised prototypical network (OUPN). **Left:** OUPN learns directly from an online visual stream. Images are processed by a deep neural network to extract representations. Representations are stored and clustered in a prototype memory. Similar features are aggregated in a concept and new concepts can be dynamically created if the current feature vector is different from all existing concepts. **Right:** The network learning uses self-supervision that encourages different augmentations of the same frame to have consistent cluster assignments.

all clusters, and mixture weights w: $p(\mathbf{z}; P) = \sum_k w_k f(\mathbf{z}; \mathbf{p}_k, \sigma^2)$. Throughout a sequence, the number of components evolves as the model makes an online decision of when to create a new cluster or remove an old one. We assume that the prior distribution for the Bernoulli variable u is constant ($u_0 \equiv \Pr(u = 1)$), and the prior for a new cluster is uniform over the entire space ($z_0 \equiv \Pr(\mathbf{z}|u = 1)$). In the following, we formulate our prototype memory as an approximation to an online EM algorithm. The full derivation is included in the supplementary materials.

3.1.1 E-STEP

Upon seeing the current input $\mathbf{z}_t$, the online clustering procedure needs to predict the cluster assignment or initiate a new cluster in the E-step.

Inferring cluster assignments. The categorical variable $\hat{y}$ infers the cluster assignment of the current input example with regard to the existing clusters.

$$\hat{y}_{t,k} = \Pr(y_t = k|\mathbf{z}_t, u = 0) = \frac{\Pr(\mathbf{z}_t|y_t = k, u = 0)\Pr(y_t = k)}{\Pr(\mathbf{z}_t, u = 0)} \tag{1}$$

$$= \frac{w_k f(\mathbf{z}_t; \mathbf{p}_{t,k}, \sigma^2)}{\sum_{k'} w_{k'} f(\mathbf{z}_t; \mathbf{p}_{t,k'}, \sigma^2)} = \text{softmax}\left(\log w_k - \frac{1}{\tau} d(\mathbf{z}_t, \mathbf{p}_{t,k})\right), \tag{2}$$

where w_k is the mixing coefficient of cluster k, $d(\cdot, \cdot)$ is the distance function, and τ is an independent learnable temperature parameter that is related to the cluster variance.

Inference on unknown classes. The binary variable $\hat{u}$ estimates the probability that the current input belongs to a new cluster:

$$\hat{u}_t = \Pr(u_t = 1|\mathbf{z}_t) = \frac{\Pr(\mathbf{z}_t|u_t = 1)\Pr(u_t = 1)}{\Pr(\mathbf{z}_t|u_t = 1)\Pr(u_t = 1) + \sum_k \Pr(\mathbf{z}_t|y_t = k, u = 0)\Pr(u = 0)} \tag{3}$$

$$= \frac{z_0 u_0}{z_0 u_0 + \sum_k w_k f(\mathbf{z}_t; \mathbf{p}_{t,k}, \sigma^2)(1 - u_0)} \geq \frac{z_0 u_0}{z_0 u_0 + \max_k f(\mathbf{z}_t; \mathbf{p}_{t,k}, \sigma^2)(1 - u_0)} \tag{4}$$

$$= \text{sigmoid}((\min_k \frac{1}{\tau} d(\mathbf{z}_t, \mathbf{p}_{t,k}) - \beta)/\gamma), \tag{5}$$

where β and γ are separate learnable parameters related to z_0 and u_0, allowing us to predict different confidence levels for unknown and known classes.

3.1.2 M-STEP

Here we infer the posterior distribution of the prototypes $\Pr(\mathbf{p}_{t,k}|\mathbf{z}_{1:t})$. We formulate an efficient recursive online update, similar to Kalman filtering, incorporating the evidence of the current input $\mathbf{z}_t$ and avoiding re-clustering the entire input history. We define $\hat{\mathbf{p}}_{t,k}$ as the posterior estimate of the mean of the k-th cluster at time step t, and $\hat{c}_{t,k}$ is the estimate of the inverse variance.

Updating prototypes. Suppose that in the E-step we have determined that $y_t = k$. Then the posterior distribution of the k-th cluster after observing $\mathbf{z}_t$ is:

$$\Pr(\mathbf{p}_{t,k}|\mathbf{z}_{1:t}, y_t = k) \propto \Pr(\mathbf{z}_t|\mathbf{p}_{t,k}, y_t = k)\Pr(\mathbf{p}_{t,k}|\mathbf{z}_{1:t-1}) \tag{6}$$

$$\approx f(\mathbf{z}_t; \mathbf{p}_{t,k}, \sigma^2) \int_{\mathbf{p}'} f(\mathbf{p}_{t,k}; \mathbf{p}', \sigma^2_{t,d}) f(\mathbf{p}'; \hat{\mathbf{p}}_{t-1,k}, \hat{\sigma}^2_{t-1,k}) \tag{7}$$

$$= f(\mathbf{z}_t; \mathbf{p}_{t,k}, \sigma^2) f(\mathbf{p}_{t,k}; \hat{\mathbf{p}}_{t-1,k}, \sigma^2_{t,d} + \hat{\sigma}^2_{t-1,k}). \tag{8}$$

The transition probability distribution $\Pr(\mathbf{p}_{t,k}|\mathbf{p}_{t-1,k})$ is a zero-mean Gaussian with variance $\hat{\sigma}^2_{t,d} = (1/\rho - 1)\hat{\sigma}^2_{t-1,k}$, where $\rho \in (0, 1]$ is some constant that we define to be the memory decay coefficient. In fact, $\hat{c}$ can be viewed as a count variable for the number of elements in each estimated cluster, subject to the decay factor ρ over time, if we further assume that $\sigma^2 = 1$, $\hat{c}_{t,k} \equiv 1/\hat{\sigma}^2_{t,k}$, and $\hat{c}_{t-1,k} \equiv 1/\hat{\sigma}^2_{t-1,k}$. The memory update equation can be formulated as follows:

$$\hat{c}_{t,k} = \mathop{\mathbb{E}}_{y_t}[\hat{c}_{t,k}|_{y_t}] = \rho\hat{c}_{t-1,k} + \hat{y}_{t,k}(1 - \hat{u}_{t,k}), \tag{9}$$

$$\hat{\mathbf{p}}_{t,k} = \mathop{\mathbb{E}}_{y_t}[\hat{\mathbf{p}}_{t,k}|_{y_t}] = \mathbf{z}_t \frac{\hat{y}_{t,k}(1 - \hat{u}_{t,k})}{\rho\hat{c}_{t-1,k} + 1} + \hat{\mathbf{p}}_{t-1,k}\left(1 - \frac{\hat{y}_{t,k}(1 - \hat{u}_{t,k})}{\rho\hat{c}_{t-1,k} + 1}\right). \tag{10}$$

$$\hat{w}_{t,k} = \mathop{\mathbb{E}}_{y_t}[\hat{w}_{t,k}|_{y_t}] = \hat{c}_{t,k} / \sum_l \hat{c}_{t,l}. \tag{11}$$

Adding and removing prototypes. Our prototype memory is a collection of tuples $(\hat{\mathbf{p}}_k, \hat{c}_k)$. We convert the probability of an observation belongs to a new cluster into a decision: if $\hat{u}_t$ exceeds a threshold α, we create a new cluster in our prototype memory. Due to the decay factor ρ, our $\hat{c}$ estimate of a cluster can decay to zero over time, which is appropriate for modeling nonstationary environments. In practice, we keep a maximum number of K clusters, and if the memory is full and we try to add another cluster at time t, we simply pop out the least relevant prototype $\mathbf{p}_{k'}$, where $k' = \arg\min(\hat{w}_k)$: $P_t = P_{t-1} \setminus \{(\hat{\mathbf{p}}_{k'}, \hat{c}_{k'})\} \cup \{(\mathbf{z}_t, 1)\}$.

3.2 Representation learning

A primary goal of our learning algorithm is to learn good visual representations through this online categorization process. In the beginning, the encoder network is randomly initialized, and the prototype memory will not produce accurate class predictions since the representations are not informative. Our overall representation learning objective has three terms:

$$\mathcal{L} = \mathcal{L}_{\text{self}} + \lambda_{\text{ent}}\mathcal{L}_{\text{ent}} + \lambda_{\text{new}}\mathcal{L}_{\text{new}}. \tag{12}$$

This loss function drives the learning of the main network parameters θ, as well as other learnable control parameters β, γ, and τ. We explain each term in detail below.

