Pretrained Encoders are All You Need

Anonymous Authors[1]

Abstract

Data-efficiency and generalization are key challenges in deep learning and deep reinforcement learning as many models are trained on large-scale, domain-specific, and expensive-to-label datasets. Self-supervised models trained on large-scale uncurated datasets have shown successful transfer to diverse settings. We investigate using pretrained image representations and spatio-temporal attention for state representation learning in Atari. We also explore fine-tuning pretrained representations with self-supervised techniques, i.e., contrastive predictive coding, spatio-temporal contrastive learning, and augmentations. Our results show that pretrained representations are at par with state-of-the-art self-supervised methods trained on domain-specific data. Pretrained representations, thus, yield data and compute-efficient state representations.

1. Introduction

Data-efficiency and generalization are key challenges in deep learning (DL) and deep reinforcement learning (RL), especially for real-world deployment and scalability. Self-supervised learning (SSL) has been used in computer vision and natural language processing (Chen et al., 2020c; Radford et al., 2021; Henaff, 2020; He et al., 2020; Devlin et al., 2019; Radford et al.) to learn from large-scale unlabeled/uncurated data and do a few-shot transfer to labeled data. Deep RL has also leveraged self-supervised learning for data-efficiency (Srinivas et al., 2020; Laskin et al., 2020).

RL in pixel space is also sample-inefficient, and state representations can help sample-efficient, robust, and generalizable RL (Lake et al., 2016; Kaiser et al., 2019; Tassa et al., 2018). Previous work investigated self-supervised state representation learning in RL, but involved training on large-scale domain-specific data (Anand et al., 2019).

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

We investigate how pre-trained models can be leveraged for state representation learning in Atari. In particular, we focus on using pretrained image representation learning and attention models. We also investigate self-supervised fine-tuning of pretrained representations using state-of-the-art (SotA) self-supervised methods.

Our results show that our pretrained representations, particularly 'zoomed in' representations, perform as well as the SotA self-supervised state representation learning models trained on large domain-specific datasets. Adding pretrained temporal attention does not particularly improve performance, possibly because the pretrained image representations already leverage attention. Adding self-supervised fine-tuning on domain-specific also does not particularly improve performance, especially in a data-constrained setting.

We make 3 key contributions: i. a new methodology for using pretrained image representations for sample-efficient and generalizable state representation learning in RL; ii. evaluations of using pretrained temporal attention models with static image representation for temporal data; iii. evaluating self-supervised fine-tuning of pretrained representations using SotA SSL techniques on domain-specific data.

2. Related Work

Our work lies at the intersection of 5 key areas: Self-Supervised Learning (SSL), Self-Supervised RL, Attention, Domain generalization, and Pretraining. Unlike previous work in Self-Supervised RL, Domain generalization, and Pretraining, we focus on state representation learning in RL, not on the RL. Unlike previous work in state representation learning for RL (Anand et al., 2019), we leverage pretrained models, not trained on domain-specific data, demonstrating data-efficient and generalizable learning. Finally, like previous work, we leverage spatial-temporal attention, particularly optical flow (Yuezhang et al., 2018), and also use SotA SSL, but only for fine-tuning pretrained representations. Our results show that good pretrained embeddings perform competitively, without necessarily needing augmentations from spatio-temporal attention or self-supervised fine-tuning.

Self-Supervised Learning (SSL): SSL has been used in natural language processing and computer vision (Devlin et al., 2018; Henaff, 2020; He et al., 2020; Chen et al., 2020a) using both contrastive ((Oord et al., 2018; Bachman

et al., 2019; He et al., 2020; Chen et al., 2020b; Radford et al., 2021) and purely predictive methods (Grill et al., 2020). SSL on large-scale uncurated datasets has shown promising few-shot transfer to diverse labeled data (Chen et al., 2020c; Radford et al., 2021). We use pretrained self-supervised models for sample-efficient state representation learning in RL. Also, we use self-supervised fine-tuning of pretrained representations on domain-specific data.

