CONTRASTIVE LEARNING THROUGH TIME

Anonymous authors
Paper under double-blind review

ABSTRACT

Human infants learn to recognize objects largely without supervision. In machine learning, contrastive learning has emerged as a powerful form of unsupervised representation learning. The utility of learned representations for downstream tasks depends strongly on the chosen augmentation operations. Taking inspiration from biology, we here study a framework for unsupervised learning of object representations we call Contrastive Learning Through Time (CLTT). CLTT simulates viewing sequences as they might be experienced by an infant while interacting with objects and avoids arbitrary augmentation operations. Instead, positive pairs are formed by successive views in such unsegmented viewing sequences. Generating viewing sequences procedurally, rather than using natural videos, gives us perfect control over the temporal structure of the input and allows us to ask the following two questions. First, can CLTT approach the performance of fully supervised learning? Second, if so, what are the required conditions on the temporal structure of the input? To answer these questions, we develop a new data set using a near-photorealistic training environment based on ThreeDWorld (TDW). We consider several state-of-the-art contrastive learning methods and demonstrate that CLTT allows linear classification performance that approaches that of the fully supervised setting if subsequent views are sufficiently likely to stem from the same object. We also consider the effect of one object being seen systematically before or after another object. We show that this leads to increased representational similarity between these objects, reminiscent of classic neurobiological findings. The data sets, code and pre-trained models for this paper can be downloaded at: (link will be added in the final version).

1 INTRODUCTION

A hallmark of biological organisms is their ability to learn to understand the world around them in a largely autonomous fashion. Consider learning about visual objects. A human infant is not exposed to object views sampled i.i.d. from some fixed distribution and conveniently labeled image by image, but forms representations of objects and categories during extended interactions with individual objects (Bambach et al., 2018) and requires hardly any (verbal) labels for this (LaTourrette & Waxman, 2019). Mimicking such learning abilities in artificial systems would represent a giant leap forward for artificial intelligence.

Self-supervised learning has emerged as a promising alternative to fully supervised approaches. In the domain of visual object recognition, recent contrastive learning approaches have obtained strong results on standard object recognition benchmarks (Chen et al., 2020a;b; Mitrovic et al., 2021; Grill et al., 2020). These approaches rely on a range of so-called augmentation operations. The basic idea is that an image is transformed through a number of operations (e.g., scaling, flipping, cropping, rotating, blurring, color distortions, pixel noise, ...) that change its appearance but not its meaning (e.g., “cat”). The key mechanism of contrastive learning is to form a representation where such augmented versions of an image are mapped on to close-by latent representations, while at the same time avoiding a “collapse” of the representation, i.e., making sure the network does not simply map all inputs to the same point in the latent space. Not surprisingly, the quality of the learned representations for downstream tasks strongly depends on the chosen augmentation operations (Grill et al., 2020).

Theories of biological learning have also addressed the requirement to learn object representations without (or with only few) labels. The classic theory of how biological organisms learn invariant

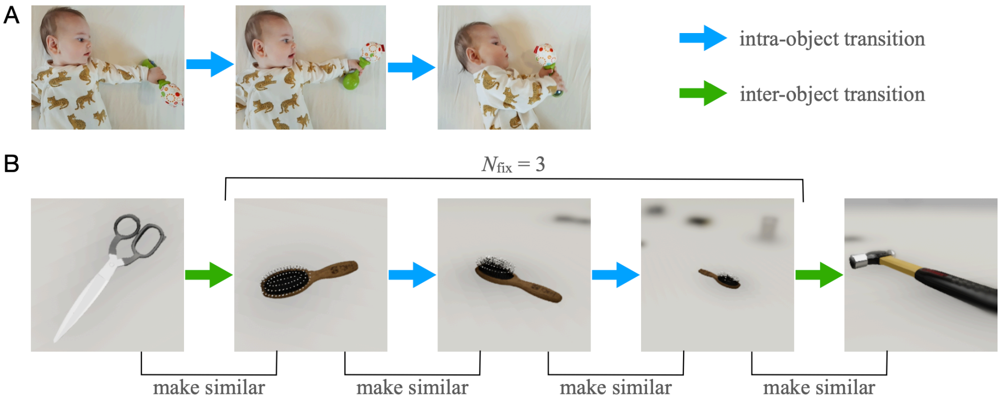

Figure 1: Contrastive Learning Through Time (CLTT). **A.** Infants learn about objects during extended interactions with these objects. Typically, they experience different views of an object before a different object comes into view. **B.** Our CLTT approach mimics the essence of such interactions. A certain number N_{fix} of object views are sampled before directing attention to another object. Latent representations of successive views are made more similar. N_{fix} determines the relative abundance of intra-object transitions vs. inter-object transitions. Importantly, CLTT does not require knowledge of which transitions are inter-object transitions. Both types of transitions are treated identically during learning, making the approach fully unsupervised.

