TEMPORAL SLOWNESS IN CENTRAL VISION DRIVES SEMANTIC OBJECT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Humans acquire semantic object representations from egocentric visual streams with minimal supervision. Importantly, the visual system processes with high resolution only the center of its field of view and learns similar representations for visual inputs occurring close in time. This emphasizes slowly changing information around gaze locations. This study investigates the role of central vision and slowness learning in the formation of semantic object representations in humans. We simulate five months of human-like visual experience using the Ego4D dataset and generate gaze coordinates with a state-of-the-art gaze prediction model. Using these predictions, we extract crops that mimic central vision and train a time-contrastive Self-Supervised Learning model on them. Our results show that combining temporal slowness and central vision improves the encoding of different semantic facets of object representations. Specifically, focusing on central vision strengthens the extraction of foreground object features, while considering temporal slowness, especially during fixational eye movements, allows the model to encode broader semantic information about objects. These findings provide new insights into the mechanisms by which humans may develop semantic object representations from natural visual experience. Our code will be made public upon acceptance.

1 INTRODUCTION

Humans develop strong semantic object representations from an unsupervised egocentric visual stream. These semantic representations reflect different non-perceptual facets of an object, such as its instance, fine-grained category, basic category or its context of occurrence. Models trained with self-supervised learning models (SSL) are reasonably good models of biological vision (Zhuang et al., 2021), but they are poor models of visual learning, as they rely on different training data and learning mechanisms than humans. As a consequence, they fail to model semantic human object similarity judgments (Mahner et al., 2025) and under-perform at recognizing objects when trained on a visual experience similar to humans' (Orhan, 2023; Orhan & Lake, 2024).

To learn semantic representations from their natural visual experience, humans may rely on two biological processes that are neglected in current models. First, the stimuli received by humans' visual cortex structurally differ from egocentric videos. The anatomy of the retina relatively amplifies the information located in the center of the field of view (Anstis, 1974; Wässle et al., 1989), meaning that high and intermediate acuity processing occurs only within several degrees from the center of the visual field, i.e. in central vision. As a consequence, central vision plays a crucial role in the formation of visual representations in areas of the visual cortex related to semantic information (Quaia & Krauzlis, 2024; Yu et al., 2015). To compensate for the relatively low acuity in peripheral vision, humans actively move their gaze onto different objects to parse their environment. Since human gazes are naturally attracted by salient objects, the visual sequence in central vision may most of the time present a few big objects. Second, a key principle of biological learning states that biological systems assign similar representations to close-in-time visual inputs. This may be important when learning from a natural experience in central vision. For instance, observing objects through different viewpoints may favor viewpoint-invariant object representations (Aubret et al., 2022a). In addition, consecutive scanning of objects within the same context may support object representations that encode their context of occurrences (Aubret et al., 2024a), a feature of human semantic object perception (Turini & Võ, 2022). For instance, the presence of a knife often reflects a "kitchen" context. However, humans produce frequent gaze movements that may disrupt the learning of these different semantic facets of objects, leaving unclear whether extracting slow information in central vision actually helps semantic object learning.

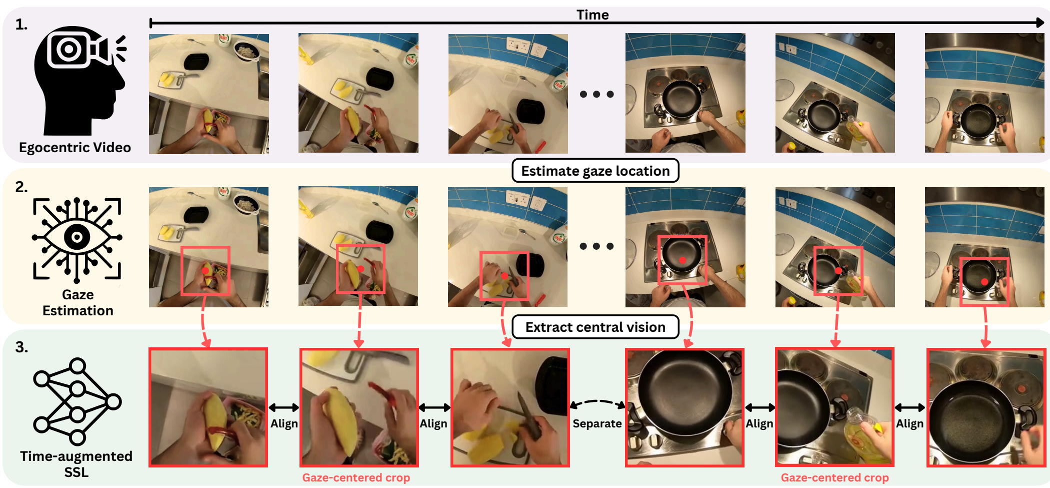

Figure 1: Illustration of our data generation and model training approach. (1) We extract frames from the egocentric dataset Ego4D (Grauman et al., 2022). (2) For each frame, we predict the human gaze location (red dot) using a state-of-the-art model (Lai et al., 2022). (3) We train a time-augmented SSL model to align representations of gaze-centered crops (red rectangle) extracted from close-in-time frames.

In this paper, we investigate the combined role of central vision and slowness learning for learning semantic object representations in humans. We simulate 5 months of human-like visual experience using Ego4D (Grauman et al., 2022), a dataset that contains 3,670 hours of videos collected with head-mounted cameras. This dataset contains gaze locations on only a subset of videos (45 hours). Thus, we apply a state-of-the-art model of human gaze prediction to generate gaze locations on the rest of the dataset (Lai et al., 2022). To simulate the importance of central vision in humans, we propose to crop a subpart of the image centered on the location of the gaze. Finally, we train a biologically inspired variant of an SSL model, which trains visual representations to slowly change, on the resulting visual sequence for one single epoch. Figure 1 summarizes our approach.