1. **Self-supervised loss:** Inspired by recent self-supervised representation learning approaches, we apply augmentations on $\mathbf{x}_t$, and encourage the clustering assignments to match across different views. Self-supervision follows three steps: First, the model makes a prediction on the augmented view, and obtains $\hat{y}$ and $\hat{u}$ (E-step). Secondly, it updates the prototype memory according to the prediction (M-step). To create a learning target, we query the original view again, and obtain $\tilde{y}$ to supervise the cluster assignment of the augmented view, $\hat{y}'$, as in distillation (Hinton et al., 2015).

$$\mathcal{L}_{\text{self}} = \frac{1}{T}\sum_t -\tilde{y}_t \log \hat{y}'_t. \tag{13}$$

   Note that both $\tilde{y}_t$ and $\hat{y}'_t$ are produced after the M-step so we can exclude the "unknown" class in the representation learning objective. We here introduce a separate temperature parameter $\tilde{\tau}$ to control the entropy of the mixture assignment $\tilde{y}_t$.

2. **Entropy loss:** In order to encourage more confident predictions we introduce another loss function $\mathcal{L}_{\text{ent}}$ that controls the entropy of the original prediction $\hat{y}$, produced in the initial E-step.

$$\mathcal{L}_{\text{ent}} = \frac{1}{T}\sum_t -\hat{y}_t \log \hat{y}_t. \tag{14}$$

3. **New cluster loss:** Lastly, our learning formulation also includes a loss for initiating new clusters $\mathcal{L}_{\text{new}}$. We define it to be a Beta prior on the expected $\hat{u}$, and we introduce a hyperparameter μ to control the expected number of clusters.

$$\mathcal{L}_{\text{new}} = \log \Pr(\mathbb{E}[\hat{u}]). \tag{15}$$

   This acts as a regularizer on the total number of prototypes: if the system is too aggressive in creating prototypes, then it does not learn to merge instances of the same class; while if it is too conservative, the representations can collapse to a trivial solution.

Relation to Online ProtoNet. The formulation of our probabilistic prototype memory is similar to Online ProtoNet (Ren et al., 2021). However, there are several main differences. First, we consider a decay term that can handle nonstationary mixtures, which is related to the variance of the transition probability. Second, our new cluster creation is unsupervised, whereas in Ren et al. (2021) only labeled examples lead to new clusters. Most importantly, our representation learning objective is also entirely unsupervised, whereas Ren et al. (2021) relies on a supervised loss.

Full algorithm. Let $\Theta = \{\theta, \beta, \gamma, \tau\}$ denote the union of the learnable parameters. Algorithm 1 outlines our proposed learning algorithm. The full list of hyperparameters are included in Appendix B.

Algorithm 1 Online Unsupervised Prototypical Learning

repeat
  $\mathcal{L}_{\text{self}} \leftarrow 0$, $p_{\text{new}} \leftarrow 0$.
  for $t \leftarrow 1 \dots T$ **do**
    Observe new input $\mathbf{x}_t$.
    Encode input, $\mathbf{z}_t \leftarrow h(\mathbf{x}_t; \theta)$.
    Compare to existing prototypes: $[\hat{u}_t, \hat{y}_t] \leftarrow$ E-step$(\mathbf{z}_t, P; \beta, \gamma, \tau)$.
    if $\hat{u}_t^0 < \alpha$ **then**
      Assign $\mathbf{z}_t$ to existing prototypes: $P \leftarrow$ M-step$(\mathbf{z}_t, P, \hat{u}_t, \hat{y}_t)$.
    else
      Recycle the least used prototype if P is full.
      Create a new prototype $P \leftarrow P \cup \{(\mathbf{z}_t, 1)\}$.
    end if
    Compute pseudo-labels: $[_, \tilde{y}_t] \leftarrow$ E-step$(\mathbf{z}_t, P; \beta, \gamma, \tilde{\tau})$.
    Augment a view: $\mathbf{x}'_t \leftarrow$ augment$(\mathbf{x}_t)$.
    Encode the augmented view: $\mathbf{z}'_t \leftarrow h(\mathbf{x}'_t; \theta)$.
    Compare the augmented view to existing prototypes: $[_, \hat{y}'_t] \leftarrow$ E-step$(\mathbf{z}'_t, P; \beta, \gamma, \tau)$.
    Compute the self-supervision loss: $\mathcal{L}_{\text{self}} \leftarrow \mathcal{L}_{\text{self}} - \frac{1}{T}\tilde{y}_t \log \hat{y}'_t$.
    Compute the entropy loss: $\mathcal{L}_{\text{ent}} \leftarrow \mathcal{L}_{\text{ent}} - \frac{1}{T}\hat{y}_t \log \hat{y}_t$.
    Compute the average probability of creating new prototypes, $p_{\text{new}} \leftarrow p_{\text{new}} + \frac{1}{T}\hat{u}_t$.
  end for
  Compute the new cluster loss: $\mathcal{L}_{\text{new}} \leftarrow -\log \Pr(p_{\text{new}})$.
  Sum up losses: $\mathcal{L} \leftarrow \mathcal{L}_{\text{self}} + \lambda_{\text{ent}}\mathcal{L}_{\text{ent}} + \lambda_{\text{new}}\mathcal{L}_{\text{new}}$.
  Update parameters: $\Theta \leftarrow$ optimize$(\mathcal{L}, \Theta)$.
until convergence
return Θ

It is worth noting that if we create a new prototype every time step, then OUPN is similar to a standard contrastive learning with an instance-based InfoNCE loss (Chen et al., 2020a; He et al., 2020); therefore it can be viewed as a generalization of this approach. Additionally, all the losses can be computed online without having to store any examples beyond the collection of prototypes.

4 Experiments

In this section, we evaluate our proposed learning algorithm on a set of visual learning tasks and evaluate the quality of the output categories. Different from prior work on visual representation learning, we focus on online non-iid image sequences to highlight the merit of method, since that is the primary scenario of interest.

Implementation details. Throughout our experiments, we make two changes on the model inference procedure defined above. First, we use cosine similarity instead of negative squared Euclidean distance for computing the mixture logits, because cosine similarity is bounded and is found to be more stable to train. Second, when we perform cluster inference, we treat the mixing coefficients w_k as constant and uniform as otherwise we find that the representations may collapse into a single large cluster.

Evaluation setup. Although our training is entirely unsupervised, we can evaluate our model under both *unsupervised* and *supervised* settings. During evaluation we present our model a sequence of new images (labeled or unlabeled) and we would like to see how well the model is able to output a successful grouping of the inputs. For the unsupervised setting, there is no class label associated with each image, and our model g directly predicts $\hat{y}_t = g(\mathbf{x}_{1:t})$. For the supervised setting, following Ren et al. (2021), we give the model the groundtruth label y_t after the prediction of $\hat{y}_t$, i.e. $\hat{y}_t = g(\mathbf{x}_{1:t}, y_{1:t-1})$.

Figure 2: An example subsequence of the *RoamingRooms* dataset, consisting of consecutive glimpses of an online agent roaming in an indoor environment, and the task is to recognize the object instances.

	AMI	AP
Supervised		
Online ProtoNet (Ren et al., 2021)	79.02	89.94
Unsupervised		
Random Network	28.25	11.68
SimCLR (Chen et al., 2020a)	50.03	52.98
SwAV (Caron et al., 2020)	42.70	37.31
SwAV+Queue (Caron et al., 2020)	48.31	50.40
OUPN (Ours)	**78.16**	**84.86**

Table 1: Novel object recognition performance on *RoamingRooms*

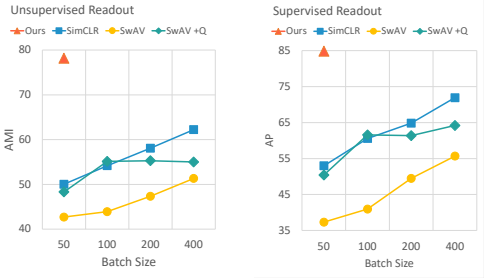

Figure 3: Comparison to SimCLR and SwAV with larger batch sizes on *RoamingRooms*

Evaluation metrics. We use the following unsupervised and supervised metrics.