representations uses the notion of time to substitute for explicit labeling (Földiák, 1991; Rolls & Milward, 2000; Wiskott & Sejnowski, 2002). Biological organisms (including the human infant mentioned above) experience objects across time, typically seeing a sequence of different views of the same object before directing their attention elsewhere (Fig. 1A). Thus, by learning a representation such that subsequent views are mapped onto close-by latent codes (Fig. 1B), a representation should emerge that maps different views of the same object onto similar latent codes, thereby establishing (partial) *invariance*. While this idea has a long history in biological theorizing, it has only recently been explored in a contrastive learning context. The basic idea is to replace the augmentation operations in contrastive learning with natural appearance variation occurring during object interactions. To systematically study this approach, we propose a new Contrastive Learning Through Time (CLTT) framework that permits perfect control over the generated viewing sequences. For our experiments, we utilize the ThreeDWorld (TDW) virtual environment (Gan et al., 2021), which allows near-photorealistic rendering. We also simulate classic biological experiments by Miyashita (1988), demonstrating that objects form similar latent representations in the brain when they are systematically seen one after the other, even if they are visually dissimilar (Miyashita, 1988). We summarize our contributions as follows:

- We develop the CLTT framework using state-of-the-art contrastive learning methods.
- We introduce novel data sets to study CLTT under controlled conditions.
- We systematically analyze the conditions for CLTT to be successful and demonstrate that it approaches fully supervised learning.
- We show that CLTT maps objects that are systematically seen in temporal succession on similar latent representations, reminiscent of classic neurobiological findings.

2 Related Work

An early demonstration that the temporal structure of visual inputs shapes object representations in primate visual cortex was given by Miyashita (1988). He showed 97 images of fractal-like objects to monkeys in always the same order. As the monkeys learned to represent these images in their visual cortices, Miyashita's observations suggested that the representations of objects which were neighbors in the sequence became aligned — even if these objects were visually dissimilar. This effect extended over a few objects, i.e., the representations of objects six steps apart in the sequence were still more similar than the average similarity.

In part motivated by such findings, there is a long history of neural network and machine learning models exploiting temporal structure for unsupervised representation learning. Földiák (1991) introduced so-called trace learning rules to explain how neurons in the mammalian visual system learn invariance properties, setting a starting point for later models considering multi-layered network architectures (Rolls & Milward, 2000). Another line of research introduced by Wiskott & Sejnowski (2002) explicitly considers the objective of extracting components from an input stream that are slowly changing. A recent variant attempts to do so in a biologically plausible fashion (Lipshutz et al., 2020).

The use of temporal learning objectives in contrastive learning has received increasing attention recently. Among the first, Mobahi et al. (2009) have proposed a method for learning object representations that combines supervised learning and unsupervised learning based on temporal coherence using a siamese neural network architecture. Subsequently, Wang & Gupta (2015) used tracking of patches in videos for unsupervised pre-training, by learning an embedding that keeps patches from the same track close in the embedding space.

An approach more closely related to ours has recently been proposed by Orhan et al. (2020). They consider learning on a longitudinal headcam video set from three developing children (Sullivan et al., 2020). They focus on a *temporal classification* approach, where they divide the videos into a finite number of contiguous segments of the same length that they call *temporal classes*. The learning objective is to predict from which of the classes a particular video frame originates. They also consider a temporal contrastive learning objective with the MoCo contrastive learning implementation of Chen et al. (2020b). This objective also aims to make the latent representations of adjacent video frames similar. However, the use of uncontrolled headcam video does not permit determining the required the temporal statistics of the visual input for the approach to work. Another related approach is that of Knights et al. (2021), who learn embeddings of video clips. Their learning objective makes latent codes of adjacent frames within a video clip similar, while making them distinct from latent codes of frames from other video clips. Note that this setup requires the video clips to be segmented, i.e., the system has access to the information where each video starts and ends rather than being exposed to an unlabeled continuous video stream as in Orhan et al. (2020) and CLTT. The same holds true for the recent approach of Feichtenhofer et al. (2021) and also Pan et al. (2021). Yet, they don't make the connection between continuous time frames and biologically-inspired vision. Finally, the interesting work of Stojanov et al. (2019) has modeled continual infant-like learning of object representations from unsegmented input streams, but they did not consider a contrastive learning approach.

3 Methods

3.1 Sampling Sequences of Object Views in CLTT

CLTT aims to mimic the stream of object views that an infant may experience during natural interactions with objects, while giving precise control over the statistical properties of this view sequence. Specifically, an object is always viewed for N_{fix} fixations before another object comes into view. Importantly, the learner does not have access to the information when a new object comes into view, making the approach fully unsupervised. Generating these view sequences involves two sampling procedures. The first describes how the next view of the *same* object is sampled during the N_{fix} fixations on the same object. We refer to this as *view sampling*. The second describes how the identity of the next object is determined at the end of the N_{fix} fixations on the same object, which we refer to as *object sampling*.