Our experiments demonstrate that learning slowly changing representations in central vision leads to more semantic object representations, compared to a standard training on the whole field of view. Our analysis shows that this stems from a complementary effect between central vision, fixational eye movements and learning temporal slowness: while central vision favors the extraction of foreground object information versus background information, learning with temporal slowness during eye fixations distills semantic information in object representations Together, our work sheds light on how humans build semantic object representations from a natural visual experience. Furthermore, our approach may inspire more semantically grounded and efficient learning strategies in embodied AI applications.

2 Related works

Egocentric SSL. The increased availability of datasets collected with head-mounted cameras (Grauman et al., 2022; Sullivan et al., 2021; Long et al., 2024; Greene et al., 2024; Ma et al., 2024) recently induced a surge for training egocentric SSL (Orhan et al., 2024). To the best of our knowledge, all these approaches train egocentric SSL models using the entire high-resolution field of view captured by head-mounted cameras (Orhan et al., 2024; He et al., 2022; Zhou et al., 2022; Emin Orhan, 2024). Previous work also found that endowing SSL models with representations that slowly change over time can slightly boost object learning (Orhan et al., 2020). A related line of work leverages egocentric data to train vision models useful for solving robotic tasks. VC-1 (Majumdar et al., 2023) is trained on egocentric and third-person videos. R3M (Nair et al.) and VIP (Ma et al.)

both notably learn slowly changing representations on Ego4D. We show in Section 4 that training on gaze-based central vision elicits better object representations. Other works try to extract the correspondences between objects' views in videos to learn visual representations (Jabri et al., 2020; Venkataramanan et al., 2024; Salehi et al., 2023; Parthasarathy et al., 2023; Gordon et al., 2020). Here, we are rather interested in understanding how the biological importance of central vision may impact egocentric SSL.

Time-based SSL. Many works previously proposed to learn similar representations for close-in-time visual inputs (Wiskott & Sejnowski, 2002; Földiák, 1991). More recently, this learning principle has been integrated into mainstream SSL methods (Aubret et al., 2022a). However, these works do not leverage in-the-wild egocentric data, such as synthetic (Aubret et al., 2022a; Schaumlöffel et al., 2023) or curated (Aubret et al., 2024b; Sanyal et al., 2023; Aubret et al., 2024a) visual sequences of interactions with objects. Other works use third-person ones (Sermanet et al., 2018), videos recorded by a car (Jayaraman & Grauman, 2015; 2016), movie video clips (Jayaraman & Grauman, 2016), chicks egocentric perspective (Pandey et al., 2024) or object-tracking datasets (Xu & Wang, 2021).

Central and peripheral vision in deep learning. Many studies modeled the changing resolution of the retina. Previous works already showed that it can make supervised representations more adversarially robust (Vuyyuru et al., 2020), improve the computational efficiency of the training process (Lukanov et al., 2021) and induce a stronger center bias (Deza & Konkle, 2020). Other works combined a bio-inspired central vision with attention mechanisms for temporally extended image recognition (e.g. Almeida et al. (2018)). In the context of SSL, Wang et al. (2021) argues that foveation can mimic the impact of the Crop/Resize data-augmentation, widespread in SSL. In line with our work, recent studies combined retina modeling with time-based SSL. Aubret et al. (2022b) showed that a progressive blur towards visual periphery can make visual representations slightly more transferable across backgrounds, and Yu et al. (2024) showed that gaze patterns in central vision may support a view-invariant object learning. However, these two works trained SSL models with a tiny number of objects (10 and 24, respectively). In contrast, using scaled human data allows us to study the role of bio-inspired learning with respect to semantic recognition abilities.

Learning context-wise object representations Only few works studied the emergence of similar visual representations for objects that co-occur in the same context. One work (Bonner & Epstein, 2021) proposed to learn similar representations for objects that co-occur in images of natural scenes. They did not study temporal co-occurrences of objects. Their impact on semantic representations was studied by Aubret et al. (2024a) with a curated dataset showing egocentric rotations around images and hand-made statistics of object transitions. Thus, it remains unclear how and whether the natural experience of humans supports the construction of context-wise object representations.

3 METHOD

We aim to study the combined impact of high-resolution central vision and the slowness principle on visual learning in humans. We use the largest-to-date dataset of egocentric videos (Ego4D) and estimate human gaze locations with a state-of-the-art model of human gaze prediction (Section 3.1). To simulate the biological importance of central vision, we simply crop the visual area around a gaze location. To model biological learning, the created sequence of visual inputs feeds an SSL model that learns slowly changing representations, which is described in Section 3.2. Figure 1 illustrates the main steps of the pipeline. Finally, we evaluate the ability of our model to capture different semantic facets of objects, using the approaches detailed in Section 3.3 and Section 3.4.

3.1 DATASET

To simulate the visual experience of humans, we use the Ego4D dataset (Grauman et al., 2022). This dataset contains 3,600 hours of videos collected through head-mounted cameras, which corresponds to approximately 5 months of visual experience. 931 participants coming from 74 worldwide locations wore a camera for one to ten hours. Thus, Ego4D arguably represents much more than 5 months of experience for a single average human in terms of diversity, although it is hard to make precise estimates. We use videos with a resolution of 540×540 pixels and extract their frames at approximately 5 fps, following previous findings that a higher fps does not boost the learning process (Sheybani et al., 2024).

During frame extraction, we create small clips of 5 seconds (25 frames) that we sequentially load into memory. We gather 24 frames of these 25 frames and split them into three sequences of 8 frames. For Ego4D videos recorded with an eye-tracker (45 hours), we do not further process the frames and associate them with ground-truth gaze location. For all other videos, we feed each sequence into GLC, a state-of-the-art model of human gaze prediction trained on the Ego4D subset that contains gaze locations (Lai et al., 2022). This model uses spatio-temporal information to generate a saliency map for each of the 8 frames. Compared to single-image saliency models (Riche & Mancas, 2016), this allows the model to generate a temporally consistent gaze location and to leverage more cues (e.g. motion). For each frame, we take as gaze location the position of the most salient pixel (x_g, y_g). Our final preprocessed dataset contains 64,380,024 images.