- **Adjusted mutual information (AMI):** In the *unsupervised* setting, we use the mutual information metric to evaluate the similarity between our prediction $\{\hat{y}_1, \ldots, \hat{y}_T\}$ the groundtruth class ID $\{y_1, \ldots, y_T\}$. Since the online clustering method admits a threshold parameter α to control the number of output clusters, therefore for each model we sweep the value of α to maximize the AMI score, to make the score threshold-invariant: $\text{AMI}_{\max} = \max_\alpha \text{AMI}(y, \hat{y}(\alpha))$.
- **Average precision (AP):** In the *supervised* setting, we followed the evaluation procedure in Ren et al. (2021) and used average precision, which combines both accuracy for predicting known classes as well as unknown ones.

Competitive methods. We compare our proposed model with the following state-of-the-art self-supervised visual representation learning methods. Since none of these competitive methods are designed to output classes with a few examples, we use our online clustering procedure to readout from these learned representations.

- **SimCLR** (Chen et al., 2020a) is a contrastive learning method with an instance-based objective that tries to classify an image instance among other augmented views of the same batch of instances. It relies on a large batch size and is often trained on well-curated datasets such as ImageNet (Deng et al., 2009).
- **SwAV** (Caron et al., 2020) is a contrastive learning method with a clustering-based objective. It has a stronger performance than SimCLR on ImageNet. The clustering is achieved through Sinkhorn-Knopp which assumes balanced assignment, and prototypes are learned by gradient descent.
- **SwAV+Queue** is a SwAV variant with an additional example queue. This setup is proposed in (Caron et al., 2020) to deal with small training batches. A feature queue that accumulates instances across batches allows the clustering process to access more data points. The queue size is set to 2000.

4.1 Indoor Home Environments

We first evaluate the algorithm using the *RoamingRooms* dataset (Ren et al., 2021) where the images are collected from indoor environments (Chang et al., 2017) using a random walking agent. The dataset contains 1.22M image frames and 7K instance classes from 6.9K random walk episodes collected using 1.2K panoramas. Each frame has an object annotation with its segmentation mask and the task here is to recognize the object instance IDs. Since our task is classification, we also send the segmentation mask as an additional channel in the input. An example episode is shown in Fig. 2.

Results. Our main results are shown in Table 1. Although both SimCLR and SwAV have shown promising results on large batch learning on ImageNet, their performances here are relatively weak compared to the supervised baseline. Adding a queue slightly improves SwAV; however, since the examples in the queue cannot be used to compute gradients, the nonstationary distribution still hampers gradient updates. In contrast, our method OUPN shows impressive performance on this benchmark: it almost matches the supervised learner in AMI, and reached almost 95% of the performance of the supervised learner in AP.

Figure 4: Image retrieval results on *RoamingRooms*. In each row, the leftmost image is the query image, and top-9 retrieved images are shown to its right. For each retrieval its cosine similarity score is in the top left; a green border signifies a correct retrieval (matching the query instance), red is a false positive, yellow a miss. Recall is the proportion of instances retrieved within the top-9.

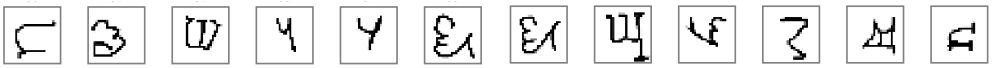

Figure 5: An example subsequence of an episode sampled from the *RoamingOmniglot* dataset

To illustrate the impact of our small batch episodes, we increase the batch size for SimCLR and SwAV, from 50 to 400, at the cost of using multiple GPUs training in parallel. The results are shown in Fig. 3. Results indicate that increasing the batch size can improve these baselines, which matches our expectation. Nevertheless, our method using a batch size of 50 is still able to outperform SimCLR and SwAV using a batch size of 400, which takes $8\times$ computational resource compared to ours. Note that the large batch experiments are designed to provide the best setting for SimCLR and SwAV to succeed. We do not intend to run our model with larger batch size since our prototype memory is a sequential module. Moreover, keeping the batch size smaller allows quicker online adaptation and less memory consumption.

Comparison to iid modes of SimCLR and SwAV. The original SimCLR and SwAV were designed to train on iid data. To study the effects of this assumption, we implemented an approximation to an iid distribution by using a large random queue that shuffles the image frames. As in the study shown in Fig. 6, we again vary the batch size for these competitive methods. Both SimCLR and SwAV thrive with iid data; the gains of iid over non-iid can be seen by comparing Fig. 6 to Fig. 3. Larger batches help both methods again here. Interestingly, our method using a batch size of 50 non-iid data again outperforms both methods using a batch size of 400 of iid data.

Visualization on image retrieval. To verify the usefulness of the learned representation, we ran an image retrieval visualization over 100 test episodes in Fig. 4, and we also provided the cosine similarity score for each retrieved image. The top retrieved images are all from the same instance of the query image, and our model sometimes achieves perfect recall. This confirms that our model can handle a certain level of view angle changes. We also investigated the missed examples and we found that these are taken from more distinct view angles. However, note that we only computed the pairwise similarity score for retrieval, and therefore it is possible that the actual clusters are more inclusive as the model incrementally computes the average embedding as the prototype. More visualizations are provided in Appendix D.

4.2 Handwritten Characters

We also evaluated our method on a different task: recognizing handwritten characters from Omniglot (Lake et al., 2015). In this experiment, images are not organized in a video-like sequence, and models have to reason more about conceptual similarity between images in order to learn grouping. Furthermore, since this is a more controllable setup, we can test our hypothesis concerning sensitivity to class imbalance by performing manipulations on the episode distribution.

Our episodes are sampled from the *RoamingOmniglot* dataset (Ren et al., 2021). An episode involves several different *contexts*, each consisting of a set of classes, and in each context, classes are sampled from a Chinese restaurant process. We use 150-frame episodes with 5 contexts.

Results. The results are reported in Table 2. Our model is able to significantly reduce the gap between supervised and unsupervised models, outperforming SimCLR and SwAV.

Effect of imbalanced distribution. We further study the effect of imbalanced cluster sizes by manipulating the class distribution in the training episodes. In the first setting, we randomly replace

Table 2: Results on *RoamingOmniglot*

	AMI	AP
Supervised		
Pretrain-Supervised	84.48	93.83
Online ProtoNet (Ren et al., 2021)	89.64	92.58
Unsupervised		
Random Network	17.66	17.01
SimCLR (Chen et al., 2020a)	59.06	73.50
SwAV (Caron et al., 2020)	62.09	75.93
SwAV+Queue (Caron et al., 2020)	67.25	81.96
OUPN (Ours)	**84.42**	**92.84**

Table 3: Results on *RoamingImageNet*

	AMI	AP
Supervised		
Pretrain-Supervised	29.44	24.39
Online ProtoNet (Ren et al., 2021)	29.73	25.38
Unsupervised		
Random Network	4.55	2.65
SimCLR (Chen et al., 2020a)	6.87	12.25
SwAV (Caron et al., 2020)	9.87	5.23
SwAV+Queue (Caron et al., 2020)	10.61	4.83
OUPN (Ours)	**19.03**	**15.05**

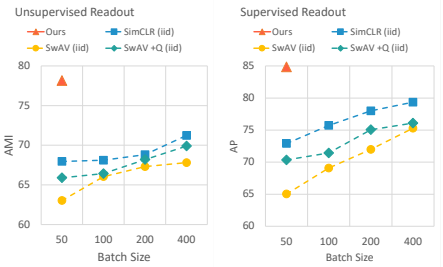

Figure 6: Comparison to IID modes of SimCLR and SwAV on *RoamingRooms* with larger training batch sizes.

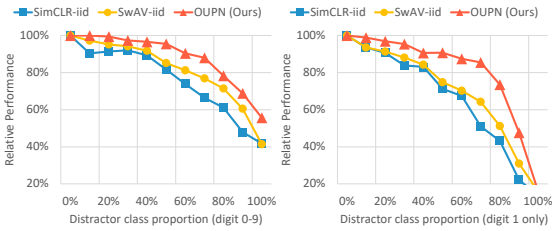

Figure 7: Robustness to imbalanced distributions by adding distractors (Omniglot mixed with MNIST images). Performance is relative to the original performance and a random baseline.