For the view sampling, we distinguish two sampling methods. Our default method is the *random walk* view sampling. Here, the next view of an object is a "neighbor" of the previous view. For example, in the TDW data set (see below) this corresponds to changing azimuth or elevation by 10° or viewing distance by 10 cm. This procedure mimics the infant gradually turning an object or moving around an object while fixating it. The second method is the *uniform* view sampling. Here, the next view is picked uniformly at random. While this does not mimic infants' viewing sequences, it provides more diversity among successive views, which is expected to aid learning.

For the object sampling we also distinguish two methods. Our default method is the *random order* method. Here, the order of the objects is a new random permutation during each training cycle. This corresponds to the case that the infant encounters objects randomly. The second method uses a *fixed order* of objects, i.e., object A is always followed by object B, etc. in all training cycles. This situation matches the neurobiological experiments by Miyashita (1988) and is expected to lead to an alignment of the latent codes of objects that are consistently seen in succession.

Training occurs in cycles. In every cycle each of the N_{obj} objects will be chosen exactly once — the order depending on the chosen object sampling method. Thus, a cycle consists of a total of $N_{\text{obj}} \times N_{\text{fix}}$ object views. The N_{fix} views for each object are determined by the chosen view sampling procedure. Several cycles of $N_{\text{obj}} \times N_{\text{fix}}$ views are stored in a buffer. For our experiments we use a batch size of 256. During learning, batches of pairs of subsequent views are sampled uniformly from the buffer. This leads to a probability $1/N_{\text{fix}}$ of sampling *inter-class* transitions and a probability of $(N_{\text{fix}} - 1)/N_{\text{fix}}$ of sampling *intra-class* transitions (compare Fig. 1). We consider an epoch to be a full run through the buffer, which is chosen to be of equal size as the underlying data set. That means that on average, every image of the data set is presented once during each epoch.

3.1.1 Contrastive Learning Algorithms for CLTT

The sampled sequences can be fed into a wide range of contrastive learning methods. Here we consider SimCLR (Chen et al., 2020a) and RELIC (Mitrovic et al., 2021) in which for every positive pair, all the other pairs in the batch are considered negative pairs. We also experiment with BYOL (Grill et al., 2020), that uses only positive pairs. We refer to the CLTT versions of these algorithms as SimCLR-TT, RELIC-TT, and BYOL-TT, respectively. In all cases, we use a ResNet-18 architecture (He et al., 2016) to transform the input images into 128-dimensional latent representations, which is followed by a single layer as the projection head for RELIC-TT and SimCLR-TT. For BYOL-TT we use a two layer projection head with batch normalization after the hidden layer.

SimCLR-TT. SimCLR (Chen et al., 2020a) is an effective contrastive learning approach for visual representation learning. By defining a wide range of augmentation operations and treating differently augmented images as positive samples, SimCLR has achieved state-of-the-art performance. Its groundbreaking success has triggered a surge of interest in augmentation-based self-supervised learning. For SimCLR-TT we replace the traditional augmentations with successive views as pairs defined in the previous subsection. For each positive pair, all remaining pairs in the batch are considered negative samples. The loss of SimCLR-TT is then defined as follows:

$$\mathcal{L}\left(z_i\right) = -\log \frac{\exp\left(\operatorname{sim}\left(z_i, z_i'\right)/\tau\right)}{\sum_{k=1, k\neq i}^{N_B}\left[\exp\left(\operatorname{sim}\left(z_i, z_k\right)\right) + \exp\left(\operatorname{sim}\left(z_i, z_k'\right)\right)\right]/\tau}, \tag{1}$$

where z_i and z_i' are the latent codes of a sampled view pair, and N_B is the number of pairs in a batch. Specifically, we use the cosine similarity as the similarity function $\operatorname{sim}(u, v) = \frac{u^\top v}{\|u\|\|v\|}$. For simplicity, we set the temperature parameter τ in the original SimCLR loss to be 1. In all our experiments, we use the AdamW optimizer (Loshchilov & Hutter, 2019) with a starting learning rate of 10^{-3}, which decays by a factor of 0.3 after every 10 epochs. The batch size for training is 256 and the buffer size is determined by the size of the data set.

RELIC-TT. RELIC (Mitrovic et al., 2021) is an approach that uses an additional penalty loss compared to SimCLR. It obtains state-of-the-art results by keeping the similarity distribution of one sample invariant against differently augmented views of other samples. Incorporating this notion into the CLTT approach, we derive RELIC-TT by adding another loss term that. The loss can be described as follows:

$$\mathcal{L}_p\left(z_i\right) = KL\left(p\left(Y \mid z_i\right), p\left(Y \mid z_i'\right)\right), \tag{2}$$

where KL is the Kullback-Leibler divergence and

$$p\left(Y = j \mid z_i\right) = \frac{\exp\left(\operatorname{sim}\left(z_i, z_j'\right)\right)}{\sum_{k=1}^{N_B} \exp\left(\operatorname{sim}\left(z_i, z'{}_k\right)\right)}. \tag{3}$$