To simulate the importance of central vision in humans, we crop a $N \times N$ squared area centered on the gaze location for each frame. The crop boundaries may go beyond the image boundaries; in this case, we minimally shift the crop such that its boundaries remain in the image.

3.2 Bio-inspired learning

Since most of the human visual experience is unsupervised, we train SSL models on the simulated experience in central vision. These models learn high-level visual representations without any explicit supervision, like human-provided labels. In this work, we focus specifically on the third version of Momentum Contrast (MoCoV3) (Chen et al., 2021), which is one of the best SSL models in the literature. The original MoCoV3 works by learning invariant representations to color- and spatial-based transformations of an image (e.g. horizontal flip, color jittering . . .). To implement the biological principle of temporal slowness, we further adapt the model to also learn slowly changing visual representations, following (Aubret et al., 2022a; Pandey et al., 2024).

For a given input image x_t in a batch, we randomly sample an indirect temporal neighbor $x_{t'}$ within a temporal window ΔT, from the same video recording. The two images capture the same scene from different moments in time, providing a temporally varied view. We compute the embeddings of images $q_t = f_q(x_t)$ and $k_{t'} = f_k(x'_t)$ using a query feature extractor f_q and a momentum feature extractor f_k, both implemented as neural networks. Finally, for a pair $(q_t, k_{t'})$, the query encoder is updated by minimizing the InfoNCE loss (van den Oord et al., 2019):

$$\mathcal{L}_{q_t} = -\log \frac{\exp\big(\mathrm{sim}(q_t, k_{t'})/\tau\big)}{\sum_{i=0}^{K} \exp\big(\mathrm{sim}(q_t, k_i)/\tau\big)} \tag{1}$$

where sim denotes cosine similarity, τ is a temperature hyperparameter, and K represents the outputs of f_k from the same training batch. Intuitively, the objective increases the similarity between representations of temporally close views (x_t and $x_{t'}$) while enhancing the dissimilarity between all views (x_t and x_i). The momentum encoder parameters θ_k are updated via exponential moving average of the query encoder θ_q: $\theta_k \leftarrow m\theta_k + (1-m)\theta_q$, with momentum coefficient m.

3.3 Evaluation of image recognition abilities

We follow standard SSL transfer protocols, evaluating frozen representations via linear probing across diverse downstream tasks grouped by semantic focus (see below). For each dataset, we train a linear classifier on top of the frozen features of the pre-trained encoder for 100 epochs. We apply the standard crop/resize and horizontal flip augmentations during training and report the accuracy on a center crop of validation images.

Object categorization: To assess the categorization ability of the models, we consider the ImageNet-1k (Russakovsky et al., 2015), ImageNet100 (Tian et al., 2020) and CIFAR100 (Krizhevsky et al., 2009) datasets, including reduced subsets of ImageNet-1k (1%, 10%) (Chen et al., 2020). We also analyze object categorization tasks that contain a tiny number of classes in Appendix D.3.

Fine-grained object categorization: Most classes in the two previous groups are for basic-level category recognition (e.g. car, trucks, bananas . . .). Here, we rather assess categorization at the supraordinate level (e.g. for cars, differentiating a 2012 VW Polo from a 2012 BMW M3). This requires a model to extract more details about an object. We consider a wide range of supraordinate categories: Flowers101 (Nilsback & Zisserman, 2008), Stanford Cars (Krause et al., 2013), Oxford Pet (Parkhi et al., 2012), FGVC-Aircraft (Maji et al., 2013), DTD (Cimpoi et al., 2014).

Instance-level object recognition: We evaluate object instance recognition when exposed in front of different backgrounds with different orientations. We use ToyBox (Wang et al., 2017), COIL100 (Nene et al., 1996), Core50 (Lomonaco & Maltoni, 2017). Core50 mostly allows us to assess the robustness of the representation to changing backgrounds, while ToyBox and COIL100 present objects in different positions and orientations. We explain in Appendix B how we split the train and test splits. We do not apply a center crop on COIL100.

Scene recognition: For scene recognition, we focus on Places365-standard (Zhou et al., 2017a). This dataset contains 1.8 million images from 365 scene categories and is commonly used to probe scene-level representations.

3.4 Evaluation of the context-wise organization of object representations

In order to evaluate whether the knowledge about 3D object co-occurrences can naturally emerge from our model, we compare the representations of our models with representations specifically built to encode objects' co-occurrence structure.

Object Co-occurrence Representations: To model the latent semantic structure of natural scenes, we extract object co-occurrence statistics from three large-scale image datasets: COCO (Lin et al., 2014), ADE20K (Zhou et al., 2017b), and Visual Genome (VG) (Krishna et al., 2017). These datasets vary in label density and semantic granularity. COCO contains coarse object categories with dense instance annotations; ADE has finer-grained, segmentation-level labels; and VG offers a rich, albeit noisy, semantic graph structure. For each dataset, we construct a co-occurrence matrix $X \in \mathbb{R}^{N \times N}$, where X_{ij} counts how often object i appears with object j in the same image. We train GloVe (Global Vectors for Word Representation) (Pennington et al., 2014) on these matrices to derive low-dimensional representations that encode this co-occurrence structure. We refer to Appendix A for more details.

Model-to-Semantics Alignment: To assess whether neural network representations encode a similar semantic structure as the co-occurrence embeddings, we perform a representation similarity analysis using Centered Kernel Alignment (CKA) (Kornblith et al., 2019).