Omniglot images with MNIST digits, with probability from 0% to 100%. For example, at 50% rate, an MNIST digit is over 300 times more likely to appear compared to any Omniglot character class, so the episodes are composed of half frequent classes and half infrequent classes. In the second setting, we randomly replace Omniglot images with MNIST digit 1 images, which makes the imbalance even greater. We compared our method to SimCLR and SwAV in the iid setup, since this is the scenario they were designed for. Results of the two settings are shown in Fig. 7, and our method is shown to be more robust under imbalanced distribution than SimCLR and SwAV. Compared to clustering-based methods like SwAV, our prototypes can be dynamically created and updated with no constraints on the number of elements per cluster. Compared to instance-based methods like SimCLR, our prototypes also samples the contrastive pairs more equally during learning. We hypothesize that these model aspects contribute to the differences in robustness.

Visualization of learned categories. We visualize the learned categories in *RoamingOmniglot* using t-SNE (Van der Maaten & Hinton, 2008), and the results are shown in the Appendix D.

4.3 IMAGENET IMAGES

Lastly, we evaluate on episodes composed of ImageNet images, and using the *RoamingImageNet* dataset (Ren et al., 2021), which has a structure analogous to *RoamingOmniglot*. This dataset's scale resembles *RoamingRooms* (around 1.3 million images), but here we focus on semantic object classes instead of instance classes. We use this benchmark as a stress test: it does not fit the target scenario of an natural online exploration, but it still provides a test whether a method can learn semantic concepts with a large intra-class variance. We developed a form of non-iid episodes, consisting of 48 frames drawn with temporal dependence from 3 contexts. Results are shown in Table 3. Although ImageNet is a more challenging benchmark, our model is still considerably better than SimCLR and SwAV.

5 CONCLUSION

Our goal is to develop learning procedures for real-world agents who operate online and in structured, nonstationary environments. Toward this goal, we develop an online unsupervised algorithm for discovering visual representations and categories. Unlike standard self-supervised learning, our algorithm embeds category formation in a probabilistic clustering module that is jointly learned with the representation encoder. Our clustering is more flexible and supports learning of new categories with very few examples. At the same time, we leverage self-supervised learning to acquire semantically meaningful representations. Our method is evaluated in both synthetic and realistic image sequences and it outperforms state-of-the-art self-supervised learning algorithms for both the non-iid sequences we are interested in as well as sequences transformed to be iid to better match assumptions of the learning algorithms.

REFERENCES

Kelsey R. Allen, Evan Shelhamer, Hanul Shin, and Joshua B. Tenenbaum. Infinite mixture prototypes for few-shot learning. In *ICML*, 2019. 3

Antreas Antoniou and Amos J. Storkey. Assume, augment and learn: Unsupervised few-shot meta-learning via random labels and data augmentation. *CoRR*, abs/1902.09884, 2019. 3

Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *ICLR*, 2020. 1, 2

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. In *NeurIPS*, 2020. 3

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, 2018. 1, 2

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 1, 2, 7, 9, 16

Francisco M. Castro, Manuel J. Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *ECCV*, 2018. 3

Angel X. Chang, Angela Dai, Thomas A. Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from RGB-D data in indoor environments. In *3DV*, 2017. 7

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020a. 1, 2, 6, 7, 9, 16

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. 2021. 2

Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297, 2020b. 2

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NIPS*, 2013. 2

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 7

Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In *ICML*, 2018. 3

Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *CVPR*, 2018. 3

Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. Boosting few-shot visual learning with self-supervision. In *ICCV*, 2019. 3

Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. Neural expectation maximization. In *NIPS*, 2017. 2

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent - A new approach to self-supervised learning. In *NeurIPS*, 2020. 2

Tyler L. Hayes, Nathan D. Cahill, and Christopher Kanan. Memory efficient experience replay for streaming learning. In *ICRA*, 2019. 3

Tyler L. Hayes, Kushal Kafle, Robik Shrestha, Manoj Acharya, and Christopher Kanan. REMIND your neural network to prevent catastrophic forgetting. In *ECCV*, 2020. 3

Jiawei He, Andreas M. Lehrmann, Joseph Marino, Greg Mori, and Leonid Sigal. Probabilistic video generation using holistic attribute control. In *ECCV*, 2018. 3

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 1, 2, 6

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. 5

Kyle Hsu, Sergey Levine, and Chelsea Finn. Unsupervised learning via meta-learning. In *ICLR*, 2019. 3

Gabriel Huang, Hugo Larochelle, and Simon Lacoste-Julien. Centroid networks for few-shot clustering and unsupervised few-shot classification. *CoRR*, abs/1902.08605, 2019. 3

Michael C. Hughes and Erik B. Sudderth. Memoized online variational inference for dirichlet process mixture models. In *NIPS*, 2013. 3

Khurram Javed and Martha White. Meta-learning representations for continual learning. In *NeurIPS*, 2019. 3

Ghassen Jerfel, Erin Grant, Tom Griffiths, and Katherine A. Heller. Reconciling meta-learning and continual learning with online mixtures of tasks. In *NeurIPS*, 2019. 3

Matthew J. Johnson, David Duvenaud, Alexander B. Wiltschko, Ryan P. Adams, and Sandeep R. Datta. Composing graphical models with neural networks for structured representations and fast inference. In *NIPS*, 2016. 3

Weonyoung Joo, Wonsung Lee, Sungrae Park, and Il-Chul Moon. Dirichlet variational autoencoder. *Pattern Recognit.*, 107:107514, 2020. 2, 16

Siavash Khodadadeh, Ladislau Bölöni, and Mubarak Shah. Unsupervised meta-learning for few-shot image classification. In *NeurIPS*, 2019. 3

Chris Dongjoo Kim, Jinseo Jeong, and Gunhee Kim. Imbalanced continual learning with partitioning reservoir sampling. In *ECCV*, 2020. 3

Rahul G. Krishnan, Uri Shalit, and David A. Sontag. Deep kalman filters. *CoRR*, abs/1511.05121, 2015. 3

Brenden M. Lake, Gautam K. Vallabha, and James L. McClelland. Modeling unsupervised perceptual category learning. *IEEE Trans. Auton. Ment. Dev.*, 1(1):35–43, 2009. 3

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015. 3, 8

Fei-Fei Li, Robert Fergus, and Pietro Perona. A bayesian approach to unsupervised one-shot learning of object categories. In *ICCV*, 2003. 3

Junnan Li, Pan Zhou, Caiming Xiong, Richard Socher, and Steven C. H. Hoi. Prototypical contrastive learning of unsupervised representations. 2021. 2

Bradley C Love, Douglas L Medin, and Todd M Gureckis. Sustain: a network model of category learning. *Psychological review*, 111(2):309, 2004. 3

Michael L. Mack, Bradley C. Love, and Alison R. Preston. Dynamic updating of hippocampal object representations reflects new conceptual knowledge. *Proceedings of the National Academy of Sciences*, 113(46):13203–13208, 2016. ISSN 0027-8424. doi: 10.1073/pnas.1614048113. 3

Carlos Medina, Arnout Devos, and Matthias Grossglauser. Self-supervised prototypical transfer learning for few-shot classification. *CoRR*, abs/2006.11325, 2020. 3

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *CVPR*, 2020. 2

Boris N. Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. TADAM: task dependent adaptive metric for improved few-shot learning. In *NIPS*, 2018. 15, 16

A. Emin Orhan, Vaibhav V. Gupta, and Brenden M. Lake. Self-supervised learning through the eyes of a child. In *NeurIPS*, 2020. 2

Deepak Pathak, Ross Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. In *CVPR*, 2017. 2

Rafael C. Pinto and Paulo Martins Engel. A fast incremental gaussian mixture model. *CoRR*, abs/1506.04422, 2015. 3

Dushyant Rao, Francesco Visin, Andrei A. Rusu, Razvan Pascanu, Yee Whye Teh, and Raia Hadsell. Continual unsupervised representation learning. In *NeurIPS*, 2019. 2, 3, 16

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, 2017. 3

Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel. Meta-learning for semi-supervised few-shot classification. In *ICLR*, 2018. 3

Mengye Ren, Michael L. Iuzzolino, Michael C. Mozer, and Richard S. Zemel. Wandering within a world: Online contextualized few-shot learning. In *ICLR*, 2021. 2, 3, 6, 7, 8, 9, 17

