Pre-Trained Image Encoder for Generalizable Visual Reinforcement Learning

Zhecheng Yuan [1] **Zhengrong Xue** [2] **Bo Yuan** [1] **Xueqian Wang** [1]
Yi Wu [1] **Yang Gao** [1] **Huazhe Xu** [3]

Abstract

Learning generalizable policies that can adapt to unseen environments remains challenging in visual Reinforcement Learning (RL). Existing approaches try to acquire a robust representation via diversifying the appearances of in-domain observations for better generalization. Limited by the specific observations of the environment, these methods ignore the possibility of exploring diverse real-world image datasets. In this paper, we investigate how a visual RL agent would benefit from the off-the-shelf visual representations. Surprisingly, we find that the early layers in an ImageNet pre-trained ResNet model could provide rather generalizable representations for visual RL. Hence, we propose **P**re-trained **I**mage **E**ncoder for **G**eneralizable visual reinforcement learning (PIE-G), a simple yet effective framework that can generalize to the unseen visual scenarios in a zero-shot manner. Extensive experiments are conducted on DMControl Generalization Benchmark, DMControl Manipulation Tasks, and Drawer World to verify the effectiveness of PIE-G. Empirical evidence suggests PIE-G can significantly outperforms previous state-of-the-art methods in terms of generalization performance. In particular, PIE-G boasts a $\mathbf{55\%}$ generalization performance gain on average in the challenging video background setting.

1. Introduction

Visual Reinforcement Learning (RL) has achieved significant success in learning complex behaviors directly from image observations (Mnih et al., 2015; Kalashnikov et al., 2018; Kostrikov et al., 2020). Despite the progress, RL agents are often plagued by the overfitting problem (Song et al., 2019), especially in high-dimensional observation space. Previous studies show that it is difficult for the visual agents to generalize to unseen scenarios (Cobbe et al., 2019; Lee et al., 2019), which severely limits their deployment in real-world applications.

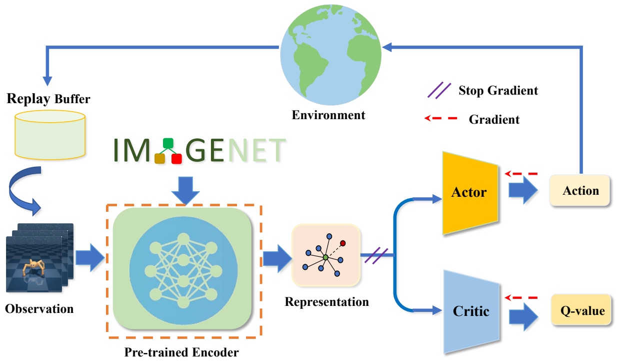


Figure 1. **Overview of PIE-G**. This figure shows the general framework of PIE-G where visual encoders embed high-dimensional images into low-dimensional representations for downstream decision-making tasks. Instead of training the encoder from scratch, PIE-G selects an ImageNet pre-trained ResNet model as the encoder and freezes its parameters during the entire training process.

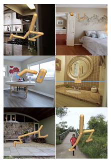
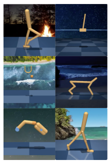
In general, visual RL methods rely on their encoders to learn a visual representation to perceive the world. Recent studies have found that data augmentation (Shorten & Khoshgoftaar, 2019) leads to more generalizable representations so that the agents can adapt to the unseen environments with different visual appearances (Raileanu et al., 2021; Wang et al., 2020). However, most of those approaches only augment the observations of the training environments (Kostrikov et al., 2020; Laskin et al., 2020; Srinivas et al., 2020), which is unable to provide enough diversity for generalization over large domain gaps. Furthermore, naively applying data augmentation may damage the robustness of learned representations and decrease training sample efficiency (Hansen et al., 2021; Yuan et al., 2022).

To overcome these drawbacks, what we require is a universal representation that can generalize to a variety of unseen scenarios. Recent works in representation learning demonstrate promising results in enabling pre-trained models to

[1]Tsinghua University [2]Shanghai Jiao Tong University [3]Stanford University. Correspondence to: Zhecheng Yuan <yuanzc20@mails.tsinghua.edu.cn>, Huazhe Xu <huazhexu@stanford.edu>.

First Workshop of Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

Table 1. **Generalization on unseen moving backgrounds.** Episode return in two types of unseen dynamic video background environments, i.e., *video easy* (*Bottom*) and *video hard* (*Top*). PIE-G achieves competitive or better performance in **9** out of **12** tasks. In *video hard* setting, we significantly outperforms other algorithms with **+55%** improvement on average.