We map object classes from the co-occurrence matrices to their corresponding WordNet synsets (Miller, 1995). Then, we retrieve representative images from the THINGS dataset (Hebart et al., 2023), which contains isolated object instances with a naturalistic appearance. For each object, we extract activations from all layers of a given model and average across object images, resulting in a single feature vector per object and layer. We then compute the linear CKA score between each layer's object representation matrix and the GloVe embedding matrix. Alternatively, we concatenate the representations across layers to compute a global CKA score, which serves as a summary measure of semantic alignment for the entire model.

We repeat all evaluations across 100 GloVe seeds. We perform a paired t-test across the seeds to compare the CKA scores of each model under identical co-occurrence conditions and access statistical significance between models. Throughout the experiment section, we report mean scores, standard deviations, and significance levels.

3.5 Implementation details

We use the *solo-learn* implementation of the model (da Costa et al., 2022), with ResNet-50 (He et al., 2016) and ViT-B/16 (Dosovitskiy et al., 2020) backbones. For the MoCoV3 loss, a two-layer MLP (hidden dimension 4096) projects features into a 256-dimensional embedding space. Models are trained for one epoch on Ego4D; training longer yielded only minor gains ($\approx +0.5\%$) at substantial computational cost, due to the large but redundant dataset sampled at 5 fps. Full hyperparameters are given in Appendix C.

4 Experiments

We aim to assess the impact of learning visual representations with bio-inspired central vision and temporal slowness. Thus, we compare our model "Bio-inspired Learning" to a baseline from previous work, "Frames Learning", which uses the full field of view and omits slowness learning during training (Orhan & Lake, 2024).

4.1 Bio-inspired learning from natural experience boosts object recognition.

Table 1: Linear probe accuracy on various datasets across two architectures, grouped by semantic category. For each semantic group, we report the average recognition accuracy. For bio-inspired vision, we use $\Delta T = 3$ for ResNet50 and $\Delta T = 1$ for ViT.

	ResNet50		ViT-B/16	
Dataset	Frames Learning	Bio-inspired Learning	Frames Learning	Bio-inspired Learning
		Category recognition		
ImgNet-1k	49.50	**49.58**	49.47	**49.86**
ImgNet-1k 10%	**35.53**	35.34	37.65	**38.10**
ImgNet-1k 1%	19.23	**20.25**	19.51	**20.10**
ImgNet-100	**70.44**	70.34	70.04	**70.12**
CIFAR100	53.53	**59.21**	61.73	**62.67**
Average	45.65	**46.94**	47.68	**48.17**
		Fine-grained recognition		
DTD	47.24	**57.06**	59.89	**62.23**
FGVCAircraft	12.83	**15.77**	**28.87**	28.60
Flowers102	43.72	**49.01**	76.35	**77.05**
OxfordIIITPet	46.68	**47.03**	54.41	**56.26**
StanfordCars	18.70	**23.25**	**33.30**	33.26
Average	33.84	**38.42**	50.56	**51.58**
		Instance recognition		
ToyBox	89.75	**92.61**	92.94	**95.03**
COIL100	64.53	**80.12**	79.24	**86.94**
Core50	22.82	**28.26**	**24.02**	23.77
Average	59.03	**67.00**	65.40	**68.58**
		Scene recognition		
Places365	**43.02**	42.95	**44.49**	39.84

Here, we investigate whether the bio-inspired mechanisms of slowness learning and central vision support the construction of object recognition abilities. We compare our models with training SSL models without temporal slowness and from raw frames. In Table 1, we observe that bio-inspired learning promotes category recognition, fine-grained recognition and instance object recognition. The improvement is particularly important for fine-grained and instance object recognition. For scene recognition, training on full frames consistently leads to better results. We further investigate this in the following paragraph. We conclude that bio-inspired learning from a natural visual experience promotes better object representations. In Appendix D.1, we support our claim with additional model architectures.

Central vision makes representations more object-centered In this section, we analyze how focusing on central vision affects the learned visual representations. First, we study the impact of the size of the gaze-based crops N on visual learning. In Figure 2, we observe a sweet spot in the intermediate crop size $[224, 336]$ for all object-centered datasets. This sweet spot is located at $N = 336$ for category and instance recognition, while $N = 224$ seems to be better for fine-grained recognition. $N = 112$ lower-bounds all semantic recognition accuracies, indicating that it probably dismisses too much information about the image. Interestingly, scene recognition accuracy consistently decreases as we make the crops smaller. We conclude that using the whole field of view elicits more scene-based representations, versus object-centered representations for central vision.

Large fields of view tend to display scenes with complex backgrounds and relatively small objects. It may be that extracting background features shortcuts object learning to satisfy the spatio-temporal invariance objective of the SSL model. To investigate that, we take the ImageNet-9 dataset, a dataset of natural images designed to investigate the background sensitivity of models (Xiao et al., 2021). We compute the category recognition accuracy of our model with a linear probe trained on ImageNet-1k

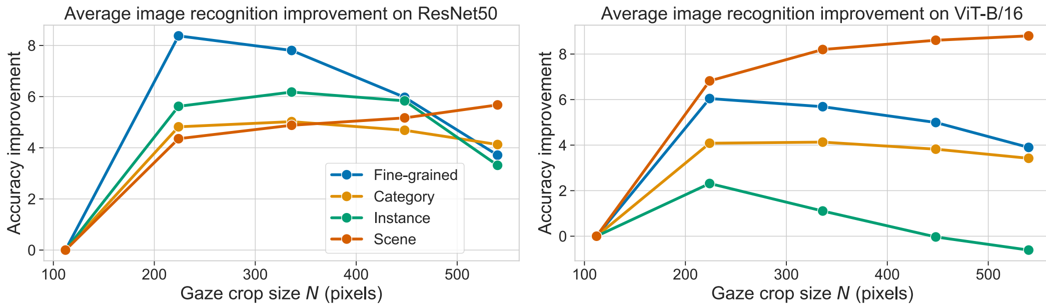

Figure 2: Impact of the gaze-based crop size on different semantic image recognition groups for ResNet50 and ViT-B/16. We compute the average improvement for each semantic group of datasets with respect to $N = 112$. We use a temporal window of $\Delta T = 3$s. Full results are provided in the Appendix Table 6.