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *NIPS*, 2017. 3

Mingzhou Song and Hongbin Wang. Highly efficient incremental estimation of gaussian mixture models for online data stream clustering. In *Intelligent Computing: Theory and Applications III*, volume 5803, pp. 174–183. International Society for Optics and Photonics, 2005. 3

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, 2020. 2

Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, and Hugo Larochelle. Meta-dataset: A dataset of datasets for learning to learn from few examples. In *ICLR*, 2020. 3

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. 2

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008. 9, 18, 19

Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *NIPS*, 2016. 15

Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *ICCV*, 2015. 2

Yuwen Xiong, Mengye Ren, Wenyuan Zeng, and Raquel Urtasun. Self-supervised representation learning from flow equivariance. In *ICCV*, 2021. 2

Xiaohang Zhan, Jiahao Xie, Ziwei Liu, Yew-Soon Ong, and Chen Change Loy. Online deep clustering for unsupervised representation learning. In *CVPR*, 2020. 2

Yizhe Zhu, Martin Renqiang Min, Asim Kadav, and Hans Peter Graf. S3VAE: self-supervised sequential VAE for representation disentanglement and data generation. In *CVPR*, 2020. 2, 3

A METHOD DERIVATION

A.1 E-STEP

Inferring cluster assignments. The categorical variable $\hat{y}$ infers the cluster assignment of the current input example with regard to the existing clusters.

$$\hat{y}_{t,k} = \Pr(y_t = k|\mathbf{z}_t, u = 0) \tag{16}$$
$$= \frac{\Pr(\mathbf{z}_t|y_t = k, u = 0)\Pr(y_t = k)}{\Pr(\mathbf{z}_t, u = 0)} \tag{17}$$
$$= \frac{w_k f(\mathbf{z}_t; \mathbf{p}_{t,k}, \sigma^2)}{\sum_{k'} w_{k'} f(\mathbf{z}_t; \mathbf{p}_{t,k'}, \sigma^2)} \tag{18}$$
$$= \frac{\exp(\log w_k - d(\mathbf{z}_t, \mathbf{p}_{t,k})/2\sigma^2)}{\sum_{k'} \exp(\log w'_k - d(\mathbf{z}_t, \mathbf{p}_{t,k'})/2\sigma^2)} \tag{19}$$
$$= \mathrm{softmax}\left(\log w_k - d(\mathbf{z}_t, \mathbf{p}_{t,k})/\tau\right), \tag{20}$$
$$= \mathrm{softmax}(v_{t,k}), \tag{21}$$

where w_k is the mixing coefficient of cluster k and $d(\cdot,\cdot)$ is the distance function, and $v_{t,k}$ is the logits. In our experiments, w_k's are kept as constant and τ is an independent learnable parameter.

Inference on unknown classes. The binary variable $\hat{u}$ estimates the probability that the current input belongs to a new cluster:

$$\hat{u}_t = \Pr(u_t = 1|\mathbf{z}_t) \tag{22}$$
$$= \frac{z_0 u_0}{z_0 u_0 + \sum_k w_k f(\mathbf{z}_t; \mathbf{p}_{t,k}, \sigma^2)(1 - u_0)} \tag{23}$$
$$= \frac{1}{1 + \frac{1}{z_0 u_0} \sum_k w_k f(\mathbf{z}_t; \mathbf{p}_{t,k}, \sigma^2)(1 - u_0)} \tag{24}$$
$$= \frac{1}{1 + \exp(\log(\frac{1}{z_0 u_0} \sum_k w_k f(\mathbf{z}_t; \mathbf{p}_{t,k}, \sigma^2)(1 - u_0)))} \tag{25}$$
$$= \frac{1}{1 + \exp(-\log(z_0) - \log(u_0) + \log(1 - u_0) + \log(\sum_k w_k f(\mathbf{z}_t; \mathbf{p}_{t,k}, \sigma^2))} \tag{26}$$
$$= \frac{1}{1 + \exp(-s + \log(\sum_k \exp(\log(w_k) - d(\mathbf{z}_t, \mathbf{p}_{t,k})/2\sigma^2)))} \tag{27}$$
$$= \mathrm{sigmoid}(s - \log(\sum_k \exp(\log(w_k) - d(\mathbf{z}_t, \mathbf{p}_{t,k})/2\sigma^2))) \tag{28}$$
$$= \mathrm{sigmoid}(s - \log(\sum_k \exp(v_{t,k}))), \tag{29}$$

where $s = \log(z_0) + \log(u_0) - \log(1 - u_0) + m\log(\sigma) + m\log(2\pi)/2$ and m is the input dimension. In our implementation, we use $\max$ here instead of $\mathrm{logsumexp}$ since we found $\max$ leads to better and more stable training performance empirically. It can be derived as a lower bound:

$$\hat{u}_t = \mathrm{sigmoid}(s - \log(\sum_k \exp(\log(w_k) - d(\mathbf{z}_t, \mathbf{p}_{t,k})/2\sigma^2))) \tag{30}$$
$$\geq \mathrm{sigmoid}(s - \log(\max_k \exp(-d(\mathbf{z}_t, \mathbf{p}_{t,k})/2\sigma^2))) \tag{31}$$
$$= \mathrm{sigmoid}(s + \min_k d(\mathbf{z}_t, \mathbf{p}_{t,k})/2\sigma^2) \tag{32}$$
$$= \mathrm{sigmoid}((\min_k d(\mathbf{z}_t, \mathbf{p}_{t,k}) - \beta)/\gamma), \tag{33}$$

where $\beta = -2s\sigma^2$, $\gamma = 2\sigma^2$. To make learning more flexible, we directly make β and γ as independent learnable parameters so that we can control the confidence level for predicting unknown classes.

A.2 M-STEP

Here we infer the posterior distribution of the prototypes $\Pr(\mathbf{p}_{t,k}|\mathbf{z}_{1:t})$. We formulate an efficient recursive online update, similar to Kalman filtering, by incorporating the evidence of the current input $\mathbf{z}_t$ and avoiding re-clustering the entire input history. We define $\hat{\mathbf{p}}_{t,k}$ as the estimate of the posterior mean of the k-th cluster at time step t, and $\hat{\sigma}^2_{t,k}$ is the estimate of the posterior variance.

Updating prototypes. Suppose that in the E-step we have determined that $y_t = k$. Then the posterior distribution of the k-th cluster after observing $\mathbf{z}_t$ is:

$$\Pr(\mathbf{p}_{t,k}|\mathbf{z}_{1:t}, y_t = k) \tag{34}$$

$$\propto \Pr(\mathbf{z}_t|\mathbf{p}_{t,k}, y_t = k)\Pr(\mathbf{p}_{t,k}|\mathbf{z}_{1:t-1}) \tag{35}$$

$$= \Pr(\mathbf{z}_t|\mathbf{p}_{t,k}, y_t = k)\int_{\mathbf{p}'}\Pr(\mathbf{p}_{t,k}|\mathbf{p}_{t-1,k} = \mathbf{p}')\Pr(\mathbf{p}_{t-1,k} = \mathbf{p}'|\mathbf{z}_{1:t-1}) \tag{36}$$

$$\approx f(\mathbf{z}_t; \mathbf{p}_{t,k}, \sigma^2)\int_{\mathbf{p}'} f(\mathbf{p}_{t,k}; \mathbf{p}', \sigma^2_{t,d})f(\mathbf{p}'; \hat{\mathbf{p}}_{t-1,k}, \hat{\sigma}^2_{t-1,k}) \tag{37}$$

$$= f(\mathbf{z}_t; \mathbf{p}_{t,k}, \sigma^2)f(\mathbf{p}_{t,k}; \hat{\mathbf{p}}_{t-1,k}, \sigma^2_{t,d} + \hat{\sigma}^2_{t-1,k}). \tag{38}$$