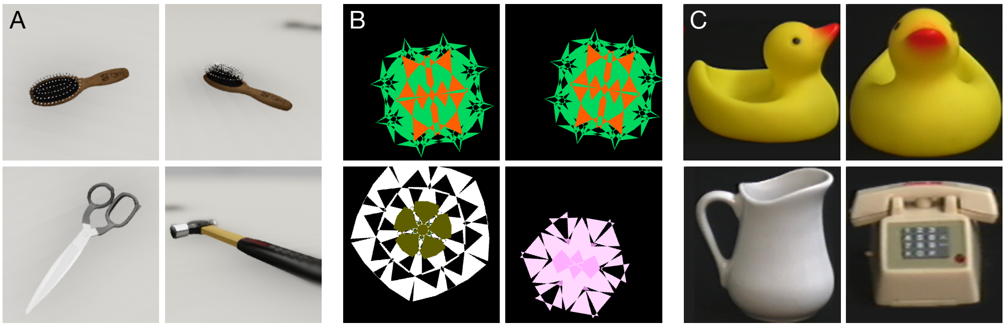

Figure 2: Data sets developed/used in our study. **A.** TDW data set containing common household objects viewed from different orientations and distances. The upper two images show the same example object ("hair brush") from two perspectives. The lower two images show two other example objects **B.** Fractal-like objects inspired by Miyashita (1988). **C.** Example objects from COIL-100.

In the definition above, z_i and z'_i are the latent codes of the first view and the second view of the sampled view pair. The similarities between z_i and all the second views in every pair are calculated and passed to a Softmax function as shown in equation 3. The same computation is done for the second view z'_i for all the other first views, resulting in totally two probability distributions denoted as $p\left(Y \mid z_i\right)$ and $p\left(Y \mid z'_i\right)$ that resembles a classification task of predicting the label Y out of totally N_B classes. And then those two distributions are pulled together by minimizing the KL divergence between them. The training configuration is the same as used in SimCLR-TT.

BYOL-TT. BYOL-TT builds on the Bootstrap Your Own Latent (BYOL) architecture by (Grill et al., 2020). BYOL has been shown to outperform other contrastive learning architectures like SimCLR (Chen et al., 2020a) or MoCov2 (Chen et al., 2020b) on the ImageNet data set (Deng et al., 2009). A key advantage of the BYOL architecture is that it works without negative pairs, which sets it apart from other contrastive learning algorithms. Instead, it uses a second so-called target network. The target network receives an augmented version of the input data like the online network. It will produce a target projection and the online network tries to predict this target projection. The loss function minimizes the similarity between the target projection and the prediction of the online network. In our BYOL-TT implementation, we use the AdamW optimizer with a learning rate of 2×10^{-4} which decays by a factor of 0.3 after every 30 epochs and a batch size of 256. The loss is given by:

$$\mathcal{L}_{\theta,\xi} = 2 - 2 \cdot \frac{\langle q_\theta(z_\theta), z'_\xi \rangle}{||q_\theta(z_\theta)||_2 \cdot ||z'_\xi||_2}, \tag{4}$$

where $q_\theta(z_\theta)$ is the prediction of the online network of z'_ξ. θ corresponds to the weights of the online network and ξ represent the weights of the target network. The optimization will be performed with respect to θ, the weights ξ of the target network will be updated using an exponential moving average of the online network. Here we choose $\tau = 0.99$ as target decay rate. In the original BYOL architecture, an input image x produces two augmented views v and v' which are then shown to the target and online network. In our case no augmentations are applied to the input image, positive pairs are formed by images next to each other in the presented data timeline. Two consecutive images in this timeline will then be used as views v and v'.

3.2 Data Sets

ThreeDWorld data set. This data set was created with the ThreeDWorld (TDW) software (Gan et al., 2021). TDW was chosen, because it enables near-photorealistic rendering including depth-of-field effects and the possibility of simulating physical interaction with objects (which we plan to utilize in the future). Our data set comprises twelve high quality 3D models of common household objects such as a hair brush, a hammer or scissors (compare Fig. 2A). They are arranged in a 4×3 grid on a white floor. To simulate viewing sequences that an infant may experience while interacting with an object such as turning an object in hand or moving relative to an object while fixating it, we render object views from different directions and distances. The individual images of the data set can then be arranged into videos simulating different viewing sequences, where an object is seen from different directions and distances. Specifically, to create the different views, we define a spherical coordinate system around the center of each object. Different views are created by changing viewing direction in terms of azimuth (from 0° to 350° in 10° steps), elevation (from 10° to 70° in 10° steps), and distance (0.3 cm to 0.6 cm in 10 cm steps). This gives rise to 1,008 views per object and a total of 12,096 images, which are down-sampled to the size of 64 by 64 pixels. We also created a test set with distinct viewing directions by shifting the azimuth and elevation by 5° (from 15° to 75° and from 5° to 355° in 10° steps). This procedure resulted in a second data set with 12,096 images. Note that due to the varying distances from the objects, sometimes not all parts of an object are visible inside the image and sometimes (parts of) other objects may be present. These appear blurry due to the simulated depth-of-field effect. By sampling different paths through this object, azimuth, elevation and distance space, we generate different viewing sequences that form the input to the CLTT algorithms presented above.

