seq-JEPA: Autoregressive Predictive Learning of Invariant-Equivariant World Models

Hafez Ghaemi **Eilif B. Muller*** **Shahab Bakhtiari***
Université de Montréal - Mila
[hafez.ghaemi,eilif.muller,shahab.bakhtiari]@umontreal.ca

Abstract

Joint-embedding predictive architecture (JEPA) is a self-supervised learning (SSL) paradigm with the capacity of world modeling via action-conditioned prediction. Previously, JEPA world models have been shown to learn action-invariant or action-equivariant representations by predicting one view of an image from another. Unlike JEPA and similar SSL paradigms, animals, including humans, learn to recognize new objects through a sequence of active interactions. To introduce *sequential* interactions, we propose *seq-JEPA*, a novel SSL world model equipped with an autoregressive memory module. Seq-JEPA aggregates a sequence of action-conditioned observations to produce a global representation of them. This global representation, conditioned on the next action, is used to predict the latent representation of the next observation. We empirically show the advantages of this sequence of action-conditioned observations and examine our sequential modeling paradigm in two settings: (1) *predictive learning across saccades*; a method inspired by the role of eye movements in embodied vision. This approach learns self-supervised image representations by processing a sequence of low-resolution visual patches sampled from image saliencies, without any hand-crafted data augmentations. (2) *invariance-equivariance trade-off*; seq-JEPA's architecture results in automatic separation of invariant and equivariant representations, with the aggregated autoregressor outputs being mostly action-invariant and the encoder output being equivariant. This is in contrast with many equivariant SSL methods that expect a single representational space to contain both invariant and equivariant features, potentially creating a trade-off between the two. Empirically, seq-JEPA achieves competitive performance on both invariance and equivariance-related benchmarks compared to existing methods. Importantly, both invariance and equivariance-related downstream performances increase as the number of available observations increases.

1 Introduction

Self-supervised learning (SSL) in latent space has made significant progress in visual representation learning in recent years [van den Oord et al. [2019], Misra and van der Maaten [2020], He et al. [2020], Chen et al. [2020], Grill et al. [2020], Chen and He [2021], Caron et al. [2020, 2021], Zbontar et al. [2021], Bardes et al. [2022], Baevski et al. [2022], Assran et al. [2023]], almost closing the gap with supervised learning in many downstream tasks. Many of these SSL methods are based on joint-embedding architectures (JEAs), or joint-embedding predictive architectures (JEPAs). In JEAs, two different views of a sample image, usually generated by hand-crafted data augmentations, are encoded to produce invariant representations via sample-contrastive [Chen et al., 2020, He et al.,

*Equal contribution

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

2020, Dwibedi et al., 2021, HaoChen et al., 2021, Yeh et al., 2022, Caron et al., 2020, 2021, Assran et al., 2022] or dimension-contrastive objectives [Zbontar et al., 2021, Ermolov et al., 2021, Bardes et al., 2022].

On the other hand, JEPAs are based on encoder-predictor architectures,where a predictor asymmetry is introduced in one of the branches. The representation learning task is framed as predicting the consequence of an action or data transformation applied to the first view resulting in the second view. However, if the predictor is not conditioned on the action/transformation, one can only hope to obtain transformation-invariant representations. Indeed, originally, this was the goal of JEPA models without action conditioning, such as BYOL [Grill et al., 2020] and SimSiam [Chen et al., 2020]. Action-conditioned world modeling is common in reinforcement learning (RL) [Ha and Schmidhuber, 2018, Hafner et al., 2019]. In generative SSL, e.g. masked autoencoders [He et al., 2022], decoder can be considered a world model conditioned on mask locations. In JEPAs, output embeddings of the encoder can be conditioned on mask locations as in I-JEPA [Assran et al., 2023], or further on data transformations as in image world models (IWM) [Garrido et al., 2024] before the predictor. IWM showed a sufficiently deep action-conditioned predictor can be used to learn transformation-equivariant representations; a property that is crucial in many fine-grained and low-level downstream tasks such as semantic segmentation. For example, a fine-grained task of distinguishing different species of birds may be impossible to solve if the model is invariant to color. Most architectures designed for equivariant SSL [Devillers and Lefort, 2022, Park et al., 2022, Garrido et al., 2023, Gupta et al., 2023, 2024, Dangovski et al., 2022] encourage both invariant and equivariant features in the same representational space, i.e. backbone encoder's output, which may result in a trade-off between invariance and equivariance-related task performance [Garrido et al., 2024].