in all settings (normal image, without background and with different ways to remove the object). We first find that training on central vision also benefits category recognition on normal images for this dataset (80% versus 75% on ResNet50). Then, we compute the recognition accuracy when removing the background (Missing background) and when removing the object (Missing object). When removing the object, we average the recognition accuracies of the different ways of removing the foreground object (cf. Xiao et al. (2021)). To obtain a measure of background and object sensitivity irrespective of the raw performance of the model, we subtract the missing object and missing background accuracies by the recognition accuracy on normal images.

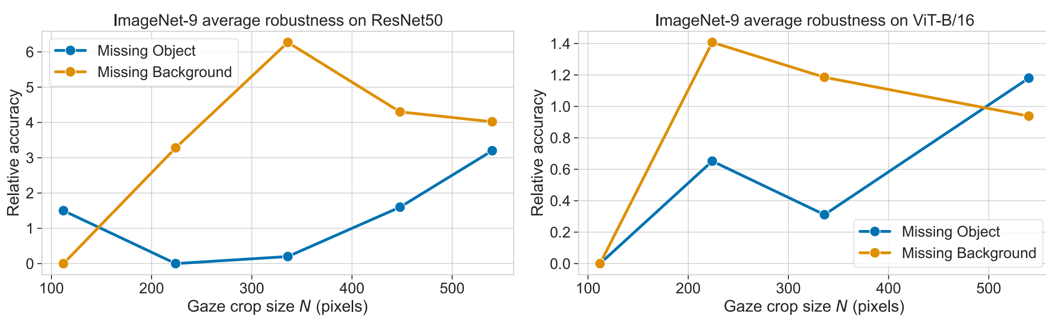

Figure 3: ImageNet-9 recognition sensitivity to missing background or missing foreground object. We show the relative improvement with respect to the worst model for the two settings. The higher, the more relatively robust is the representation to missing backgrounds or missing objects. We use a temporal window $\Delta T = 3$s.

In Figure 3, we clearly observe that training on an intermediate size of central vision ($N \in \{224, 336\}$) allows to rely more on the foreground object (missing background), and less on the background (missing object). We assume this is because training on central vision removes much of the background information while often keeping the foreground object intact. For $N = 112$, there is an opposite trend, presumably because this is too small to display objects fully. Overall, we conclude that central vision leads to better object-centered representations because it learns to extract less background information.

Slowness learning supports object learning Previous work on standard videos suggests that learning representations that slowly vary for up to $t = 1$ seconds can be beneficial for visual learning (Xu & Wang, 2021). However, focusing on gaze-based central vision during egocentric learning provides semantically different temporal dynamics compared to using the whole field of view of, e.g., a movie clip. In Figure 4, we present the impact of the level of temporal slowness on visual representations trained with central vision. We observe that temporal slowness is critical for learning representations with respect to all the semantic aspects investigated ($\Delta T = 0$ versus $\Delta T = 3$ for

Resnet50 and $\Delta T = 1$ for ViT). We note an exception for category recognition with ViT-B/16, for which the reason is currently unclear to us. We provide detailed results in Appendix E, which further show that the best temporal window ΔT is overall consistent for datasets within a semantic group.

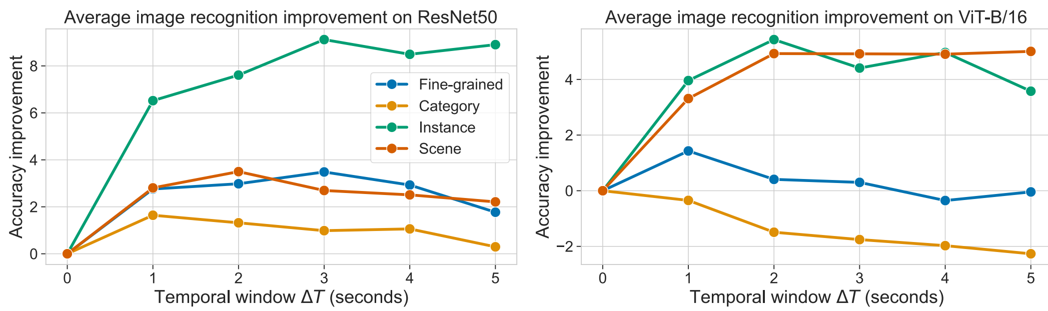

Figure 4: Impact of the temporal window of slowness learning on different semantic image recognition groups. We compute the average improvement for each semantic group of datasets with respect to $\Delta T = 0$ second. We use a crop of size $N = 224$. Full results are provided in the Appendix Table 7.

4.2 Human fixational eye movements support object learning

Human gaze behaviors are characterized by relatively long fixations interleaved by short and distant saccadic eye movements, a pattern captured by the gaze estimation model (cf. Appendix D.4). To systematically investigate the importance of human slow/fast fixational eye movements for object learning, we group fixational eye movements as a sequence of movements with speeds inferior to $P/200\,\mathrm{px\,ms^{-1}}$, where P denotes the maximum eye movement in $200\,\mathrm{ms}$. Then, during training, we ensure that two temporal neighbors, even if spaced in time, belong to the same fixation. We also consider another extreme setting ($P = 0$) that completely discards eye movements, for which we always crop the center of the frame. Note that the resulting visual sequences always vary over time due to head movements of the camera wearer.