Miyashita-style data set. The Miyashita-style data set draws inspiration from the neuroscience experiments from Miyashita (1988). We adhered to the generation procedure described in (Miyashita et al., 1991) and wrote python code to generate 100 different fractals of size 64×64. Each of those fractals is unique and highly distinguishable from all others. For our experiments, we give each fractal its own class label and allow for a certain degree of variability to simulate fixation inaccuracies during viewing of the images. Each fixation applies a transformation with a random rotation between -10 and $+10$ degrees, a random rescaling between 90 and 100 $\%$ and a random translation of up to 15 $\%$ in x- and y-direction (compare Fig. 2B). The fractals are used exclusively with the *fixed order* procedure, i.e., there is a predefined order of fractals that is repeatedly shown to the network.

COIL-100 data set. To validate the CLTT approach on real images, we employ the Columbia Object Image Library (COIL-100) database (Nene et al., 1996). It is composed of color images of 100 objects, each viewed from 72 different directions by placing the objects on a motorized turntable making a total of 7,200 images. The objects are seen against a homogeneous black background (compare Fig. 2C).

4 Results

4.1 CLTT approaches representation quality of supervised learning

Results with ThreeDWorld data set. We evaluate the proposed family of CLTT methods, namely SimCLR-TT, RELIC-TT, BYOL-TT on the novel ThreeDWorld data set (Fig. 3). We also add a supervised method as baseline, which has access to the true label of each image. We focus on the *random walk* sampling of views and the *random sequence* procedure for sampling objects. We train the network for 100 epochs and vary N_{fix} to study its effect. In order to evaluate the quality of learned representations, we use a Linear Least Squares (LLS) classifier to test linear separability. In general, a good representation, like that resulting from supervised learning, should be linearly separable and have high LLS classification accuracy. In Fig. 3A we show the LLS classification accuracy for $N_{\text{fix}} = 5$ as a function of training epoch. Both SimCLR-TT and RELIC-TT achieve the same or even slightly better performance than the supervised approach. Figure 3B shows the final LLS classification accuracy after 100 training epochs for different values of N_{fix}. Noticeably, SimCLR-TT and RELIC-TT perform on par with the supervised approach with $N_{\text{fix}} \geq 5$, while BYOL-TT achieves the best score at $N_{\text{fix}} = 5$ and larger N_{fix} seems to have no positive effect on BYOL-TT. Figure 3C depicts a PacMAP (Wang et al., 2021) visualization of the latent space resulting from training with RELIC-TT for 100 epochs using $N_{\text{fix}} = 30$. The objects form well separated clusters.

Results with COIL-100 data set. To evaluate CLTT on real (rather than computer rendered) images and a data set with a larger number of objects, we use the classic COIL-100 data set. Note that the COIL-100 data set does not have separate validation and test sets, thus we evaluated performance on the training set in the following experiments. This approach may have led to some over-fitting in the supervised setting, but this effect would only make it harder for our unsupervised method to

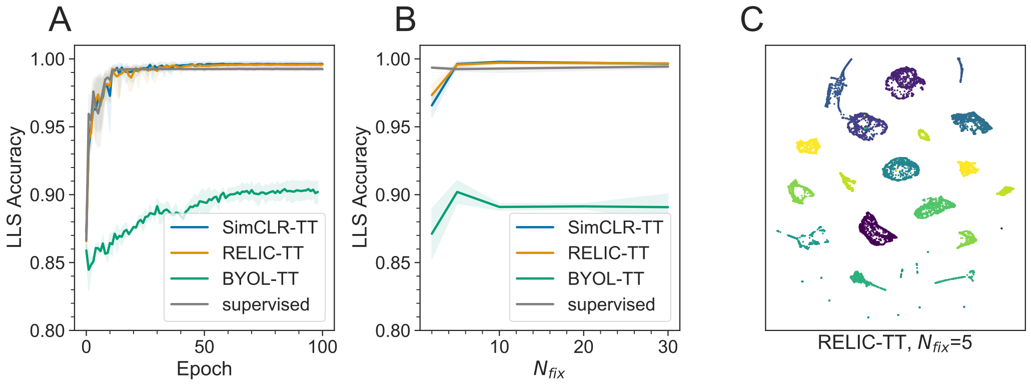

Figure 3: Results for CLTT on the TDW data set. **A.** Comparison of Linear Least Squares (LLS) classification accuracy as a function of training epoch for the different algorithms and with $N_{\text{fix}} = 5$. **B.** Final LLS classification accuracy as a function of N_{fix}. **C.** Visualization of the clustering of representations in the latent space using PacMAP, here shown for RELIC-TT with $N_{\text{fix}} = 5$. Each color corresponds to one distinct object. Shaded areas in panels A and B represent the standard deviation based on three individual runs.