Most of the SSL models discussed above operate by comparing only two views of an image, overlooking the potential advantages of integrating multiple views to learn both action-invariant and action-equivariant representations. Models of object learning in animals have emphasized the role of sequential actions and observations for both object representation learning [Tarr et al., 1998, Harman et al., 1999, Vuilleumier et al., 2002] and planning [Rao, 2024]. To adapt JEPAs to take advantage of such action-observation sequences, we propose seq-JEPA in which the usual predictor world model is preceded by an autoregressive memory module to aggregate the latent representations of recent observations, a mechanism that can be loosely compared with working memory in humans [Baddeley, 2003]. We show that there is a positive correlation between the number of available observations (length of the input sequence) and its downstream performance in both invariance and equivariance-related tasks. To summarize, our contributions are as follows:

- We propose seq-JEPA, an SSL method with a memory-enhanced autoregressive predictive world model. Our model is able to effectively aggregate recent action-conditioned observations to achieve competitive performance on invariance- and equivariance-related tasks on 3DIEBench and STL-10.
- We show that seq-JEPA separates invariant and equivariant representations, avoiding the potential trade-off between invariance- and equivariance-related performances which can happen when using a single representational space for both types of tasks.
- Utilizing seq-JEPA's structure, we propose predictive learning across saccades, a method to learn image representations inspired by embodied vision. To do so, we aggregate a sequence of low-resolution saccadic patches sampled from image saliency maps with no hand-crafted data augmentations.

2 Related Work

Self-supervised predictive architectures. In addition to JEPAs (Fig. 2.a) discussed in previous section, another line of predictive SSL architectures predict the representation of the next observation in a sequence yet with a contrastive objective and using negative samples [van den Oord et al., 2019, Gupta et al., 2024, Schneider et al., 2021, Aubret et al., 2024] (see Fig. 2.b). Predictive learning in latent space is also used to train world models in RL for better sample efficiency [Schwarzer et al., 2021] or to create an auxiliary intrinsic reward function [Sekar et al., 2020, Guo et al., 2022]. Among these models, architecturally, BYOL-Explore [Guo et al., 2022] would be closest to seq-JEPA.

Equivariant representation learning. To learn equivariant representations in SSL, one way is to modify the type of transformations used to tailor to a downstream task [Xiao et al., 2021, Dangovski et al., 2022]. Another way is to use an equivariance predictor to predict the effect of a transformation in the embedding space by conditioning the predictor on transformation parameters [Devillers and Lefort, 2022, Park et al., 2022, Garrido et al., 2023, Gupta et al., 2023, 2024]. Most of the former methods add an auxiliary predictor with an additional equivariance loss to avoid collapse to invariance. Action-conditioned JEPAs [Garrido et al., 2023] have a single prediction loss but suffer a trade-off between invariance- and equivariance-related performance based on the size of the predictor, i.e., a larger predictor would result in more equivariant world models but with a drop in invariant linear probe performance [Garrido et al., 2024].