Setting	DMControl Tasks	DrQ	DrQ-v2	SVEA	TLDA	**PIE-G**
	Cartpole-Swingup	138±9	130±3	393±45	286±47	**401**±21 (+2.0%)
	Walker-Stand	289±49	151±13	834±46	602±51	**852**±56 (+2.2%)
	Walker-Walk	104±22	34±11	377±93	271±55	**600**±28 (+59.2%)
	Ball_in_cup-Catch	92±23	97±27	403±174	257±57	**786**±47 (+95.0%)
	Cheetah-Run	32±13	23±5	105±37	90±27	**154**±17 (+46.6%)
	Finger-Spin	71±45	21±4	335±58	241±29	**762**±59 (+127%)
	Cartpole-Swingup	485±105	267±41	**782**±27	671±57	614±60
	Walker-Stand	873±83	560±48	961±8	**973**±6	**965**±6
	Walker-Walk	682±89	175±117	819±71	873±34	**887**±22
	Ball_in_cup-Catch	318±157	454±60	871±106	892±68	**922**±20
	Cheetah-Run	102±30	64±22	249±20	**336**±57	305±60
	Finger-Spin	533±119	456±19	808±33	744±18	**837**±107

provide strong priors for downstream tasks (He et al., 2020; Devlin et al., 2018). The pre-trained models contain representations obtained from a wide range of existing real-world image datasets. These representations are proved to be robust to noises and capable of distinguishing salient features despite the diversity and the variability (Donahue et al., 2014). Based on the observations, we would like to ask the following question: is it possible to train a visual RL agent that is augmented with pre-trained visual representations so that it can better generalize to novel tasks?

Towards answering the question, the main contribution of this paper is a surprising discovery that the off-the-shelf features of frozen models trained with ImageNet can be used as universal representations for visual RL. Based on such findings, we present **P**retrained **I**mage **E**ncoder for **G**eneralizable visual reinforcement learning (PIE-G), a visual RL framework that allows agents to obtain enhanced generalization ability via integrating the extracted representations from a pre-trained ResNet (He et al., 2016) encoder into RL training. Straightforward as the framework appears, PIE-G enjoys thoughtful details and nuanced design choices to acquire representations that are suitable for control and generalizable to novel scenarios. Specifically, we show that the choice of early layer features and the ever-updating Batch Normalization (BatchNorm) (Ioffe & Szegedy, 2015) are crucial for the performance gain.

2. Method

In this section, we introduce PIE-G, a simple yet effective framework for visual RL which benefits from the pre-trained encoders on other domains to facilitate the generalization ability.

2.1. Pre-trained Encoder

PIE-G explicitly leverages the pre-trained models as the representation extractor without any modification. The pre-trained encoder projects high-dimensional image observations into low-dimensional embeddings that are later used by RL policies. Note that PIE-G is as simple as importing a pre-trained ResNet model from the *torchvision* (Marcel & Rodriguez, 2010) library. This avoids the design of any auxiliary tasks to acquire useful representations.

For all the training tasks on different benchmarks, the encoder's parameters are frozen to obtain universal visual representations. Since the pre-trained model contains the priors from a wide range of real-world images, we hypothesize that the inherited power from a pre-trained model may help to capture and distinguish the main components of different tasks' observations regardless of the changes of visual appearances or deformed shapes, and will further improve the generalization abilities of RL agents.

To validate our hypothesis, we first encode each observation independently to obtain embeddings. Then, the embeddings from the second layer of the pre-trained model are fused as input features to the policy networks (Shang et al., 2021; Pari et al., 2021). Moreover, we enable BatchNorm (Ioffe & Szegedy, 2015) to keep updating the running mean and running standard deviation during the policy training. The key findings are: 1) early layers of a neural network would provide better representations for visual RL generalization, which resonates with prior works in imitation learning (Parisi et al., 2022); 2) the always updating statistics in BatchNorm helps better adapt to the shift in observation space and thus improve the generalization ability.