In Figure 5, we observe that bio-inspired training with estimated human gaze ($P = \infty$) outperforms its gaze-agnostic ($P = 0$) counterparts in most cases. Notably, for instance recognition, gaze-crops yield substantial improvements of +4.17% for ResNet and +1.14% for ViT. This suggests that gaze-directed views provide beneficial cues for distinguishing specific objects over time. This shows that using simulated gaze locations of humans (weakly) boosts semantic object learning. When training is restricted to fixational eye movements ($P \in \{45, 30\}$, we observe that the model learns better object representations, relative to category, fine-grained and instance recognition. This indicates that saccadic eye movements are harmful for learning about objects. In practice, humans are aware of whether their movement is fixational or saccadic (O'regan & Noë, 2001): our finding also suggests that this knowledge can be leveraged to improve semantic object learning.

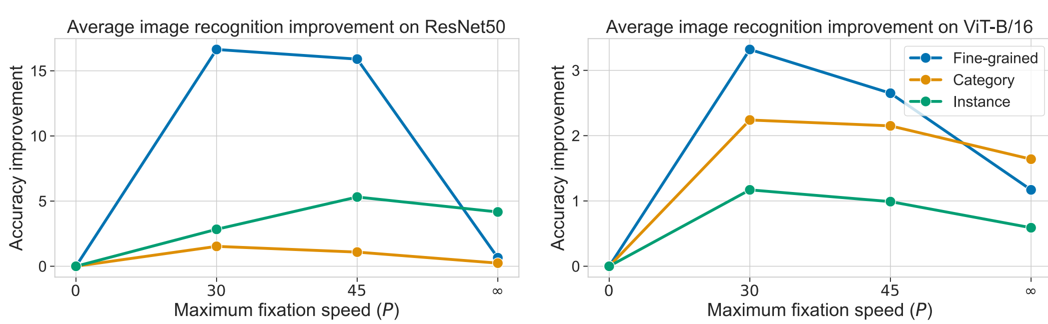

Figure 5: Impact of varying the maximal eye movement during a fixation. We report the average improvement for each semantic dataset group relative to the $P = 0$ baseline. All models are trained under the same settings as the bio-inspired models described in section 4.1.

4.3 Slowness learning better associates objects that occur in the same context

In this section, we explore how slowness learning and central vision jointly shape the semantic similarity between categories of objects. In Table 2, bio-inspired learning yields object representations whose inter-object similarities more closely align with object co-occurrence statistics, capturing contextual relationships more effectively than "Frames learning". This effect is moderate with ResNet50 ($+0.010$ CKA score), but important with ViT ($+0.028$ CKA score). We further observe that models trained with slowness learning produce significantly higher CKA similarity with co-occurrence embeddings than static models. This suggests that temporal slowness is essential for semantic learning.

In practice, the context-based organization of objects may be captured by learning both spatial and temporal co-occurrences. The former may be particularly pronounced when the image contains several objects, such as in the full frame. The latter may result from slowness learning when consecutively observing different objects (Aubret et al., 2024a). Thus, we investigate their relative role in shaping context-based object representations. In Table 2, we observe an increase in CKA similarity as we remove the focus on central vision (w/o Central vision), presumably because co-occurring objects are often spatially distributed and may not appear within the parafovea-sized crop. However, we observe a minor gap between "Bio-inspired Learning" suggesting that the integration of central vision over time compensates for limited spatial view. We support the generalization of our findings across datasets in Appendix D.2. In general, we conclude that slowness learning suffice to build context-based object representations.

Table 2: CKA similarity between learned representations and GloVe-based object co-occurrence embeddings, computed on the COCO dataset. Higher values indicate stronger semantic alignment.

	Frames Learning	Bio-inspired Learning	w/o Slowness	w/o Central Vision
ResNet50	0.315 ± 0.004	$\mathbf{0.325 \pm 0.004}$	0.320 ± 0.004	0.335 ± 0.003
ViT-B/16	0.453 ± 0.004	$\mathbf{0.481 \pm 0.004}$	0.406 ± 0.004	0.487 ± 0.003

5 Conclusion

We investigated whether the relative importance of central vision (vs. peripheral) and the biological learning principle of slowness jointly support semantic object learning in humans. We simulated humans' gaze locations on the largest-to-date dataset of egocentric videos and extracted the visual areas surrounding the gaze locations. Then, we trained a variant of a mainstream SSL model that learns slowly changing visual representations. Our extensive experiments demonstrate that extracting slowly changing information in central vision allows visual representations to better encode different semantic facets of human perception. This includes the between-object similarity based on their context of co-occurrences, their basic category, fine-grained (or supraordinate) category and their instance identity. Our analysis shows that central vision elicits the extraction of more object-related features than background features. In addition, we found that fixational eye movements specifically support such bio-inspired learning.

Humans remain far more efficient in learning semantic visual representations. For example, the accuracy of the Top-5 linear probe with ImageNet-1k 1% barely goes beyond 40%, compared to about 90% for humans (Russakovsky et al., 2015; Orhan, 2023). We foresee several perspectives. First, although visual semantic learning continues far beyond the maturation of the retina and eye movements, the visuo-motor experience of infants in early development differs from the experience of adults modeled in the present work (Ayzenberg & Behrmann, 2024). Future work will have to investigate whether this early experience synergizes with slowness learning and central vision. Second, the gaze estimation model extracts only bottom-up saliency maps of adult views; modeling the learning of eye movements may offer a way to curate the experience for visual learning. Third, improving the model may require more realistic retinal processing in egocentric SSL models (Wang et al., 2021), with a more gradual attenuation of visual information towards the periphery. Yet, our work shows that prioritizing slow information in central vision is key for learning strong semantic representations, marking a step toward understanding their human development.

REFERENCES

Ana Filipa Almeida, Rui Figueiredo, Alexandre Bernardino, and José Santos-Victor. Deep networks for human visual attention: A hybrid model using foveal vision. In *ROBOT 2017: Third Iberian Robotics Conference: Volume 2*, pp. 117–128. Springer, 2018.