3 Method

Architecture. We now describe the learning procedure of seq-JEPA (Fig. 1c). Assuming a sequence of actions or transformations $\{a_i\}_{i=1}^{M+1}$, possibly generated from a policy that can also be learnable, a set of sequential observations $\{x_i\}_{i=1}^{M+1}$ are produced (denote Δ_{a_{i+1},a_i} as the relative action/transformation that transforms x_i to x_{i+1}). A backbone encoder f_θ (we use a ResNet-18) encodes these observations to output embeddings $\{z_i\}_{i=1}^{M}$. At this stage, the first M embeddings are concatenated by learnable relative action embeddings, $\Delta_{\hat{a}_{i+1},\hat{a}_i}$. These concatenated embeddings are then fed as tokens to the autoregressive memory module g_ϕ (we use a three-layer transformer encoder [Vaswani et al., 2017]). This module also receives a learnable token called `[AGG]` (similar to the `[CLS]` token in supervised transformers). The output embedding corresponding to `[AGG]` is $z_{AGG} = g_\phi((z_1, \Delta_{\hat{a}_2,\hat{a}_1}), (z_2, \Delta_{\hat{a}_3,\hat{a}_2})..., ((z_M, 0))$ and serves as the global representation of the observation sequence. The vector z_{AGG} is then concatenated with the action embedding $\hat{a}_{M+1}$ and is fed to a predictor MLP module h_ψ which outputs the prediction of the subsequent observation embedding $\hat{z}_{M+1} = h_\psi((z_{AGG}, \Delta_{\hat{a}_{M+1},\hat{a}_M}))$. The loss to be minimized is simply based on the negative cosine similarity between the original observation embedding (passed through a stop-gradient (`sg`) to avoid representational collapse) and the predicted embedding,

$$\mathcal{L}_{seq-JEPA} = 1 - \frac{\hat{z}_{M+1}}{\|\hat{z}_{M+1}\|_2} \cdot \frac{\mathtt{sg}(z_{M+1})}{\|\mathtt{sg}(z_{M+1})\|_2}. \tag{1}$$

Actions and Observations. We experiment with three different modes of action-observation pairs for seq-JEPA. In the first mode, observations are transformed image views using hand-crafted augmentations typically used in augmentation-invariant SSL with action being encoded transformation parameters (Fig. 2.d). In the second mode, observations are rendering images from the 3D Invariant Equivariant Benchmark (3DIEBench) [Garrido et al., 2023] with actions being the object rotation parameters (Fig. 2.c). In addition to these two settings, we also propose predictive learning across saccades (PLS) (Fig. 2.a) to learn representations without hand-crafted augmentations with a similar flavor to I-JEPA [Assran et al., 2023] but with differences in terms of architecture and view generation. PLS is enabled by our use of an autoregressive memory module and is not dependent on the architecture of the encoder (e.g. masking tokens of a vision transformer in I-JEPA). PLS takes advantage of small cheap-to-process saccadic patches (glances); specifically, we adopt two principles from embodied vision, i.e., saliency maps [Itti et al., 1998, Li, 2002, Zhaoping, 2014], and inhibition of return (IoR) [Posner et al., 1985] to do away with the constraints of I-JEPA context and target creation. Inspired by the V1 Saliency Hypothesis (V1SH) [Li, 2002, Zhaoping, 2014], we assume that the agent possesses a (pre-defined or learnable) internal saliency map of its visual field to actively guide its visual attention. We sample $M + 1$ fixations (actions) from the saliency map policy with IoR (to avoid overlapping saccades), and feed seq-JEPA with the first M glances along with relative action embeddings to predict the next glance representation. To generate saliency maps we use the pre-trained DeepGaze IIE [Linardos et al., 2021]. See Appendix for details.

4 Experiments and Discussion

Augmentation-based STL-10 and 3DIEBench. For first and second mode of action-observation discussed above, we use STL-10 and 3DIEBench respectively and compare both invariant and equivariant performance of seq-JEPA with existing invariant (SimCLR [Chen et al., 2020], VICReg [Bardes et al., 2022], and SimSiam [Chen and He, 2021]) and equivariant (SIE [Garrido et al., 2024], SEN [Park et al., 2022], and EquiMod [Devillers and Lefort, 2022]) SSL methods. Top-1 linear probe accuracy on frozen representations is reported as the invariance-related performance metric. As