If we assume that the transition probability distribution $\Pr(\mathbf{p}_{t,k}|\mathbf{p}_{t-1,k})$ is a zero-mean Gaussian with variance $\sigma^2_{t,d} = (1/\rho - 1)\hat{\sigma}^2_{t-1,k}$, where $\rho \in (0, 1]$ is some constant that we defined to be the memory decay coefficient, then the posterior estimates are:

$$\hat{\mathbf{p}}_{t,k}|_{y_t=k} = \frac{\mathbf{z}_t\hat{\sigma}^2_{t-1,k}/\rho + \hat{\mathbf{p}}_{t-1,k}\sigma^2}{\sigma^2 + \hat{\sigma}^2_{t-1,k}/\rho}, \quad \hat{\sigma}^2_{t,k}|_{y_t=k} = \frac{\sigma^2\hat{\sigma}^2_{t-1,k}/\rho}{\sigma^2 + \hat{\sigma}^2_{t-1,k}/\rho}. \tag{39}$$

If $\sigma^2 = 1$, and $\hat{c}_{t,k} \equiv 1/\hat{\sigma}^2_{t,k}$, $\hat{c}_{t-1,k} \equiv 1/\hat{\sigma}^2_{t-1,k}$, it turns out we can formulate the update equation as follows, and $\hat{c}_{t,k}$ can be viewed as a count variable for the number of elements in each estimated cluster, subject to the decay factor ρ over time:

$$\hat{c}_{t,k}|_{y_t=k} = \rho\hat{c}_{t-1,k} + 1, \tag{40}$$

$$\hat{\mathbf{p}}_{t,k}|_{y_t=k} = \mathbf{z}_t\frac{1}{\hat{c}_{t,k}|_{y_t=k}} + \hat{\mathbf{p}}_{t-1,k}\frac{\rho\hat{c}_{t-1,k}}{\hat{c}_{t,k}|_{y_t=k}}. \tag{41}$$

If $y_t \neq k$, then the prototype posterior distribution simply gets diffused at timestep t:

$$\Pr(\mathbf{p}_{t,k}|z_{1:t}, y_t \neq k) \approx f(\mathbf{p}_{t,k}; \hat{\mathbf{p}}_{t-1,k}, \hat{\sigma}^2_{t-1,k}/\rho) \tag{42}$$

$$\hat{c}_{t,k}|_{y_t\neq k} = \rho\hat{c}_{t-1,k}, \tag{43}$$

$$\hat{\mathbf{p}}_{t,k}|_{y_t\neq k} = \hat{\mathbf{p}}_{t-1,k}. \tag{44}$$

Finally, our posterior estimates at time t are computed by taking the expectation over y_t:

$$\hat{c}_{t,k} = \mathop{\mathbb{E}}_{y_t}[\hat{c}_{t,k}|_{y_t}] \tag{45}$$

$$= \hat{c}_{t,k}|_{y_t=k}\Pr(y_t = k|\mathbf{z}_t) + \hat{c}_{t,k}|_{y_t\neq k}\Pr(y_t \neq k|\mathbf{z}_t) \tag{46}$$

$$= (\rho\hat{c}_{t-1,k} + 1)\hat{y}_{t,k}(1 - \hat{u}_{t,k}) + \rho\hat{c}_{t-1,k}(1 - \hat{y}_{t,k}(1 - \hat{u}_{t,k})), \tag{47}$$

$$= \rho\hat{c}_{t-1,k} + \hat{y}_{t,k}(1 - \hat{u}_{t,k}), \tag{48}$$

$$\hat{\mathbf{p}}_{t,k} = \mathop{\mathbb{E}}_{y_t}[\hat{\mathbf{p}}_{t,k}|_{y_t}] \tag{49}$$

$$= \hat{\mathbf{p}}_{t,k}|_{y_t=k}\Pr(y_t = k|\mathbf{z}_t) + \hat{\mathbf{p}}_{t,k}|_{y_t\neq k}\Pr(y_t \neq k|\mathbf{z}_t) \tag{50}$$

$$= \left(\mathbf{z}_t\frac{1}{\hat{c}_{t,k}|_{y_t=k}} + \hat{\mathbf{p}}_{t-1,k}\frac{\rho\hat{c}_{t-1,k}}{\hat{c}_{t,k}|_{y_t=k}}\right)\hat{y}_{t,k}(1 - \hat{u}_{t,k}) + \hat{\mathbf{p}}_{t-1,k}(1 - \hat{y}_{t,k}(1 - \hat{u}_{t,k})) \tag{51}$$

$$= \mathbf{z}_t\frac{\hat{y}_{t,k}(1 - \hat{u}_{t,k})}{\rho\hat{c}_{t-1,k} + 1} + \hat{\mathbf{p}}_{t-1,k}\left(1 - \hat{y}_{t,k}(1 - \hat{u}_{t,k}) + \hat{y}_{t,k}(1 - \hat{u}_{t,k})\frac{\rho\hat{c}_{t-1,k}}{\rho\hat{c}_{t-1,k} + 1}\right) \tag{52}$$

$$= \mathbf{z}_t\frac{\hat{y}_{t,k}(1 - \hat{u}_{t,k})}{\rho\hat{c}_{t-1,k} + 1} + \hat{\mathbf{p}}_{t-1,k}\left(1 - \frac{\hat{y}_{t,k}(1 - \hat{u}_{t,k})}{\rho\hat{c}_{t-1,k} + 1}\right). \tag{53}$$

Since $\hat{c}_{t,k}$ is also our estimate on the number of elements in each cluster, we can use it to estimate the mixture weights,

$$\hat{w}_{t,k} = \frac{\hat{c}_{t,k}}{\sum_{k'} \hat{c}_{t,k}}. \tag{54}$$

Note that in our experiments the mixture weights are not used and we assume that each cluster has an equal mixture probability.

B Experiment Details

We provide additional implementation details in Tab. 4, 5 and 6.

Table 4: Hyperparameter settings for *RoamingRooms*

Hyperparameter	Values
Random crop area range	0.08 - 1.0
Random color strength	0.5
Backbone	ResNet-12 (Oreshkin et al., 2018)
Num channels	[32, 64, 128, 256]
τ init	0.1
β init	-12.0
γ init	1.0
Num prototypes K	150
Memory decay ρ	0.995
Sequence length / batch size	50 (100 eval)
Beta mean μ	0.5
Entropy loss λ_{ent}	0.0
New cluster loss λ_{new}	0.5
Threshold α	0.5
Pseudo label temperature ratio $\tilde{\tau}/\tau$	0.1
Learning rate schedule	[40k, 60k, 80k]
Learning rate	[1e-3, 1e-4, 1e-5]
Optimizer	Adam

Table 5: Hyperparameter settings for *RoamingOmniglot*

Hyperparameter	Values
Random crop area range	0.08 - 1.0
Random color	None
Backbone	Conv-4 (Vinyals et al., 2016)
Num channels	[64, 64, 64, 64]
τ init	0.1
β init	-12.0
γ init	1.0
Num prototypes K	150
Memory decay ρ	0.995
Sequence length / batch size	150
Beta mean μ	0.5
Entropy loss λ_{ent}	1.0
New cluster loss λ_{new}	1.0
Threshold α	0.5
Pseudo label temperature ratio $\tilde{\tau}/\tau$	0.2
Learning rate schedule	[40k, 60k, 80k]
Learning rate	[1e-3, 1e-4, 1e-5]
Optimizer	Adam

Table 7: Unsupervised iid learning on Omniglot using an MLP

Method	3-NN Error	5-NN Error	10-NN Error
VAE (Joo et al., 2020)	92.34±0.25	91.21±0.18	88.79±0.35
SBVAE (Joo et al., 2020)	86.90±0.82	85.10±0.89	82.96±0.64
DirVAE (Joo et al., 2020)	76.55±0.23	73.81±0.29	70.95±0.29
CURL (Rao et al., 2019)	78.18±0.47	75.41±0.34	72.51±0.46
SimCLR (Chen et al., 2020a)	44.35±0.55	42.99±0.55	44.93±0.55
SwAV (Caron et al., 2020)	**42.66±0.55**	**42.08±0.55**	44.78±0.55
OUPN (Ours)	43.75±0.55	**42.13±0.55**	**43.88±0.55**

Table 6: Hyperparameter settings for *RoamingImageNet*

Hyperparameter	Values
Random crop area range	0.08 - 1.0
Random color strength	0.5
Backbone	ResNet-12 (Oreshkin et al., 2018)
Num channels	[32, 64, 128, 256]
τ init	0.1
β init	-12.0
γ init	1.0
Num prototypes K	600
Memory decay ρ	0.99
Sequence length / batch size	48
Beta mean μ	0.5
Entropy loss λ_{ent}	0.5
New cluster loss λ_{new}	0.5
Threshold α	0.5
Pseudo label temperature ratio $\tilde{\tau}/\tau$	0.0 (i.e. one-hot pseudo labels)
Learning rate schedule	[40k, 60k, 80k]
Learning rate	[1e-3, 1e-4, 1e-5]
Optimizer	Adam

B.1 Metric Details

For each method, we used the same nearest centroid algorithm for online clustering. For unsupervised readout, at each timestep, if the closest centroid is within threshold α, then we assign the new example to the cluster, otherwise we create a new cluster. For supervised readout, we assign examples based on the class label, and we create a new cluster if and only if the label is a new class. Both readout will provide us a sequence of class IDs, and we will use the following metrics to compare our predicted class IDs and groundtruth class IDs. Both metrics are designed to be threshold invariant.