Self-Supervised RL: SSL has been used in model-based (Kaiser et al., 2019; Ha & Schmidhuber, 2018; Schrittwieser et al., 2020; Hafner et al., 2019) and model-free RL (Jaderberg et al., 2016; Shelhamer et al., 2016; Oord et al., 2018). Learning is either reconstruction-based (Jaderberg et al., 2016; Higgins et al., 2017; Yarats et al., 2019) or constrastive (Sermanet et al., 2018; Warde-Farley et al., 2018; Anand et al., 2019; Oord et al., 2018; Srinivas et al., 2020), and the predictions are in pixel space (Jaderberg et al., 2016) or latent space (Oord et al., 2018; Guo et al., 2020). Image augmentations (Srinivas et al., 2020; Laskin et al., 2020), spatio-temporal structures (Sermanet et al., 2018; Aytar et al., 2018; Oord et al., 2018; Anand et al., 2019), and scene or object representations (Burgess et al., 2019; Zhu et al., 2018; Greff et al., 2019; van Steenkiste et al., 2019) have also been used. Representation learning has also been decoupled from RL (Anand et al., 2019; Stooke et al., 2020). We focus on state representation learning in RL, instead of the RL itself, and unlike previous work, leverage pretrained self-supervised representations not trained on domain-specific data.

Attention: Attention has been used in RL (Zhang et al., 2018; Gregor et al., 2018; Manchin et al., 2019; Salter et al., 2019), especially to add robustness and interpretability (Sorokin et al., 2015; Mott et al., 2019). Optical flow-based attention has also been used in RL (Yuezhang et al., 2018). We explore spatio-temporal attention, including optical flow, to improve image representations for temporal data.

Domain generalization: Transfer learning uses representations from one domain to generalize to different domains (Rusu et al., 2016; Oquab et al., 2014; Donahue et al., 2013). Domain adaptation adapts from one domain to another domain using data from the target domain (Bousmalis et al., 2016; Ganin et al., 2016; Wulfmeier et al., 2017), whereas domain randomisation covers a distribution of environments during training to generalize (Sadeghi & Levine, 2016; Andrychowicz et al., 2020; Viereck et al., 2017; Held et al., 2017; Tobin et al., 2017; Peng et al., 2018). We use pretrained models, not trained on Atari-specific data, for sample-efficient state representation learning in RL.

Pretraining: Unsupervised pretraining has been leveraged in RL, e.g., by maximizing diversity of states (Liu & Abbeel, 2021) or skills (Eysenbach et al., 2018; Hansen et al., 2019; Sharma et al., 2019). Intrinsic rewards for exploration (Pathak et al., 2016; Sekar et al., 2020), predicting state dynamics (Anderson et al., 2015), and reward-free representations (Schwarzer et al.) have also been used. Transferable skills can also be learned (Campos et al., 2021). We use pretrained models for state representation learning in RL.

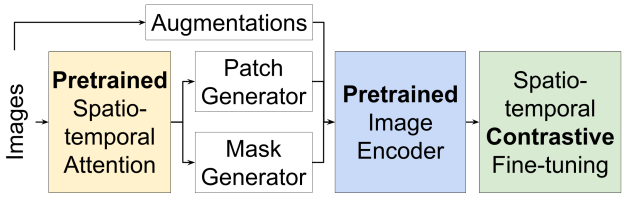

Figure 1. PEARL: Pretrained Encoder & Attention for Representation Learning.

3. Approach

State representation learning in RL: Deep reinforcement learning has leveraged the expressive power of deep learning to create end-to-end models for RL in high dimensional spaces. RL in high dimensional spaces, e.g., pixels spaces, however, is sample-inefficient (Lake et al., 2016; Kaiser et al., 2019) and learning policies from state representations could be more **sample-efficient, robust, and generalizable** (Tassa et al., 2018; Liu et al., 2021; Eslami et al., 2018). Thus, we decouple state representation learning from RL (Stooke et al., 2020) and aimed to investigate state representation learning in RL (Anand et al., 2019).