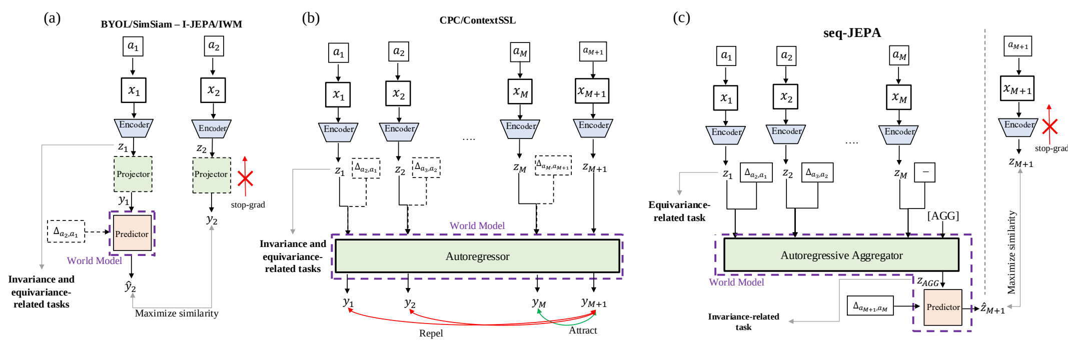

Figure 1: Different approaches to world modeling in SSL; (a) JEPAs predict the effect of a single transformation in latent space, (b) Contrastive world models such as Contrastive Predictive Coding and ContextSSL predict the next part of an input sequence via contrastive learning, (c) Seq-JEPA's world model aggregates a sequence of observations and conditions the aggregated output to predict the representation of next observation.

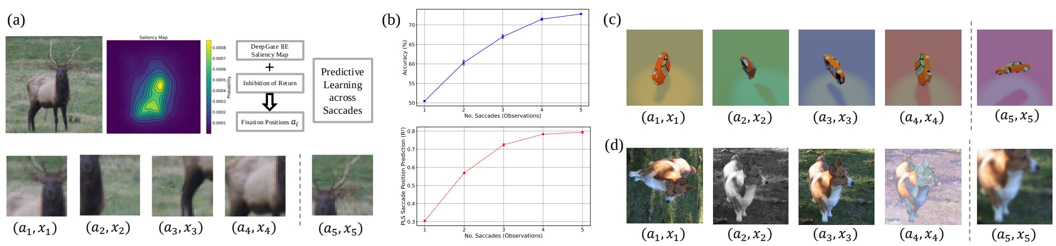

Figure 2: (a) Predictive learning across saccades (PLS) for learning image representations (b) Both invariant and equivariant performance in PLS increases with the number of available observations to seq-JEPA (c-d) Geometric transformations and hand-crafted augmentations as other modes of action-observation for seq-JEPA

the equivariance metric, we report R^2 of a regression head trained to predict relative transformation parameters between a pair of observations with their concatenated representations as regressor's input. For seq-JEPA, we used the output of either `[AGG]` token or the ResNet for invariance, and the ResNet output for equivariance tasks. All other methods use the ResNet output for both tasks. Tab. 1 shows these metrics for both STL-10 and 3DIEBench. As can be seen, seq-JEPA achieves competitive invariant and equivariant performance in both settings without sacrificing one for the other. Although seq-JEPA does not achieve the best performance in all settings, it shows a high level of invariance among equivariant models, and a high level of equivariance among invariant models. Most interestingly, two different locations in the network, i.e., `AGG` and ResNet outputs, have taken specialized roles: the `[AGG]` output is specialized for the invariance tasks while the ResNet output is equivariant.

Predictive Learning across Saccades. In PLS, we report performance with different number of saccades (Fig. 2.b). Equivariance metric is the R^2 of a regressor that predicts the relative position of two patches in the image given their concatenated representations. As can be seen, both invariant and equivariant performances increase with the number of available saccadic observations which indicates the ability of seq-JEPA's world model to integrate more information to improve its performance. See Appendix for PLS ablations on saliency map, IoR, and action conditioning.

Limitations and Future Work. The current version of seq-JEPA has only been tested on static image datasets and perception tasks that do not require planning. Future works include larger-scale experiments, planning tasks with intrinsic motivation or combined with RL, and integrating multi-modality into our framework.