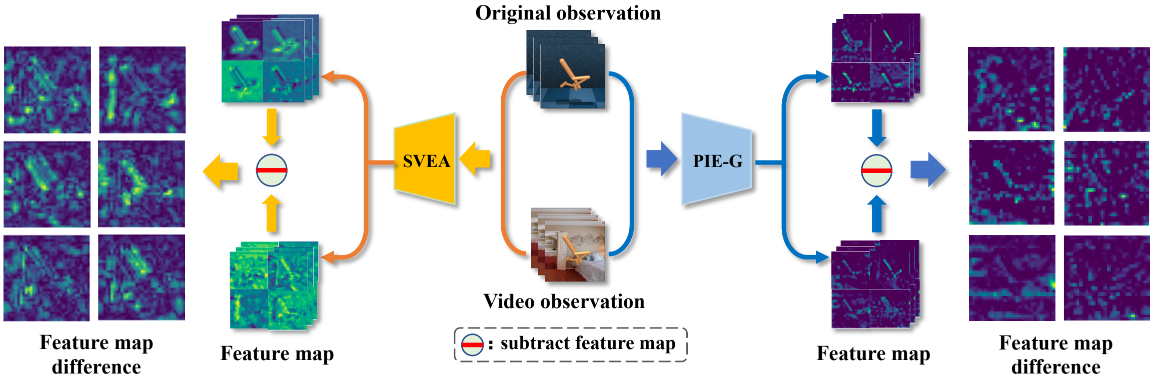


Figure 2. **Visualized feature map differences of two inputs from the same state with different backgrounds.** The difference of the feature maps with PIE-G as the encoder is closer to zero than that with SVEA, indicating PIE-G enjoys better generalization ability.

2.2. Reinforcement Learning Backbone

We implement DrQ-v2 (Yarats et al., 2021) as the base visual reinforcement learning algorithm. DrQ-v2 is the state-of-the-art method for visual continuous control tasks, which adopts DDPG (Lillicrap et al., 2015) coupling with clipped Double Q-learning (Fujimoto et al., 2018) to alleviate the overestimation bias of target Q-value.

We emphasize that PIE-G does not need any other proprioceptive states and sensory information as the inputs besides the representations extracted from original image observations. In the setting of generalization, we follow stronger augmentation methods (e.g., mixup) of SVEA (Hansen et al., 2021) and DrAC (Raileanu et al., 2021) to further boost the performance. It is worth mentioning that since the gradient is stopped before it reaches the encoder, all the data augmentation techniques discussed here do not affect the pre-trained visual representation. Meanwhile, unlike Rutav et al. (Shah & Kumar, 2021) and Simone et al. (Parisi et al., 2022), we purely train the agent in a standard RL paradigm without any expert's demonstration.

3. Experiments

In this section, we investigate the following questions: (1) Can PIE-G improve the agent's generalization ability? Specifically, how well does PIE-G deal with moving or unseen video backgrounds, and deformed shapes of robots? (2) How do the choice of layers in the encoder and the use of BatchNorm affect the performance?

3.1. Evaluation on Generalization Ability

Generalization on unseen or moving backgrounds. We then evaluate PIE-G on the more challenging settings: *video easy* and *video hard* in DMC-GB. The *video hard* setting consists of more complicated and fast-switching video backgrounds that are drastically different from the training environments. Notably, even the reference plane of the ground is removed in this setting.

The comparison results are shown in Table 1. PIE-G achieves better or comparable performance with the prior state-of-the-art methods in **9** out of **12** instances. In particular, PIE-G gains significant improvement in the *video hard* setting over all the previous methods with **+55%** improvement on average. For example, in the *Finger Spin*, *Cup Catch*, and *Walker Walk* tasks, PIE-G outperforms the best of the other methods by substantial margins **127.0%**, **95.0%**, and **59.2%** respectively.

Attempting to explain the success, we visualize the difference of the normalized feature maps extracted from the encoder whose inputs are two Walkers of the same pose but with different backgrounds, as is shown in Figure 2. Ideally, a well-generalizable encoder would map the observations of the two Walkers to exactly the same embedding, and therefore the difference should be zero. In practice, as shown in Figure 2, the encoder of PIE-G produces a difference much closer to zero than that of SVEA. Numerically, we calculate the average pixel intensity in the difference of normalized feature maps, and the intensity is decreased by 50.9% with PIE-G than that with SVEA.

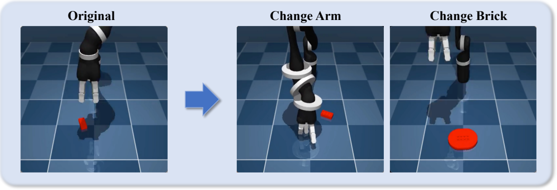


Figure 3. **Visualization of the deformed shape.** Aiming at evaluating the agent's robustness of the shape, we deform the robot's arm and the target brick.