Pretrained models with self-supervised fine-tuning: Pretrained models have been leveraged for few-shot learning (Chen et al., 2020c; Henaff, 2020; Radford et al., 2021), but not for state representation learning in RL. State representation learning in RL has leveraged SSL (Anand et al., 2019), but involved domain-specific and data-intensive training. We aimed to leverage pretrained models for sample-efficient state representation learning in RL. We decided on 3 key explorations: *i. Pretrained image representations*, including 'zooming in' via grid-based patches; *ii. Spatio-temporal attention*, e.g., using optical flow; *iii. Self-supervised fine-tuning* on domain-specific data via constrastive losses.

Our framework, **PEARL** (Pretrained Encoder and Attention for Representation Learning), has 3 components (Figure 1).

3.1. Pretrained Image Representations

Recent work compared image representations from different supervised, self-supervised, and weakly supervised models, and found that CLIP (Radford et al., 2021), a weakly supervised image representation learning model, produced the best representations for few-shot learning (Khan et al., 2021). Thus, we selected CLIP as our image encoder.

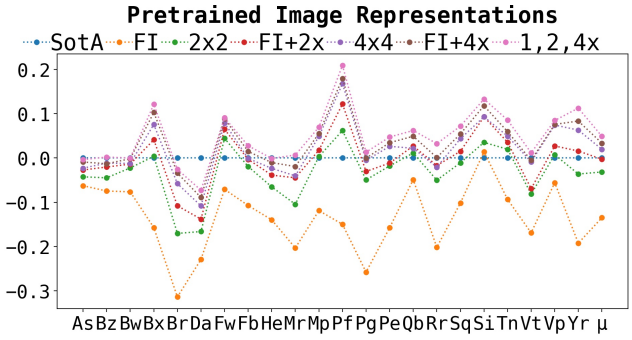

Figure 2. SotA vs pretrained representations for patches (vanilla).

In addition to image representations for full-image, we also considered 'zooming in' via grid-based patches, i.e., equally-sized non-overlapping patches covering the whole image.

3.2. Spatio-temporal Attention

For spatial attention, we considered a SotA supervised object detection model, i.e., EfficientDet (Tan et al., 2019), and a SotA SSL model, i.e., Dino (Caron et al., 2021).

For temporal attention, we decided to compare optical flow using RAFT, a SotA pretrained model (Teed & Deng, 2020), with image difference using structural similarity (Wang et al., 2004). We used image difference particularly because objects appear/disappear in video animations and optical flow may not be able to capture sudden flowless changes.

Finally, we compare two types of attention: i. mask-based attention, which highlights the relevant regions in the full image; ii. patch-based attention, which crops out the relevant patches and calculates embeddings for each by zooming in.

3.3. Self-supervised Fine-tuning

We considered 3 SotA methods for self-supervised fine-tuning of pretrained representations using domain-specific data: i. Image augmentations (color jitter, random crop, and gaussian blur) (Chen et al., 2020b); ii. Spatio-temporal constrastive learning (ST-DIM) (Anand et al., 2019); iii. Contrastive predictive learning (CPC) (Oord et al., 2018).

4. Experiments

We outline our experiments and findings below. Similar to the SotA work (Anand et al., 2019), our evaluations use a linear probe (Alain & Bengio, 2018) with a {70, 10, 20}% {train, validation, and test}-split with 50k data points and early stopping. We share the F1 score for each Atari game.