Acknowledgments and Disclosure of Funding

This project was supported by funding from NSERC (Discovery Grants RGPIN-2022-05033 to E.B.M, and RGPIN-2023-03875 to S.B.), the Canada CIFAR AI Chairs Program, the Canada Excellence Research Chairs (CERC) Program, the Institute for Data Valorization (IVADO), the CHU Sainte-

Table 1: Performance metrics for 3DIEBench and STL-10

3DIEBench			
Conditioning	**Method**	**Rotation Pred. (R^2)**	**top-1 Acc.**
-	SimCLR	0.34	78.39
	VICReg	0.14	76.13
	BYOL	0.04	77.32
Rotation	SIE	**0.67**	77.06
	SEN	0.50	83.72
	EquiMod	0.53	**84.90**
	Seq-JEPA (Ours)	0.56	79.64 (`[AGG]`), 70.09 (**ResNet**)
STL-10 (Augmentations)			
Conditioning	**Method**	**Transform Pred. (R^2)**	**top-1 Acc.**
-	SimCLR	0.07	**79.81**
	VICReg	0.04	77.12
	BYOL	0.09	78.21
Transformation	SIE	0.11	75.88
	SEN	0.06	77.91
	EquiMod	0.08	78.40
	Seq-JEPA (Ours)	**0.13**	79.12 (`[AGG]`), 75.26 (**ResNet**)

Justine Research Center (CHUSJRC), Fonds de Recherche du Québec–Santé (FRQS), the Quebec Institute for Artificial Intelligence (Mila), and Google. This research was also supported in part by Calcul Québec (calculquebec.ca) and the Digital Research Alliance of Canada (alliancecan.ca).

References

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Michael Rabbat, and Nicolas Ballas. Masked Siamese Networks for Label-Efficient Learning, April 2022. URL `http://arxiv.org/abs/2204.07141`. arXiv:2204.07141 [cs, eess].

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15619–15629, Vancouver, BC, Canada, June 2023. IEEE. ISBN 9798350301298. doi: 10.1109/CVPR52729.2023.01499. URL `https://ieeexplore.ieee.org/document/10205476/`.

Arthur Aubret, Céline Teulière, and Jochen Triesch. Self-supervised visual learning from interactions with objects, July 2024. URL `http://arxiv.org/abs/2407.06704`. arXiv:2407.06704 [cs].

Alan Baddeley. Working memory: looking back and looking forward. *Nature Reviews Neuroscience*, 4(10):829–839, October 2003. ISSN 1471-0048. doi: 10.1038/nrn1201. URL `https://www.nature.com/articles/nrn1201`. Publisher: Nature Publishing Group.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language. In *Proceedings of the 39th International Conference on Machine Learning*, pages 1298–1312. PMLR, June 2022. URL `https://proceedings.mlr.press/v162/baevski22a.html`. ISSN: 2640-3498.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=xm6YD62D1Ub`.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/hash/70feb62b69f16e0238f741fab228fec2-Abstract.html`.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Caron_Emerging_Properties_in_Self-Supervised_Vision_Transformers_ICCV_2021_paper.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607. PMLR, November 2020. URL https://proceedings.mlr.press/v119/chen20j.html. ISSN: 2640-3498.

Xinlei Chen and Kaiming He. Exploring Simple Siamese Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Chen_Exploring_Simple_Siamese_Representation_Learning_CVPR_2021_paper.html.

Rumen Dangovski, Li Jing, Charlotte Loh, Seungwook Han, Akash Srivastava, Brian Cheung, Pulkit Agrawal, and Marin Soljacic. Equivariant Self-Supervised Learning: Encouraging Equivariance in Representations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=gKLAAfiytI.

Alexandre Devillers and Mathieu Lefort. EquiMod: An Equivariance Module to Improve Visual Instance Discrimination. In *The Eleventh International Conference on Learning Representations*, September 2022. URL https://openreview.net/forum?id=eDLwjKmtYFt.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a Little Help From My Friends: Nearest-Neighbor Contrastive Learning of Visual Representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9588–9597, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Dwibedi_With_a_Little_Help_From_My_Friends_Nearest-Neighbor_Contrastive_Learning_ICCV_2021_paper.html.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for Self-Supervised Representation Learning. In *Proceedings of the 38th International Conference on Machine Learning*, pages 3015–3024. PMLR, July 2021. URL https://proceedings.mlr.press/v139/ermolov21a.html. ISSN: 2640-3498.