AMI: For unsupervised evaluation, we consider the adjusted mutual information score. Suppose we have two clustering $U = \{U_i\}$ and $V = \{V_j\}$, and U_i and V_j are set of example IDs, and N is the total number of examples. U and V can be viewed as discrete probability distribution over cluster IDs. Therefore, the mutual information score between U and V is:

$$\text{MI}(U, V) = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} \frac{|U_i \cap V_j|}{N} \log \left(\frac{N|U_i \cap V_j|}{|U_i||V_j|} \right) \tag{55}$$

$$= \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{n_{ij}}{N} \log \left(\frac{N n_{ij}}{a_i b_j} \right). \tag{56}$$

Table 8: Effect of mem. size K

K	*RoamingOmniglot* AMI	*RoamingOmniglot* AP	*RoamingRooms* AMI	*RoamingRooms* AP
50	89.19	95.12	75.33	82.42
100	**90.54**	95.83	76.70	83.51
150	90.24	**95.92**	77.07	84.00
200	90.36	95.68	76.81	**84.45**
250	89.87	95.69	**77.83**	84.33

Table 9: Effect of decay rate ρ

ρ	*RoamingOmniglot* AMI	*RoamingOmniglot* AP	*RoamingRooms* AMI	*RoamingRooms* AP
0.9	51.12	64.19	65.07	75.50
0.95	79.78	89.30	74.33	81.92
0.99	89.43	95.54	76.97	84.05
0.995	**90.80**	**95.90**	**77.78**	**85.02**
0.999	86.27	93.69	38.89	39.37

Table 10: Effect of λ_{new}

λ_{new}	*RoamingOmniglot* AMI	*RoamingOmniglot* AP	*RoamingRooms* AMI	*RoamingRooms* AP
0.0	38.26	93.40	19.49	73.93
0.1	86.60	93.50	67.25	71.69
0.5	89.89	95.28	**78.04**	**84.85**
1.0	**90.06**	**95.81**	77.59	84.36
2.0	88.74	95.73	77.62	84.72

The adjusted MI score[1] normalizes the range between 0 and 1, and subtracts the baseline from random chance:

$$\text{AMI}(U, V) = \frac{MI(U, V) - \mathbb{E}[MI(U, V)]}{\frac{1}{2}(H(U) + H(V)) - \mathbb{E}[MI(U, V)]}, \tag{57}$$

where $H(\cdot)$ denotes the entropy function, and $\mathbb{E}[MI(U, V)]$ is the expected mutual information by chance [2]. Finally, we sweep the threshold α to get a threshold invariant score:

$$\text{AMI}_{\text{max}} = \max_{\alpha} \text{AMI}(y, \hat{y}(\alpha)). \tag{58}$$

AP: For supervised evaluation, we used the AP metric. The AP metric is also threshold invariant, and it takes both output $\hat{u}$ and $\hat{y}$ into account. First it sorts all the prediction based on its unknown score $\hat{u}$ in ascending order. Then it checks whether $\hat{y}$ makes the correct prediction. For the N top ranked instances in the sorted list, it computes: precision@N and recall@N among the known instances.

- precision@N = $\frac{1}{N}\sum_n \mathbb{1}[\hat{y}_n = y_n]$,
- recall@N = $\frac{1}{K}\sum_n \mathbb{1}[\hat{y}_n = y_n]$,

where K is the true number of known instances among the top N instances. Finally, AP is computed as the area under the curve of (y=precision@N, x=recall@N). For more details, see Appendix A.3 of Ren et al. (2021).

C Additional Experimental Results

C.1 Comparison to Reconstruction-Based Methods

We additionally provide Tab. 7 to show a comparison with CURL in the iid setting. We used the same MLP architecture and applied it on the Omniglot dataset using the same data split. Reconstruction-based methods lag far behind self-supervised learning methods. Our method is on par with SimCLR and SwAV.

C.2 Additional studies on hyperparameters

We study the effect of certain hyperparameters of our model in Tab. 8-14, using the validation set of *RoamingRooms* and *RoamingOmniglot*. Shown in Tab. 8, the memory size K does not affect the final performance, but interestingly the decay factor ρ is more important. In Tab. 9, the best performance is achieved with the memory decay rate ρ between 0.99 and 0.995. As shown in Tab. 10, the new cluster loss is necessary since the model performance drops significantly when the value of λ_{new} is below 0.5.

Tab. 11, 12, 13, and 14 show additional studies on the effect of the cluster creation threshold α, the temperature parameter $\tilde{\tau}$, the entropy loss parameter λ_{ent}, and the Beta prior mean. The Beta mean μ is computed as the following: Suppose a and b are the parameters of the Beta distribution, and μ is the mean. We fix $a = 4\mu$ and $b = 4 - a$.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted_mutual_info_score.html

[2] https://en.wikipedia.org/wiki/Adjusted_mutual_information

Table 11: Effect of threshold α

α	*RoamingOmniglot* AMI	*RoamingOmniglot* AP	*RoamingRooms* AMI	*RoamingRooms* AP
0.3	82.75	90.57	52.60	58.71
0.4	81.59	90.94	59.69	66.11
0.5	**89.65**	**95.22**	**77.96**	**84.34**
0.6	87.01	93.87	64.65	69.49
0.7	86.08	92.94	66.60	73.54

Table 12: Effect of $\tilde{\tau}$

$\tilde{\tau}$ / τ	*RoamingOmniglot* AMI	*RoamingOmniglot* AP	*RoamingRooms* AMI	*RoamingRooms* AP
0.05	89.23	95.01	77.44	84.38
0.10	89.71	95.21	**77.89**	**84.99**
0.20	**89.78**	**95.31**	77.82	84.57
0.50	89.40	95.13	76.81	83.90
1.00	89.62	95.16	0.00	19.91

Table 13: Effect of λ_{ent}

λ_{ent}	*RoamingOmniglot* AMI	*RoamingOmniglot* AP	*RoamingRooms* AMI	*RoamingRooms* AP
0.00	82.45	90.66	**76.64**	**84.11**
0.25	87.31	93.85	76.61	83.16
0.50	87.98	94.21	75.46	81.78
0.75	88.77	94.74	74.76	79.91
1.00	**89.70**	**95.14**	75.32	80.29

Table 14: Effect of mean μ of the Beta prior

μ	*RoamingOmniglot* AMI	*RoamingOmniglot* AP	*RoamingRooms* AMI	*RoamingRooms* AP
0.3	84.14	93.19	68.75	72.58
0.4	86.59	93.10	69.19	73.86
0.5	**89.89**	**95.24**	**77.61**	**84.64**
0.6	85.93	93.81	64.21	73.23
0.7	26.22	92.08	48.58	64.28

D Additional Visualization Results

We visualize the clustering mechanism and the learned image embeddings on *RoamingRooms* in Fig. 8 and 9. The results suggest that our model can handle a certain level of view point changes by grouping different view points of the same object into a single cluster. It also shows that our model is instance-sensitive: for example, the headboard, pillows, and the blanket are successfully separated.

In Fig. 10 and 11, we visualize the learned categories in *RoamingOmniglot* using t-SNE (Van der Maaten & Hinton, 2008). Different colors represents different ground-truth classes. Our method is able to learn meaningful embeddings and roughly group items of similar semantic meanings together.