match the performance of the supervised counterpart. We compare performance of SimCLR-TT, RELIC-TT, and supervised methods for different values of N_{fix}. We also compare the two sampling strategies: *random walk sampling* where successive views are generated via a random walk in the space of viewing directions and *uniform sampling* where successive views are picked uniformly at random across all possible viewing directions. Figure 4A shows the LLS accuracy of the different algorithms and sampling strategies as a function of the number of training epochs for $N_{\text{fix}} = 30$. The SimCLR-TT method outperforms RELIC-TT. Furthermore, we observe that the *uniform sampling* leads to better results than the *random walk* sampling. This is not surprising, as it creates more diversity in the training data by allowing for very different views of an object to be grouped as a positive pair during contrastive learning. Also, although learning is somewhat slower, the performance of SimCLR-TT and RELIC-TT with uniform sampling approaches that of fully supervised learning. Figure 4B compares the final LLS classification accuracy of the algorithms after 50 training epochs for different values of N_{fix}. The performance improves monotonically with N_{fix} as expected and approaches that of the fully supervised setting. Figure 4C visualizes the clustering of latent representations with PacMAP for RELIC-TT with $N_{\text{fix}} = 30$ and the *random walk sampling* procedure. The 100 different objects have formed distinct clusters in the latent space. Finally, in Appendix A.1, we compare *uniform sampling* and *random walk sampling* for the three methods with $N_{\text{fix}} = 2$ and $N_{\text{fix}} = 10$. As expected, the results show that *uniform sampling* achieves better performance.

4.2 CLTT ALIGNS LATENT CODES OF SUCCESSIVELY VIEWED OBJECTS

Results with Miyashita-style data set. To relate our framework to biological findings we use the Miyashita-style data set combined with our SimCLR-TT approach. We train the network for 10^6 stimuli presentations (100 epochs, buffer size = 10,000) and vary N_{fix}. Here, the buffer-size is larger than the number of fractals, because the data set is dynamically generated using the aforementioned fixations. In line with results from neuroscience, fractals that were presented in succession evoke more similar activations in the latent space than fractals that are far apart in the predefined sequence. Fig. 5A depicts the mean cosine similarity between a fractal's latent representation and that of its two n-th nearest neighbors along the sequence (in the positive and negative time direction). Note that neighbor zero has a similarity of one, as it is de-facto the same activation pattern. The curves for all values of N_{fix} display the property of a downward slope that is clearly separated from the baseline (dashed lines). This indicates that over time not only immediate neighbors, but also more distant fractals become associated. In fact for $N_{\text{fix}} = 2$ we see significant deviations ($p < .01$, two-sided Kolmogorov-Smirnov two-sample test) for the nine next neighbors before merging with the baseline. The number of deviating neighbors shrinks with increasing N_{fix}, but significant effects still can be observed. This experiment illustrates a fundamental property of CLTT that can be directly

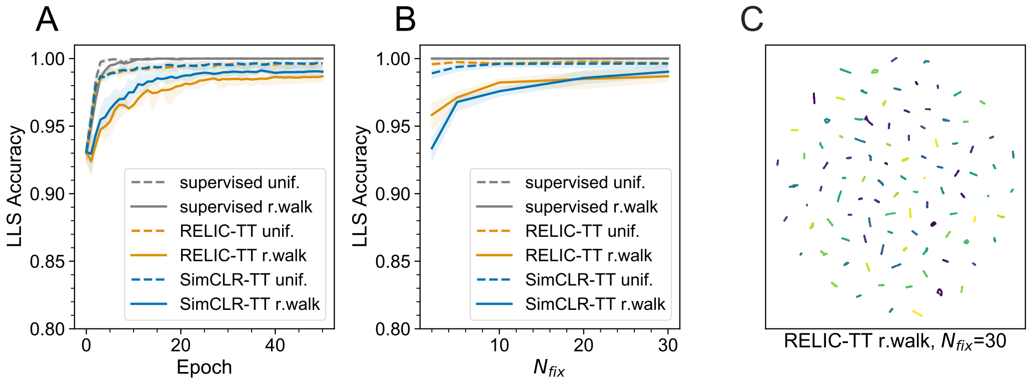

Figure 4: Results for CLTT on the COIL-100 data set. **A.** Comparison of LLS classification accuracy for the different algorithms and sampling strategies with $N_{\text{fix}} = 30$. **B.** Comparison the influence of N_{fix} on LLS classification accuracy under the two sampling strategies. **C.** Visualization of the clustering of representations in the latent space using PacMAP, here shown for RELIC-TT with $N_{\text{fix}} = 30$ and the *random walk sampling* procedure. Different colors correspond to different objects, but due to the large number of objects, colors have been reused multiple times. Shaded areas in panels A and B represent the standard deviation based on five individual runs.

related to biological findings. We also compare our results to supervised learning, which can solve the classification of 100 fractals, but the resulting latent representations do not display any structure of temporal associations and thus do not differ significantly from its baseline (not shown).