Quentin Garrido, Laurent Najman, and Yann Lecun. Self-supervised learning of Split Invariant Equivariant representations. In *Proceedings of the 40th International Conference on Machine Learning*, pages 10975–10996. PMLR, July 2023. URL https://proceedings.mlr.press/v202/garrido23b.html. ISSN: 2640-3498.

Quentin Garrido, Mahmoud Assran, Nicolas Ballas, Adrien Bardes, Laurent Najman, and Yann LeCun. Learning and Leveraging World Models in Visual Representation Learning, March 2024. URL http://arxiv.org/abs/2403.00504. arXiv:2403.00504 [cs].

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 *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html.

Zhaohan Daniel Guo, Shantanu Thakoor, Miruna Pîslar, Bernardo Avila Pires, Florent Altché, Corentin Tallec, Alaa Saade, Daniele Calandriello, Jean-Bastien Grill, Yunhao Tang, Michal Valko, Rémi Munos, Mohammad Gheshlaghi Azar, and Bilal Piot. BYOL-Explore: Exploration by Bootstrapped Prediction, June 2022. URL http://arxiv.org/abs/2206.08332. arXiv:2206.08332 [cs, stat].

Sharut Gupta, Joshua Robinson, Derek Lim, Soledad Villar, and Stefanie Jegelka. Structuring Representation Geometry with Rotationally Equivariant Contrastive Learning. In *The Twelfth International Conference on Learning Representations*, October 2023. URL https://openreview.net/forum?id=lgaFMvZHSJ.

Sharut Gupta, Chenyu Wang, Yifei Wang, Tommi Jaakkola, and Stefanie Jegelka. In-Context Symmetries: Self-Supervised Learning through Contextual World Models, May 2024. URL http://arxiv.org/abs/2405.18193. arXiv:2405.18193 [cs].

David Ha and Jürgen Schmidhuber. Recurrent World Models Facilitate Policy Evolution. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/2de5d16682c3c35007e4e92982f1a2ba-Abstract.html.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to Control: Learning Behaviors by Latent Imagination. In *International Conference on Learning Representations*, September 2019. URL https://openreview.net/forum?id=S1lOTC4tDS.

Jeff Z. HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable Guarantees for Self-Supervised Deep Learning with Spectral Contrastive Loss. In *Advances in Neural Information Processing Systems*, volume 34, pages 5000–5011. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/27debb435021eb68b3965290b5e24c49-Abstract.html.

Karin L. Harman, G.Keith Humphrey, and Melvyn A. Goodale. Active manual control of object views facilitates visual recognition. *Current Biology*, 9(22):1315–1318, November 1999. ISSN 09609822. doi: 10.1016/S0960-9822(00)80053-6. URL https://linkinghub.elsevier.com/retrieve/pii/S0960982200800536.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, Seattle, WA, USA, June 2020. IEEE. ISBN 978-1-72817-168-5. doi: 10.1109/CVPR42600.2020.00975. URL https://ieeexplore.ieee.org/document/9157636/.

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*, pages 16000–16009, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/He_Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper.

L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, November 1998. ISSN 1939-3539. doi: 10.1109/34.730558. URL https://ieeexplore.ieee.org/document/730558/?arnumber=730558. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

Zhaoping Li. A saliency map in primary visual cortex. *Trends in Cognitive Sciences*, 6(1):9–16, January 2002. ISSN 1364-6613, 1879-307X. doi: 10.1016/S1364-6613(00)01817-9. URL https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(00)01817-9. Publisher: Elsevier.

Akis Linardos, Matthias Kummerer, Ori Press, and Matthias Bethge. DeepGaze IIE: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12899–12908, Montreal, QC, Canada, October 2021. IEEE. ISBN 978-1-66542-812-5. doi: 10.1109/ICCV48922.2021.01268. URL https://ieeexplore.ieee.org/document/9711473/.