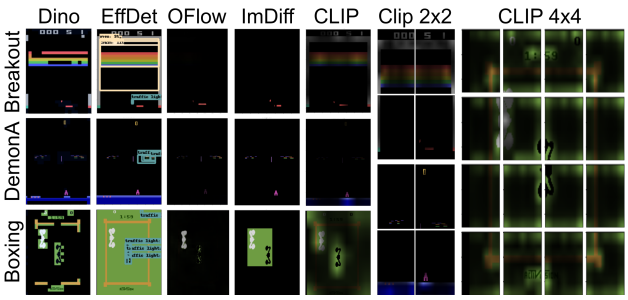

Figure 3. Spatio-temporal attention masks from pretrained models.

4.1. Pretrained Image Representations

Setup: We tried 3 nxn-grid patch sizes (1x1, 2x2, and 4x4). Each patch generated a 512-size embedding using CLIP and we concatenated the embeddings. We tried 6 different configurations: 1. 1x1, i.e., full image (FI); 2. 2x2; 3. 1x1+2x2; 4. 4x4; 5. 1x1+4x4; 6. 1x1+2x2+4x4. Figure 2 compares the results for our 6 configurations to the SotA results. SotA refers to the top performance of all the models in (Anand et al., 2019) – ST-DIM was top in all, except 4 games, CPC was top in 3, and Pixel-Pred was top in 4.

Results: On average, using at least 5 embeddings from 1 full image and 4 2x2 patches, CLIP's pretrained embeddings perform better than SotA. Even our worst configuration (1x1) is on average better than a VAE trained on domain-specific data (Anand et al., 2019). With 21 patches (1x+2x+4x), our model performs better than the SotA in all, except 2 games, and has an average of 5% better performance than SotA. Overall, the performance improves with increasing number of patches. Thus, zooming in gives performance improvements at the cost of bigger embedding sizes.

4.2. Spatial-Temporal Attention

Setup: We compared Optical Flow Mask (FM) with Image Difference Mask (DM), and also, each mask combined with full image, i.e., FM+ (FM+FI) and DM+ (DM+FI). We evaluated patch-wise attention using full image combined with 4 patches selected from 4x4 and 2x2 patches, weighted by Optical Flow (FP5) and Image Difference (DP5). Each image 'combination' refers to a concatenation of the image embeddings. We compared each of our 3 settings with their equivalent size vanilla embeddings, i.e., embeddings using just the pretrained encoder (Section 4.1): i. DM and FM with FI (1x512 embedding); ii. DM+ and FM+ with FI+FI (2x512 embedding); iii. DP5 and FP5 with FI+2x2 (5x512 embedding). We did not evaluate EfficientDet and Dino as the attention masks did not look promising (Figure 3).

Results: Our results (Figure 4) show that optical flow masks (FM) alone are on average better than image difference

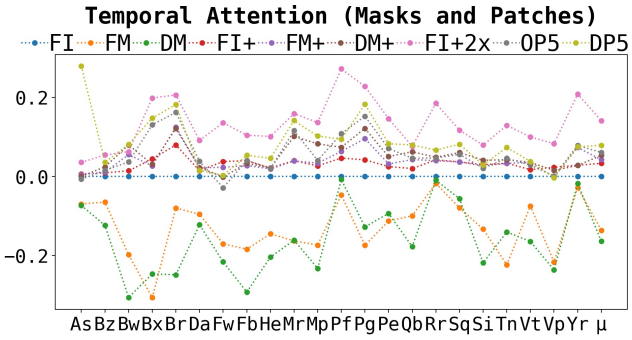

Figure 4. Temporal attention masks and patches vs vanilla results.