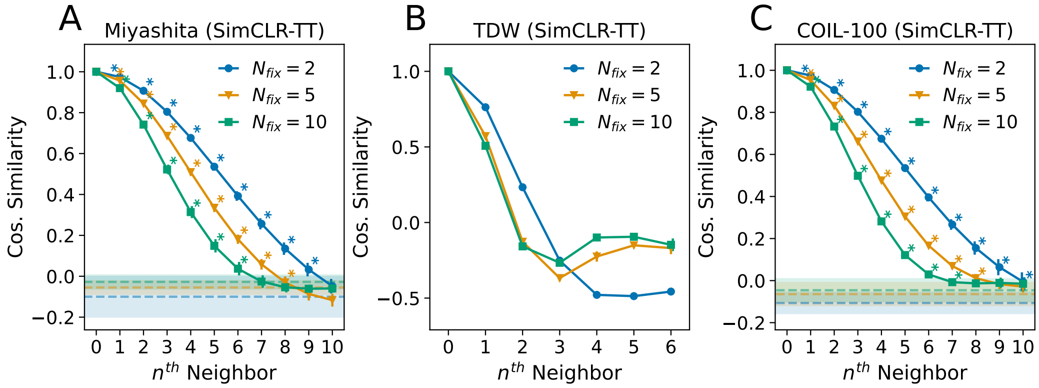

Figure 5: Latent codes of successively viewed objects. Cosine similarity of neighboring objects along the timeline as seen in the latent space for different values of N_{fix}. Colored dashed lines and envelopes represent the mean and standard deviation of the neighbors not shown, i.e., $n \geq 10$. Error-bars depict the standard deviation based on three independent runs. **A.** Miyashita fractals. **B.** TDW objects. **C.** COIL-100 objects.

Results with ThreeDWorld data set. We tested if the alignment of latent representations of objects that are consistently shown in succession could also be observed with the TDW data set. For this we sampled objects using the *fixed order* procedure, i.e., the objects were always seen in the same sequence order. We focused on SimCLR-TT for this experiment. Figure 5B replicates the effect from the Fractals data set (compare Fig. 5A). Figure 5C shows the equivalent experiment for the COIL-100 data set. The fixed order aids classification and SimCLR-TT achieves excellent classification performance for different values of N_{fix}. Thus, the network is able to learn good representations of the input data even with low values of N_{fix} when the order in which the objects are shown to the network is fixed.

5 Discussion

We have developed a general framework for contrastive learning through time (CLTT). CLTT emulates viewing sequences as may be experienced by infants and uses a temporal contrastive loss that maps subsequent inputs occurring during such interactions onto close-by latent representations ("close in time, will align"). To systematically investigate this approach, we have created a new data set using the ThreeDWorld environment (Gan et al., 2021), allowing us to flexibly simulate different kinds of viewing scenarios in a near photo-realistic fashion. We also validated our approach using a new fractal-like data set inspired by biological experiments and the classic COIL-100 computer vision data set. We have demonstrated that CLTT can approach the quality of representations learned using full supervision. For this it is important that intra-object transitions (successive fixations fall on the same object) dominate over inter-object transitions (successive fixations fall on different objects). We have also shown that CLTT produces effects reminiscent of classic biological findings showing that inputs that occur close in time are mapped onto close-by latent representations by the brain.

Our work has a number of limitations. First, while the use of TDW gives us perfect control over all parameters of viewing sequences including scene geometry, sequence of views, lighting conditions, etc., the approach needs to be validated in the real world. Using the COIL-100 data set has been a first step in this direction. To address this more thoroughly, future work could consider first person videos from infants wearing head-mounted cameras (Bambach et al., 2018; Orhan et al., 2020). Second, we have assumed well-separated objects without major occlusions in front of uniform backgrounds. Learning in a cluttered environment is expected to be harder, but it may benefit from foveated vision and additional attentional and figure-ground separation mechanisms, which we plan to incorporate in future work. Third, our eye movement model, which kept gaze on the same object for a certain number of fixations before redirecting it elsewhere, is overly simplistic. In the future, it will be interesting to use more refined models that better reflect (measured) gaze sequences of children and adults. Indeed, it is an interesting question what gaze sequences are particularly beneficial for learning and if and how infants and artificial vision systems can optimize viewing sequences to maximize their own learning progress. This links CLTT to research on intrinsic motivation and (artificial) curiosity. Exploring these issues will bring us closer to building artificial vision systems that can learn truly autonomously.

Acknowledgments

Removed for double blind review.

References

Sven Bambach, David Crandall, Linda Smith, and Chen Yu. Toddler-inspired visual object learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/48ab2f9b45957ab574cf005eb8a76760-Paper.pdf.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020a.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3299–3309, June 2021.

Peter Földiák. Learning invariance from transformation sequences. *Neural computation*, 3(2):194–200, 1991. doi: 10.1162/neco.1991.3.2.194.