Ishan Misra and Laurens van der Maaten. Self-Supervised Learning of Pretext-Invariant Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Misra_Self-Supervised_Learning_of_Pretext-Invariant_Representations_CVPR_2020_paper.html.

Jung Yeon Park, Ondrej Biza, Linfeng Zhao, Jan-Willem Van De Meent, and Robin Walters. Learning Symmetric Embeddings for Equivariant World Models. In *Proceedings of the 39th International Conference on Machine Learning*, pages 17372–17389. PMLR, June 2022. URL `https://proceedings.mlr.press/v162/park22a.html`. ISSN: 2640-3498.

Michael I Posner, Robert D Rafal, Lisa S Choate, and Jonathan Vaughan. Inhibition of return: Neural basis and function. *Cognitive neuropsychology*, 2(3):211–228, 1985.

Rajesh P. N. Rao. A sensory–motor theory of the neocortex. *Nature Neuroscience*, 27(7):1221–1235, July 2024. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-024-01673-9. URL `https://www.nature.com/articles/s41593-024-01673-9`.

Felix Schneider, Xia Xu, Markus R Ernst, Zhengyang Yu, and Jochen Triesch. Contrastive Learning Through Time. In *SVRHM Workshop@ NeurIPS.*, 2021.

Max Schwarzer, Ankesh Anand, Rishab Goel, R. Devon Hjelm, Aaron Courville, and Philip Bachman. Data-Efficient Reinforcement Learning with Self-Predictive Representations. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=uCQfPZwRaUu&fbclid=IwAR3FMvlynXXYEMJaJzPki1x1wC9jjA3aBDC_moWxrI91hLaDvtk7nnnIXT8`.

Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to Explore via Self-Supervised World Models. In *Proceedings of the 37th International Conference on Machine Learning*, pages 8583–8592. PMLR, November 2020. URL `https://proceedings.mlr.press/v119/sekar20a.html`. ISSN: 2640-3498.

Michael J. Tarr, Pepper Williams, William G. Hayward, and Isabel Gauthier. Three-dimensional object recognition is viewpoint dependent. *Nature Neuroscience*, 1(4):275–277, August 1998. ISSN 1546-1726. doi: 10.1038/1089. URL `https://www.nature.com/articles/nn0898_275`. Publisher: Nature Publishing Group.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding, January 2019. URL `http://arxiv.org/abs/1807.03748`. arXiv:1807.03748 [cs, stat].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html`.

P. Vuilleumier, R. N. Henson, J. Driver, and R. J. Dolan. Multiple levels of visual object constancy revealed by event-related fMRI of repetition priming. *Nature Neuroscience*, 5(5):491–499, May 2002. ISSN 1546-1726. doi: 10.1038/nn839. URL `https://www.nature.com/articles/nn839`. Publisher: Nature Publishing Group.

Tete Xiao, Xiaolong Wang, Alexei A. Efros, and Trevor Darrell. What Should Not Be Contrastive in Contrastive Learning. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=CZ8Y3NzuVzO`.

Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. Decoupled Contrastive Learning. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 668–684, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19809-0. doi: 10.1007/978-3-031-19809-0_38.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In *Proceedings of the 38th International Conference on Machine Learning*, pages 12310–12320. PMLR, July 2021. URL `https://proceedings.mlr.press/v139/zbontar21a.html`. ISSN: 2640-3498.

Li Zhaoping. The V1 hypothesis—creating a bottom-up saliency map for preattentive selection and segmentation. In *Understanding Vision*, pages 189–314. Oxford University PressOxford, 1 edition, May 2014. ISBN 978-0-19-956466-8 978-0-19-177250-4. doi: 10.1093/acprof:oso/9780199564668.003.0005. URL `https://academic.oup.com/book/8719/chapter/154784147`.

A Details and ablations for predictive learning across saccades.