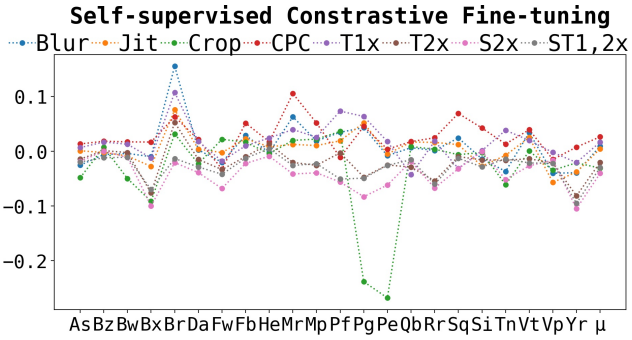

Figure 5. Performance changes using self-supervised fine-tuning

(DM) masks. However, image difference masks combined with the full image (DM+) are better than optical flow masks combined with the full image (FM+). Also, both FM+ and DM+ are both slightly better than the two full image embeddings concatenated (FI+), and the difference is biggest in games like Pong and Montezuma's Revenge. Finally, full image + 4 non-grid patches from image difference (DP5) are slightly better than full image + 4 non-grid patches from optical flow (FP5), but they are both worse than the full image + 4 2x2 grid patches. Thus, n grid-based patches may be better than n non-grid patches, which may zoom in more but not cover the full image. Also, temporal masks combined with full image are slightly better than two copies of the full image. However, the difference is not significant and is game-dependent, possibly because our encoder model, CLIP, already incorporates 'good enough' attention (Fig 3).

4.3. Self-supervised Fine-tuning

We evaluated if state-of-the-art self-supervised methods could be used to fine-tune pretrained embeddings using domain-specific data. We evaluated three self-supervised methods: i. Augmentations (random crop, color jitter, and gaussian blur) with an MLP layer and pairwise contrastive loss (Chen et al., 2020c); ii. ST-DIM (Anand et al., 2019) with separate temporal (T-DIM), spatial (S-DIM), and spatio-temporal contrast (ST-DIM); iii. CPC, replacing CPC's original encoder with our CLIP encoder and then adding a linear layer followed by a Gated Recurrent Unit. We compare each of the evaluations with their equivalent embedding-size baselines from Section 4.1 – FI for Blur, Jit, Crop, CPC, and T-DIM-1x (T1x); 2x2 for T-DIM 2x2 (T2x) and S-DIM (s2x); 1x1+2x2 for ST-DIM 1x1+2x2 (ST1,2x).

Our results (Figure 5) show that on average, fine-tuning with CPC leads to the most improvement (3%), followed by T-DIM-1x1 and gaussian blur. Color jitter, on average, makes no change and the rest lead to a 2-4% drop in performance compared to the equivalent vanilla cases (Section 4.1). Maximum improvements are using gaussian blur in Breakout (15.5%), CPC in Montezuma's revenge (10.8%), and T-DIM in Breakout (6.3%), but most improvements are game and method-dependent. Thus, overall, there are no significant improvements using self-supervised fine-tuning, possibly due to data or model constraints and especially since we are not fine-tuning the pretrained CLIP model.

5. Conclusion

Data-efficiency and generalization are key challenges in deep learning and deep RL. Most deep RL models are trained from scratch on domain-specific data, but recent research shows that self-supervised models trained on large-scale uncurated data have promising few-shot transfer.

We investigated the use of pretrained models for state representation learning in RL. Our results show that pretrained self-supervised models, not trained on domain-specific data, give competitive performance compared to SotA self-supervised models, trained on large-scale domain-specific data. Thus, pretrained models enable data-efficient and generalizable state representation learning for RL.

Moreover, our framework, PEARL (Pretrained Encoder and Attention for Representation Learning), investigates not only using pretrained image representations but also pretrained spatio-temporal attention. Our pretrained image representation model also uses attention and our results show that attention helps state representation learning.

Finally, even though self-supervised fine-tuning on domain-specific data did not significantly improve pretrained representations, it could be because we froze our pretrained model and had a few trainable layers with limited data.

State representations are key in RL. We believe that 'Pretrained Encoders are All You Need' (PEAYN) for data-efficient and generalizable state representation learning in RL, and will hopefully be a stepping stone to data-efficient, generalizable, and interpretable RL. No PEAYN, no gain :)

References

Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes. *arXiv:1610.01644 [cs, stat]*, November 2018. URL http://arxiv.org/abs/1610.01644. arXiv: 1610.01644.