Generalization on deformed shapes. To verify agent's robustness in terms of the deformed shapes, we modify the shapes of the jaco arm and the target objects in the manipulation tasks, as shown in Figure 3. Figure 4(a) demonstrates

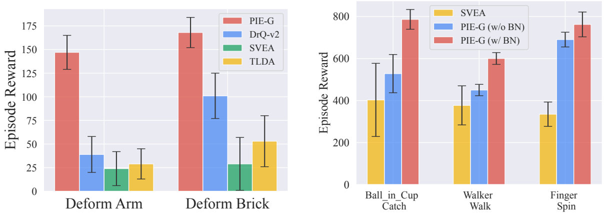
Task	Setting	DrQ-v2	SVEA	**PIE-G**
Drawer-Close	Training	**98%**	70%	**99%**
	Wood	32%	49%	**59%**
	Metal	46%	69%	**95%**
	Blanket	8%	**72%**	**71%**

Task	Setting	DrQ-v2	SVEA	**PIE-G**
Drawer-Open	Training	**100%**	75%	**97%**
	Wood	2%	47%	**79%**
	Metal	53%	71%	**97%**
	Blanket	5%	37%	**85%**

Table 2. **Generalization on Drawer World.** Evaluation on distracting textures. PIE-G is robust to the texture changing.

that PIE-G also improves the agents' generalization ability with various shapes while other methods could barely generalize to these changes. We attribute this to the lack of shape changing in previous data augmentation techniques. Conversely, our pre-trained encoder is learned from a multitude of real-world images with various poses and shapes, thus enhancing its generalization ability on deformed shapes.

Furthermore, we conduct experiments on the *DrawerWorld* benchmark to test the agent's generalization ability in manipulation tasks with different background textures. *Success Rate* is adopted as the evaluation metric for its goal-conditioned nature. Table 2 illustrates that PIE-G can achieve better or comparable generalization performance in all the settings with **+24%** boost on average while other approaches may suffer from the CNN's sensitivity in the face of various textures (Geirhos et al., 2018).



(a) (b)

Figure 4. **Generalization Performance.** *Left*: The results demonstrate that PIE-G is well-generalizable in the face of deformed shapes. *Right*: This figure shows that ever-updating BatchNorm is beneficial for better performance.

3.2. Choice of layers

In convolutional neural networks, the later layers capture high-level semantic features, while the early layers are responsible for extracting low-level information (Ma et al., 2015; Zeiler & Fergus, 2014; Lin et al., 2017). Table 3 investigate how much control tasks can benefit from the features extracted from different layers. As shown in Figure 5, the early layers preserve rich details of edges and corners, while the later layers only provide very abstract information.

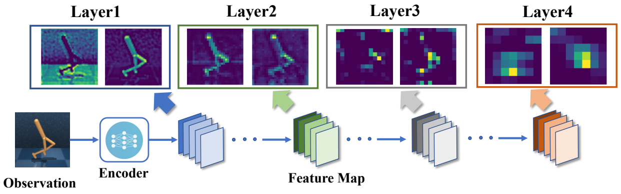


Figure 5. **Visualization of the feature maps of different layers .** The feature map of Layer 2 largely preserves the outline of the Walker that is advantageous to the control tasks, and at the same time discards redundant details.

Intuitively, for control tasks, a trade-off is required between low-level details and high-level semantics. Table 3 shows that the Layer 2 gains better performance than the other layers.

Task	Layer 1	Layer 2	Layer 3	Layer 4
Walker Walk	840±32	**884±20**	845±27	306±31
Cheetah Run	**366±56**	**369±53**	294±60	111±19
Walker Stand	953±8	**964±7**	**957±7**	625±116

Table 3. **Different layers.** We employ the feature map of different layers of a ResNet model as the visual representation. Among them, the Layer 2 exhibits the best generalization performance.

3.3. Batch Normalization

Batch Normalization (BatchNorm) (Ioffe & Szegedy, 2015) is a popular technique in computer vision. However, it is not widely adopted in RL algorithms. In contrast to conventional wisdom, BatchNorm is found to be useful and important in PIE-G. Specifically, we find that calculating the mean and variance of the observations during evaluation rather than using the statistics from training data would boost the performance. Figure 4(b) demonstrates that, in the most challenging settings, PIE-G with the use of BatchNorm can further improve the generalization performance. This is largely because the distribution of observations is determined by the agent, violating the assumption of independent and identical distribution (i.i.d.). This use of BatchNorm also reassures the recommendation from Ioffe et al. (Ioffe & Szegedy, 2015) that recomputation of the statistical means and variances allows the BatchNorm layer to generalize to new data distributions.