Stuart M Anstis. Chart demonstrating variations in acuity with retinal position. *Vision research*, 14 (7):589–592, 1974.

Arthur Aubret, Markus Ernst, Céline Teulière, and Jochen Triesch. Time to augment self-supervised visual representation learning. *arXiv preprint arXiv:2207.13492*, 2022a.

Arthur Aubret, Céline Teulièr, and Jochen Triesch. Toddler-inspired embodied vision for learning object representations. In *2022 IEEE International Conference on Development and Learning (ICDL)*, pp. 81–87. IEEE, 2022b.

Arthur Aubret, Timothy Schaumlöffel, Gemma Roig, and Jochen Triesch. Learning object semantic similarity with self-supervision. pp. 1–6, 2024a.

Arthur Aubret, Céline Teulière, and Jochen Triesch. Self-supervised visual learning from interactions with objects. *arXiv preprint arXiv:2407.06704*, 2024b.

Vladislav Ayzenberg and Marlene Behrmann. Development of visual object recognition. *Nature Reviews Psychology*, 3(2):73–90, 2024.

Michael F Bonner and Russell A Epstein. Object representations in the human brain reflect the co-occurrence statistics of vision and language. *Nature communications*, 12(1):4081, 2021.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9640–9649, 2021.

Ming-Chang Chiu, Pin-Yu Chen, and Xuezhe Ma. Better may not be fairer: A study on subgroup discrepancy in image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4956–4966, 2023.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014. doi: 10.1109/CVPR.2014.461.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Victor Guilherme Turrisi da Costa, Enrico Fini, Moin Nabi, Nicu Sebe, and Elisa Ricci. solo-learn: A library of self-supervised methods for visual representation learning. *Journal of Machine Learning Research*, 23(56):1–6, 2022. URL http://jmlr.org/papers/v23/21-1155.html.

Arturo Deza and Talia Konkle. Emergent properties of foveated perceptual systems. *arXiv preprint arXiv:2006.07991*, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

A Emin Orhan. Hvm-1: Large-scale video models pretrained with nearly 5000 hours of human-like video data. *arXiv e-prints*, pp. arXiv–2407, 2024.

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

Daniel Gordon, Kiana Ehsani, Dieter Fox, and Ali Farhadi. Watching the world go by: Representation learning from unlabeled videos. *arXiv preprint arXiv:2003.07990*, 2020.

Priya Goyal, Piotr Dollár, Ross B. Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *ArXiv preprint*, abs/1706.02677, 2017.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19012, 2022.

Michelle R Greene, Benjamin J Balas, Mark D Lescroart, Paul R MacNeilage, Jennifer A Hart, Kamran Binaee, Peter A Hausamann, Ronald Mezile, Bharath Shankar, Christian B Sinnott, et al. The visual experience dataset: Over 200 recorded hours of integrated eye movement, odometry, and egocentric video. *arXiv preprint arXiv:2404.18934*, 2024.

Tanmay Gupta, Alexander Schwing, and Derek Hoiem. Vico: Word embeddings from visual co-occurrences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7425–7434, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Martin N Hebart, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife*, 12:e82580, 2023.

Allan Jabri, Andrew Owens, and Alexei Efros. Space-time correspondence as a contrastive random walk. *Advances in neural information processing systems*, 33:19545–19560, 2020.

Dinesh Jayaraman and Kristen Grauman. Learning image representations tied to ego-motion. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1413–1421, 2015.

Dinesh Jayaraman and Kristen Grauman. Slow and steady feature analysis: higher order temporal coherence in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3852–3861, 2016.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMLR, 2019.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013. doi: 10.1109/ICCVW.2013.77.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Bolin Lai, Miao Liu, Fiona Ryan, and James Rehg. In the eye of transformer: Global-local correlation for egocentric gaze estimation. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022. URL https://bmvc2022.mpi-inf.mpg.de/0227.pdf.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pp. 740–755. Springer, 2014.

Vincenzo Lomonaco and Davide Maltoni. Core50: a new dataset and benchmark for continuous object recognition. In *Conference on robot learning*, pp. 17–26. PMLR, 2017.

Bria Long, Violet Xiang, Stefan Stojanov, Robert Z Sparks, Zi Yin, Grace E Keene, Alvin WM Tan, Steven Y Feng, Chengxu Zhuang, Virginia A Marchman, et al. The babyview dataset: High-resolution egocentric videos of infants' and young children's everyday experiences. *arXiv preprint arXiv:2406.10447*, 2024.

Hristofor Lukanov, Peter König, and Gordon Pipa. Biologically inspired deep learning model for efficient foveal-peripheral vision. *Frontiers in Computational Neuroscience*, 15:746204, 2021.

Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, et al. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. *arXiv preprint arXiv:2406.09905*, 2024.

Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. In *The Eleventh International Conference on Learning Representations*.

Florian P Mahner, Lukas Muttenthaler, Umut Güçlü, and Martin N Hebart. Dimensions underlying the representational alignment of deep neural networks with humans. *Nature Machine Intelligence*, 7(6):848–859, 2025.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *Technical Report*, abs/1306.5151, 2013.

Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Tingfan Wu, Jay Vakil, et al. Where are we in the search for an artificial visual cortex for embodied intelligence? *Advances in Neural Information Processing Systems*, 36: 655–677, 2023.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. In *6th Annual Conference on Robot Learning*.

Ryoichi Nakashima and Satoshi Shioiri. Why do we move our head to look at an object in our peripheral region? lateral viewing interferes with attentive search. *PloS one*, 9(3):e92284, 2014.

Sameer A Nene, Shree K Nayar, Hiroshi Murase, et al. Columbia object image library (coil-20). 1996.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729, 2008. doi: 10.1109/ICVGIP.2008.47.