To perform predictive learning across saccades, given an image dataset, we use DeepGaze IIE Linardos et al. [2021], a model trained on human visual fixations, to generate the saliency maps corresponding to the images. Specifically, DeepGaze receives an input image and outputs a negative log-likelihood saliency map which can be converted to a probability distribution (policy) over the pixels. This distribution can be interpreted as how likely humans are to center their gaze on a given pixel. We extract and add these saliency maps as a fourth channel to images and store them as a dataset to speed up training. To simulate a fixation, we sample a pixel from this policy and crop a square patch centered on this pixel from the image with a width equal to third of the original image, e.g. 32 for STL-10 square images. Furthermore, to implement IoR, which helps minimize the overlap between subsequent glances and information redundancy in the prediction objective, we set the probabilities of a circular area with a diameter equal to patch width surrounding the center of the saccade to zero before sampling the subsequent saccade. To generate the PLS observations, we sample $M + 1$ fixations (actions) from the saliency map policy with IoR, shuffle the fixations, and feed the corresponding patches to the network. In Table 2, the results of ablation experiments for the PLS setting are presented. It can be seen that removing each component in PLS results in a performance drop.

Table 2: Predictive learning across saccades ablations on STL-10

Method	Classification (top-1)	Position prediction (R^2)
seq-JEPA ($M = 5$)	72.81	0.794
w/o action conditioning	71.64	0.14
w/o IoR	68.51	0.761
w/o saliency map (uniform distribution)	68.12	0.748

B Implementation details.

We train all methods for 1000 epochs using a batch size of 512 with the AdamW optimizer, a learning rate of 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of 0.01. We use a cosine schedule to decay the learning rate to 10^{-6} following a linear warmup for 20 epochs starting from a learning rate of 10^{-4}. To apply augmentations for STL-10, we follow the protocol of EquiMod [Devillers and Lefort, 2022] and their method to encode the transformation parameters. Each experiment was conducted on 1 NVIDIA A100 GPU with 40GB of accelerator RAM. A single run of seq-JEPA on 3DIEBench with a sequence length of 5 takes around 3 hours.

Evaluation protocol. For linear probing, we follow the SSL protocol and train a linear classifier on top of frozen representations with a batch size of 256 and using the Adam optimizer with default hyperparameters. For action prediction, we follow the protocol in SIE [Garrido et al., 2023] and train an MLP regressor with intermediate dimensions of 1024-1024-d, where d is the size of the action vector (4 for quaternion rotations, 2 for relative saccade position, and 18 for augmentation parameters)

Below, we describe the architectural details and hyperparameters of each method used in the paper:

Seq-JEPA. In our default setting, we use a ResNet-18 (not pre-trained) as the encoder. The relative action/transformation parameters are passed from a learnable linear projector of size 32. We use a transformer encoder architecture [Vaswani et al., 2017] with three layers and four attention heads as our autoregressor. The predictor MLP has intermediate dimensions of 512-512.

SimCLR. We use a temperature parameter of $\tau = 0.5$ with a projection MLP with 2048-2048-2048 intermediate dimensions.

VICReg. We use $\lambda_{\text{inv}} = \lambda_V = 10$, $\lambda_C = 1$, and a projection head of 2048-2048-2048 intermediate dimensions.

BYOL. We use a a projection head of 2048-2048-2048 intermediate dimensions. The predictor has intermediate dimensions of 512-512. For EMA, we use a starting momentum of 0.996 and increase it to one through training.

SIE. For both invariant and equivariant projection heads, we use intermediate dimensions of 1024-1024-1024. For the loss coefficients, we use $\lambda_{\text{inv}} = \lambda_V = 10$, $\lambda_{\text{equi}} = 4.5$, and $\lambda_C = 1$.

SEN. We use a temperature parameter of $\tau = 0.5$ with a projection MLP with 2048-2048-2048 intermediate dimensions.

EquiMod. We use the version based on SimCLR (both invariant and equivariant losses are contrastive with $\tau = 0.1$ and have equal weights). The projection head has 1024-1024-128 intermediate dimensions. The transformation parameters are normalized and passed through a linear projection layer of size 128.