Anand, A., Racah, E., Ozair, S., Bengio, Y., Côté, M.-A., and Hjelm, R. D. Unsupervised state representation learning in atari. *arXiv preprint arXiv:1906.08226*, 2019.

Anderson, C. W., Lee, M., and Elliott, D. L. Faster reinforcement learning after pretraining deep networks to predict state dynamics. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. IEEE, 2015.

Andrychowicz, O. M., Baker, B., Chociej, M., Jozefowicz, R., McGrew, B., Pachocki, J., Petron, A., Plappert, M., Powell, G., Ray, A., et al. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020.

Aytar, Y., Pfaff, T., Budden, D., Paine, T. L., Wang, Z., and de Freitas, N. Playing hard exploration games by watching youtube. *arXiv preprint arXiv:1805.11592*, 2018.

Bachman, P., Hjelm, R. D., and Buchwalter, W. Learning representations by maximizing mutual information across views. *arXiv preprint arXiv:1906.00910*, 2019.

Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., and Erhan, D. Domain separation networks. *arXiv preprint arXiv:1608.06019*, 2016.

Burgess, C. P., Matthey, L., Watters, N., Kabra, R., Higgins, I., Botvinick, M., and Lerchner, A. Monet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019.

Campos, V., Sprechmann, P., Hansen, S., Barreto, A., Kapturowski, S., Vitvitskyi, A., Badia, A. P., and Blundell, C. Coverage as a principle for discovering transferable behavior in reinforcement learning. *arXiv preprint arXiv:2102.13515*, 2021.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020a.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv:2002.05709 [cs, stat]*, June 2020b. URL http://arxiv.org/abs/2002.05709. arXiv: 2002.05709.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. Big Self-Supervised Models are Strong Semi-Supervised Learners. *arXiv:2006.10029 [cs, stat]*, October 2020c. URL http://arxiv.org/abs/2006.10029. arXiv: 2006.10029.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May 2019. URL http://arxiv.org/abs/1810.04805. arXiv: 1810.04805.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. Decaf: A deep convolutional activation feature for generic visual recognition. arxiv e-prints. 2013.

Eslami, S. A., Rezende, D. J., Besse, F., Viola, F., Morcos, A. S., Garnelo, M., Ruderman, A., Rusu, A. A., Danihelka, I., Gregor, K., et al. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018.

Eysenbach, B., Gupta, A., Ibarz, J., and Levine, S. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Greff, K., Kaufman, R. L., Kabra, R., Watters, N., Burgess, C., Zoran, D., Matthey, L., Botvinick, M., and Lerchner, A. Multi-object representation learning with iterative variational inference. In *International Conference on Machine Learning*, pp. 2424–2433. PMLR, 2019.

Gregor, M., Nemec, D., Janota, A., and Pirník, R. A visual attention operator for playing pac-man. In *2018 ELEKTRO*, pp. 1–6. IEEE, 2018.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

Guo, Z. D., Pires, B. A., Piot, B., Grill, J.-B., Altché, F., Munos, R., and Azar, M. G. Bootstrap latent-predictive representations for multitask reinforcement learning. In *International Conference on Machine Learning*, pp. 3875–3886. PMLR, 2020.

Ha, D. and Schmidhuber, J. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Hafner, D., Lillicrap, T., Ba, J., and Norouzi, M. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

Hansen, S., Dabney, W., Barreto, A., Van de Wiele, T., Warde-Farley, D., and Mnih, V. Fast task inference with variational intrinsic successor features. *arXiv preprint arXiv:1906.05030*, 2019.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Held, D., McCarthy, Z., Zhang, M., Shentu, F., and Abbeel, P. Probabilistically safe policy transfer. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5798–5805. IEEE, 2017.

Henaff, O. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pp. 4182–4192. PMLR, 2020.