J Kevin O'regan and Alva Noë. A sensorimotor account of vision and visual consciousness. *Behavioral and brain sciences*, 24(5):939–973, 2001.

A Emin Orhan. Scaling may be all you need for achieving human-level object recognition capacity with human-like visual experience. *arXiv preprint arXiv:2308.03712*, 2023.

A Emin Orhan and Brenden M Lake. Learning high-level visual representations from a child's perspective without strong inductive biases. *Nature Machine Intelligence*, 6(3):271–283, 2024.

A Emin Orhan, Wentao Wang, Alex N Wang, Mengye Ren, and Brenden M Lake. Self-supervised learning of video representations from a child's perspective. *arXiv preprint arXiv:2402.00300*, 2024.

Emin Orhan, Vaibhav Gupta, and Brenden M Lake. Self-supervised learning through the eyes of a child. *Advances in Neural Information Processing Systems*, 33:9960–9971, 2020.

Lalit Pandey, Samantha Wood, and Justin Wood. Are vision transformers more data hungry than newborn visual systems? *Advances in Neural Information Processing Systems*, 36, 2024.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, 2012. doi: 10.1109/CVPR.2012.6248092.

Nikhil Parthasarathy, SM Eslami, Joao Carreira, and Olivier Henaff. Self-supervised video pretraining yields robust and more human-aligned visual representations. *Advances in Neural Information Processing Systems*, 36:65743–65765, 2023.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Christian Quaia and Richard J Krauzlis. Object recognition in primates: What can early visual areas contribute? *Frontiers in Behavioral Neuroscience*, 18:1425496, 2024.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021.

Nicolas Riche and Matei Mancas. Bottom-up saliency models for still images: A practical review. *From Human Attention to Computational Attention: A Multidisciplinary Approach*, pp. 141–175, 2016.

Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019–1025, 1999.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Mohammadreza Salehi, Efstratios Gavves, Cees GM Snoek, and Yuki M Asano. Time does tell: Self-supervised time-tuning of dense image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16536–16547, 2023.

Deepayan Sanyal, Joel Michelson, Yuan Yang, James Ainooson, and Maithilee Kunda. A computational account of self-supervised visual learning from egocentric object play. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45, 2023.

Timothy Schaumlöffel, Arthur Aubret, Gemma Roig, and Jochen Triesch. Caregiver talk shapes toddler vision: A computational study of dyadic play. In *2023 IEEE International Conference on Development and Learning (ICDL)*, pp. 67–72. IEEE, 2023.

Heiko H Schütt, Lars OM Rothkegel, Hans A Trukenbrod, Ralf Engbert, and Felix A Wichmann. Disentangling bottom-up versus top-down and low-level versus high-level influences on eye movements over time. *Journal of vision*, 19(3):1–1, 2019.

Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 1134–1141. IEEE, 2018.

Saber Sheybani, Himanshu Hansaria, Justin Wood, Linda Smith, and Zoran Tiganj. Curriculum learning with infant egocentric videos. *Advances in Neural Information Processing Systems*, 36, 2024.

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*, 5:20–29, 2021.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 776–794. Springer, 2020.

Jacopo Turini and Melissa Le-Hoa Võ. Hierarchical organization of objects in scenes is reflected in mental representations of objects. *Scientific Reports*, 12(1):20068, 2022.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL https://arxiv.org/abs/1807.03748.

Shashanka Venkataramanan, Mamshad Nayeem Rizve, João Carreira, Yuki Asano, and Yannis Avrithis. Is imagenet worth 1 video? learning strong image encoders from 1 long unlabelled video. In *ICLR 2024-Twelfth International Conference on Learning Representations*, pp. 1–21, 2024.

Manish Reddy Vuyyuru, Andrzej Banburski, Nishka Pant, and Tomaso Poggio. Biologically inspired mechanisms for adversarial robustness. *Advances in Neural Information Processing Systems*, 33: 2135–2146, 2020.

Binxu Wang, David Mayo, Arturo Deza, Andrei Barbu, and Colin Conwell. On the use of cortical magnification and saccades as biological proxies for data augmentation. In *SVRHM 2021 Workshop@ NeurIPS*, 2021.

Xiaohan Wang, Fernanda M. Eliott, James Ainooson, Joshua H. Palmer, and Maithilee Kunda. An object is worth six thousand pictures: The egocentric, manual, multi-image (emmi) dataset. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017.

Heinz Wässle, Ulrike Grünert, Jürgen Röhrenbeck, and Brian B Boycott. Cortical magnification factor and the ganglion cell density of the primate retina. *Nature*, 341(6243):643–646, 1989.

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

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations*, 2021.

Jiarui Xu and Xiaolong Wang. Rethinking self-supervised correspondence learning: A video frame-level similarity perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10075–10085, 2021.

Yang You, Igor Gitman, and Boris Ginsburg. Scaling SGD batch size to 32k for imagenet training. *ArXiv preprint*, abs/1708.03888, 2017. URL http://arxiv.org/abs/1708.03888.

H-H Yu, TA Chaplin, and MGP Rosa. Representation of central and peripheral vision in the primate cerebral cortex: Insights from studies of the marmoset brain. *Neuroscience Research*, 93:47–61, 2015.

Zhengyang Yu, Arthur Aubret, Marcel C Raabe, Jane Yang, Chen Yu, and Jochen Triesch. Active gaze behavior boosts self-supervised object learning. *arXiv preprint arXiv:2411.01969*, 2024.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017a.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 633–641, 2017b.

Pan Zhou, Yichen Zhou, Chenyang Si, Weihao Yu, Teck Khim Ng, and Shuicheng Yan. Mugs: A multi-granular self-supervised learning framework. *arXiv preprint arXiv:2203.14415*, 2022.

Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C Frank, James J DiCarlo, and Daniel LK Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, 2021.