4. Conclusion

In this work, we propose PIE-G, a simple yet effective framework that leverages off-the-shelf features of ImageNet pre-trained ResNet models for better generalization in visual RL. Extensive experiments on a variety of tasks in three RL environments confirm the merits of universal visual representations, which endow the agents with better generalization performance. In addition, we show that the choice of layers and the use of BatchNorm are crucial for the performance gain.

References

Cobbe, K., Klimov, O., Hesse, C., Kim, T., and Schulman, J. Quantifying generalization in reinforcement learning. In *International Conference on Machine Learning*, pp. 1282–1289. PMLR, 2019.

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.

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. In *International conference on machine learning*, pp. 647–655. PMLR, 2014.

Fujimoto, S., Hoof, H., and Meger, D. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pp. 1587–1596. PMLR, 2018.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Hansen, N., Su, H., and Wang, X. Stabilizing deep q-learning with convnets and vision transformers under data augmentation. *Advances in Neural Information Processing Systems*, 34, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. PMLR, 2015.

Kalashnikov, D., Irpan, A., Pastor, P., Ibarz, J., Herzog, A., Jang, E., Quillen, D., Holly, E., Kalakrishnan, M., Vanhoucke, V., et al. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on Robot Learning*, pp. 651–673. PMLR, 2018.

Kostrikov, I., Yarats, D., and Fergus, R. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. *arXiv preprint arXiv:2004.13649*, 2020.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Laskin, M., Lee, K., Stooke, A., Pinto, L., Abbeel, P., and Srinivas, A. Reinforcement learning with augmented data. *Advances in Neural Information Processing Systems*, 33: 19884–19895, 2020.

Lee, K., Lee, K., Shin, J., and Lee, H. Network randomization: A simple technique for generalization in deep reinforcement learning. *arXiv preprint arXiv:1910.05396*, 2019.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

Ma, C., Huang, J.-B., Yang, X., and Yang, M.-H. Hierarchical convolutional features for visual tracking. In *Proceedings of the IEEE international conference on computer vision*, pp. 3074–3082, 2015.

Marcel, S. and Rodriguez, Y. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1485–1488, 2010.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533, 2015.

Pari, J., Muhammad, N., Arunachalam, S. P., Pinto, L., et al. The surprising effectiveness of representation learning for visual imitation. *arXiv preprint arXiv:2112.01511*, 2021.

Parisi, S., Rajeswaran, A., Purushwalkam, S., and Gupta, A. The unsurprising effectiveness of pre-trained vision models for control. *arXiv preprint arXiv:2203.03580*, 2022.

Raileanu, R., Goldstein, M., Yarats, D., Kostrikov, I., and Fergus, R. Automatic data augmentation for generalization in reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Shah, R. and Kumar, V. Rrl: Resnet as representation for reinforcement learning. *arXiv preprint arXiv:2107.03380*, 2021.

Shang, W., Wang, X., Srinivas, A., Rajeswaran, A., Gao, Y., Abbeel, P., and Laskin, M. Reinforcement learning with latent flow. *Advances in Neural Information Processing Systems*, 34, 2021.

Shorten, C. and Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

Song, X., Jiang, Y., Tu, S., Du, Y., and Neyshabur, B. Observational overfitting in reinforcement learning. *arXiv preprint arXiv:1912.02975*, 2019.

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

Wang, K., Kang, B., Shao, J., and Feng, J. Improving generalization in reinforcement learning with mixture regularization. *Advances in Neural Information Processing Systems*, 33:7968–7978, 2020.

Yarats, D., Fergus, R., Lazaric, A., and Pinto, L. Mastering visual continuous control: Improved data-augmented reinforcement learning. *arXiv preprint arXiv:2107.09645*, 2021.

Yuan, Z., Ma, G., Mu, Y., Xia, B., Yuan, B., Wang, X., Luo, P., and Xu, H. Don't touch what matters: Task-aware lipschitz data augmentation for visual reinforcement learning. *arXiv preprint arXiv:2202.09982*, 2022.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.