Higgins, I., Pal, A., Rusu, A., Matthey, L., Burgess, C., Pritzel, A., Botvinick, M., Blundell, C., and Lerchner, A. Darla: Improving zero-shot transfer in reinforcement learning. In *International Conference on Machine Learning*, pp. 1480–1490. PMLR, 2017.

Jaderberg, M., Mnih, V., Czarnecki, W. M., Schaul, T., Leibo, J. Z., Silver, D., and Kavukcuoglu, K. Reinforcement learning with unsupervised auxiliary tasks. *arXiv preprint arXiv:1611.05397*, 2016.

Kaiser, L., Babaeizadeh, M., Milos, P., Osinski, B., Campbell, R. H., Czechowski, K., Erhan, D., Finn, C., Kozakowski, P., Levine, S., et al. Model-based reinforcement learning for atari. *arXiv preprint arXiv:1903.00374*, 2019.

Khan, M., Srivatsa, P., Rane, A., Chenniappa, S., Hazariwala, A., and Maes, P. Personalizing pre-trained models, 2021.

Lake, B., Ullman, T., Tenenbaum, J., and Gershman, S. Building machines that learn and think like people.(c), 1–89. doi: 10.1017. *S0140525X16001837*, 2016.

Laskin, M., Lee, K., Stooke, A., Pinto, L., Abbeel, P., and Srinivas, A. Reinforcement learning with augmented data. *arXiv preprint arXiv:2004.14990*, 2020.

Liu, G., Zhang, C., Zhao, L., Qin, T., Zhu, J., Li, J., Yu, N., and Liu, T.-Y. Return-based contrastive representation learning for reinforcement learning. *arXiv preprint arXiv:2102.10960*, 2021.

Liu, H. and Abbeel, P. Behavior from the void: Unsupervised active pre-training. *arXiv preprint arXiv:2103.04551*, 2021.

Manchin, A., Abbasnejad, E., and van den Hengel, A. Reinforcement learning with attention that works: A self-supervised approach. In *International Conference on Neural Information Processing*, pp. 223–230. Springer, 2019.

Mott, A., Zoran, D., Chrzanowski, M., Wierstra, D., and Rezende, D. J. Towards interpretable reinforcement learning using attention augmented agents. *arXiv preprint arXiv:1906.02500*, 2019.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Oquab, M., Bottou, L., Laptev, I., and Sivic, J. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1717–1724, 2014.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., and Efros, A. A. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016.

Peng, X. B., Andrychowicz, M., Zaremba, W., and Abbeel, P. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 3803–3810. IEEE, 2018.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. 2021.

Rusu, A., Vecerik, M., Rothörl, T., Heess, N., Pascanu, R., and Hadsell, R. Sim-to-real robot learning from pixels with progressive nets. arxiv. *arXiv preprint arXiv:1610.04286*, 2016.

Sadeghi, F. and Levine, S. Cad2rl: Real single-image flight without a single real image. *arXiv preprint arXiv:1611.04201*, 2016.

Salter, S., Rao, D., Wulfmeier, M., Hadsell, R., and Posner, I. Attention-privileged reinforcement learning. *arXiv preprint arXiv:1911.08363*, 2019.

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839): 604–609, 2020.

Schwarzer, M., Rajkumar, N., Noukhovitch, M., Anand, A., Charlin, L., Hjelm, D., Bachman, P., and Courville, A. Pretraining reward-free representations for data-efficient reinforcement learning.

Sekar, R., Rybkin, O., Daniilidis, K., Abbeel, P., Hafner, D., and Pathak, D. Planning to explore via self-supervised world models. In *International Conference on Machine Learning*, pp. 8583–8592. PMLR, 2020.

Sermanet, P., Lynch, C., Chebotar, Y., Hsu, J., Jang, E., Schaal, S., Levine, S., and Brain, G. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1134–1141. IEEE, 2018.