Chuang Gan, Jeremy Schwartz, Seth Alter, Damian Mrowca, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, Megumi Sano, Kuno Kim, Elias Wang, Michael Lingelbach, Aidan Curtis, Kevin Tyler Feigelis, Daniel Bear, Dan Gutfreund, David Daniel Cox, Antonio Torralba, James J. DiCarlo, Joshua B. Tenenbaum, Josh Mcdermott, and Daniel LK Yamins. ThreeDWorld: A platform for interactive multi-modal physical simulation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL `https://openreview.net/forum?id=db1InWAwW2T`.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf`.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Joshua Knights, Ben Harwood, Daniel Ward, Anthony Vanderkop, Olivia Mackenzie-Ross, and Peyman Moghadam. Temporally coherent embeddings for self-supervised video representation learning. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 8914–8921. IEEE, 2021.

Alexander LaTourrette and Sandra R Waxman. A little labeling goes a long way: Semi-supervised learning in infancy. *Developmental science*, 22(1):e12736, 2019. doi: 10.1111/desc.12736.

David Lipshutz, Charles Windolf, Siavash Golkar, and Dmitri Chklovskii. A biologically plausible neural network for slow feature analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 14986–14996. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/ab73f542b6d60c4de151800b8abc0a6c-Paper.pdf`.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Bkg6RiCqY7`.

Jovana Mitrovic, Brian McWilliams, Jacob C Walker, Lars Holger Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=9p2ekP904Rs`.

Yasushi Miyashita. Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature*, 335(6193):817–820, 1988. doi: 10.1038/335817a0.

Yasushi Miyashita, Sei-Ichi Higuchi, Kuniyoshi Sakai, and Naohiko Masui. Generation of fractal patterns for probing the visual memory. *Neuroscience Research*, 12(1):307–311, 1991. doi: 10.1016/0168-0102(91)90121-E.

Hossein Mobahi, Ronan Collobert, and Jason Weston. Deep learning from temporal coherence in video. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 737–744, 2009.

Sameer A Nene, Shree K Nayar, Hiroshi Murase, et al. Columbia object image library (coil-100). Technical report, Columbia University, Feb 1996.

Emin Orhan, Vaibhav Gupta, and Brenden M Lake. Self-supervised learning through the eyes of a child. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9960–9971. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/7183145a2a3e0ce2b68cd3735186b1d5-Paper.pdf`.

Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11205–11214, June 2021.

Edmund T Rolls and T Milward. A model of invariant object recognition in the visual system: learning rules, activation functions, lateral inhibition, and information-based performance measures. *Neural computation*, 12(11):2547–2572, 2000. doi: 10.1162/089976600300014845.

Stefan Stojanov, Samarth Mishra, Ngoc Anh Thai, Nikhil Dhanda, Ahmad Humayun, Chen Yu, Linda B Smith, and James M Rehg. Incremental object learning from contiguous views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8777–8786, 2019.

Jessica Sullivan, Michelle Mei, Andrew Perfors, Erica Wojcik, and Michael C Frank. Saycam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open Mind*, pp. 1–10, 2020. doi: 10.1162/opmi_a_00039.

Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization. *Journal of Machine Learning Research*, 22(201):1–73, 2021. URL `http://jmlr.org/papers/v22/20-1061.html`.

Laurenz Wiskott and Terrence J Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14(4):715–770, 2002. doi: 10.1162/089976602317318938.

A APPENDIX

A.1 COIL-100: ADDITIONAL EXPERIMENTAL RESULTS

In this section, we supply additional results on the COIL-100 data set. Figure A.1. compares the two view sampling methods for the different algorithms for $N_{\text{fix}} = 2$ and $N_{\text{fix}} = 10$. It shows improved LSS accuracy for the uniform sampling strategy. This is not suprising, since it creates more diverse views than the random walk sampling strategy. It also shows that N_{fix} has a larger influence on the *random walk sampling*.

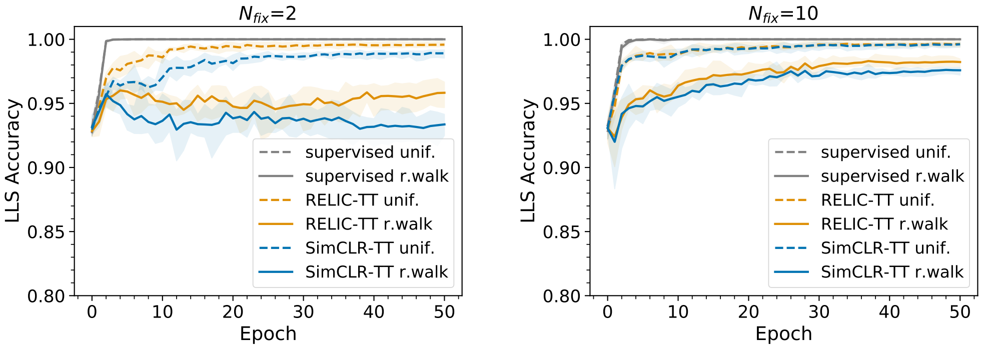

Figure A.1: LLS Classification Accuracy results of different N_{fix} for COIL-100 with *random walk sampling* and *uniform sampling*.