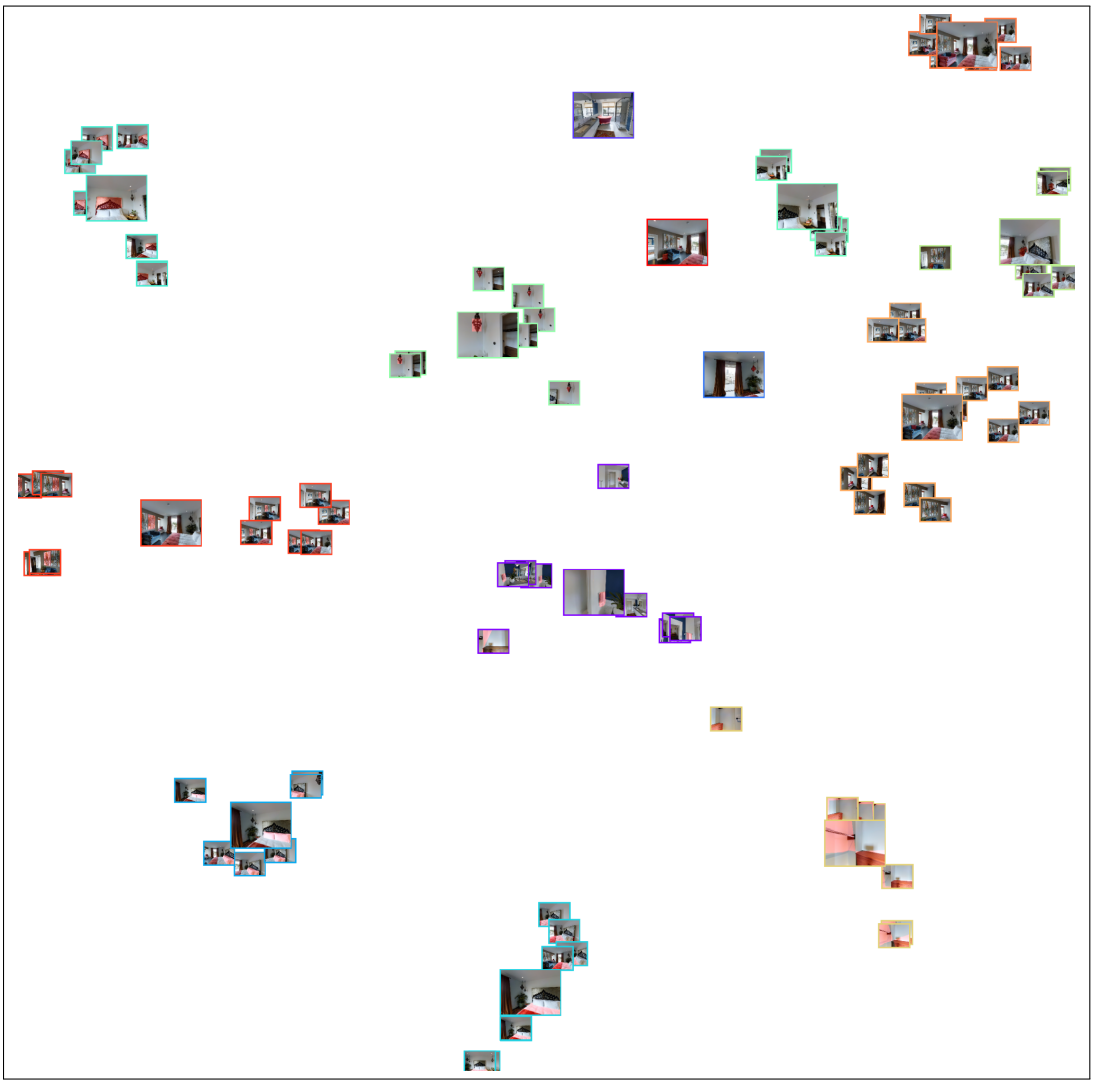

Figure 8: Embeddings and clustering outputs of an example episode (1). Embeddings are extracted from the trained CNN and projected to 2D space using t-SNE (Van der Maaten & Hinton, 2008). The main object in each image is highlighted in a red mask. The nearest example to each cluster centroid is enlarged. Image border colors indicate the cluster assignment.

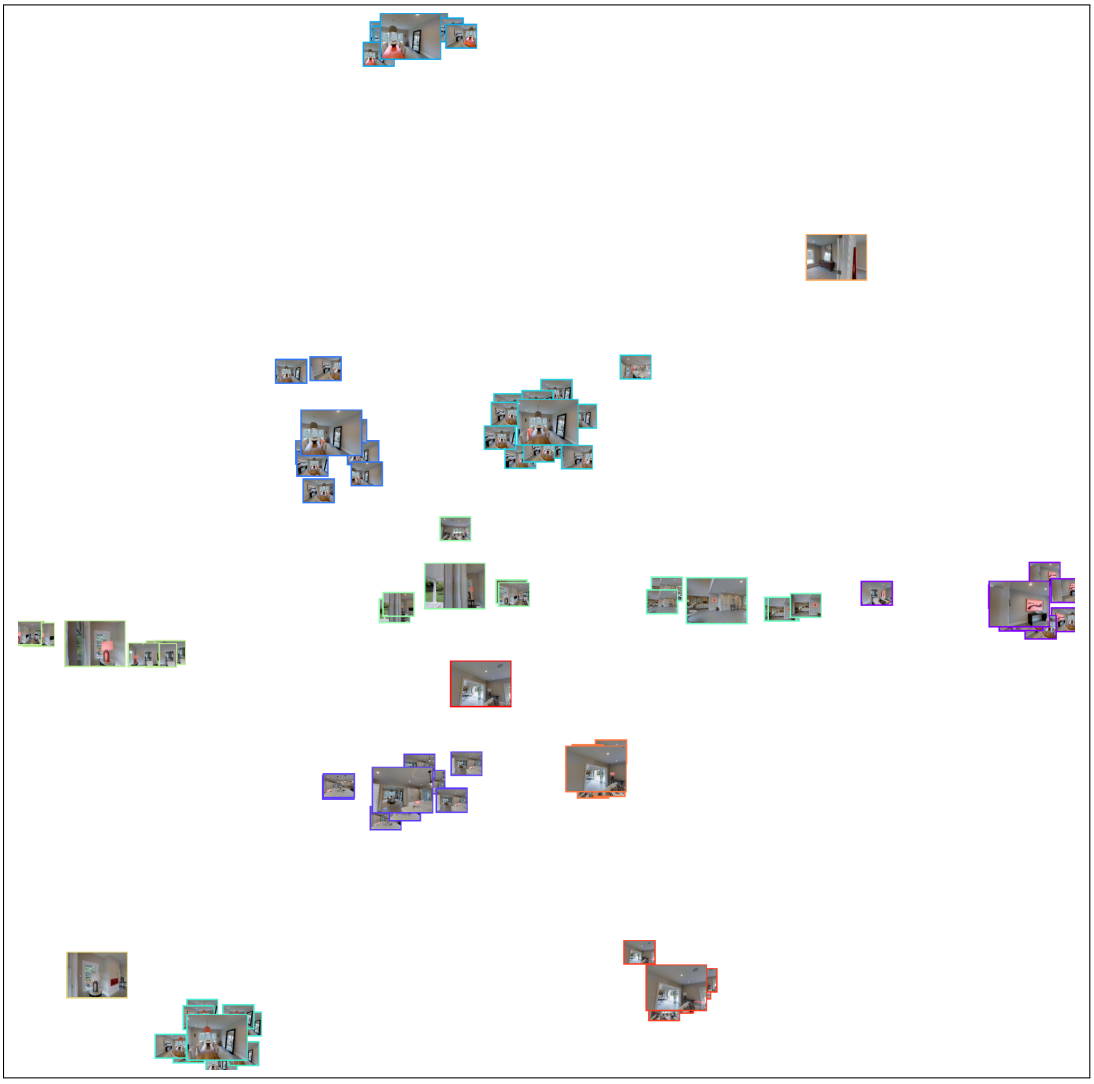

Figure 9: Embeddings and clustering outputs of another example episode (2).

Figure 10: Embedding visualization of an unsupervised training episode of *RoamingOmniglot*. Different colors denote the ground-truth class IDs.

Figure 11: Embedding visualization of an test episode of *RoamingOmniglot*.