Sharma, A., Gu, S., Levine, S., Kumar, V., and Hausman, K. Dynamics-aware unsupervised discovery of skills. *arXiv preprint arXiv:1907.01657*, 2019.

Shelhamer, E., Mahmoudieh, P., Argus, M., and Darrell, T. Loss is its own reward: Self-supervision for reinforcement learning. *arXiv preprint arXiv:1612.07307*, 2016.

Sorokin, I., Seleznev, A., Pavlov, M., Fedorov, A., and Ignateva, A. Deep attention recurrent q-network. *arXiv preprint arXiv:1512.01693*, 2015.

Srinivas, A., Laskin, M., and Abbeel, P. Curl: Contrastive unsupervised representations for reinforcement learning. *arXiv preprint arXiv:2004.04136*, 2020.

Stooke, A., Lee, K., Abbeel, P., and Laskin, M. Decoupling representation learning from reinforcement learning. *arXiv preprint arXiv:2009.08319*, 2020.

Tan, M., Pang, R., and Quoc, V. L. Efficientdet: Scalable and efficient object detection. arxiv e-prints, page. *arXiv preprint arXiv:1911.09070*, 2, 2019.

Tassa, Y., Doron, Y., Muldal, A., Erez, T., Li, Y., Casas, D. d. L., Budden, D., Abdolmaleki, A., Merel, J., Lefrancq, A., et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

Teed, Z. and Deng, J. Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, pp. 402–419. Springer, 2020.

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., and Abbeel, P. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 23–30. IEEE, 2017.

van Steenkiste, S., Greff, K., and Schmidhuber, J. A perspective on objects and systematic generalization in model-based rl. *arXiv preprint arXiv:1906.01035*, 2019.

Viereck, U., Pas, A., Saenko, K., and Platt, R. Learning a visuomotor controller for real world robotic grasping using simulated depth images. In *Conference on Robot Learning*, pp. 291–300. PMLR, 2017.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Warde-Farley, D., Van de Wiele, T., Kulkarni, T., Ionescu, C., Hansen, S., and Mnih, V. Unsupervised control through non-parametric discriminative rewards. *arXiv preprint arXiv:1811.11359*, 2018.

Wulfmeier, M., Posner, I., and Abbeel, P. Mutual alignment transfer learning. In *Conference on Robot Learning*, pp. 281–290. PMLR, 2017.

Yarats, D., Zhang, A., Kostrikov, I., Amos, B., Pineau, J., and Fergus, R. Improving sample efficiency in model-free reinforcement learning from images. *arXiv preprint arXiv:1910.01741*, 2019.

Yuezhang, L., Zhang, R., and Ballard, D. H. An initial attempt of combining visual selective attention with deep reinforcement learning. *arXiv preprint arXiv:1811.04407*, 2018.

Zhang, R., Liu, Z., Zhang, L., Whritner, J. A., Muller, K. S., Hayhoe, M. M., and Ballard, D. H. Agil: Learning attention from human for visuomotor tasks. In *Proceedings of the european conference on computer vision (eccv)*, pp. 663–679, 2018.

Zhu, G., Huang, Z., and Zhang, C. Object-oriented dynamics predictor. *arXiv preprint arXiv:1806.07371*, 2018.

Abbreviation	Game Name
As	Asteroids
Bz	Berzerk
Bw	Bowling
Bx	Boxing
Br	Breakout
Da	DemonAttack
Fw	Freeway
Fb	Frostbite
He	Hero
Mr	Montezuma's Revenge
Mp	MsPacman
Pf	Pitfall
Pg	Pong
Pe	PrivateEye
Qb	Qbert
Rr	Riverraid
Sq	Seaquest
Si	SpaceInvaders
Tn	Tennis
Vt	Venture
Vp	VideoPinball
Yr	YarsRevenge
μ	Average

Table 1. Caption

A. Atari Game Abbreviations

Table 1 shows the list of all games and their respective abbreviations used in our paper.