時系列予測モデルを導入した価値関数に基づく強化学習

○西片智広 †，山内悠嗣 †

○ Tomohiro NISHIKATA † and Yuji YAMAUCHI †

†：中部大学，tr21010-0146@sti.chubu.ac.jp，yuu@isc.chubu.ac.jp

＜要約＞ 強化学習は環境とエージェントの相互のやり取りにより，一定期間における報酬の期待値を表す価値が最大となるように学習することで，エージェントが取るべき行動を獲得できる教師なし学習手法の 1 つである．高い価値を得るためには，未知である将来の状態において最適な行動を選択する必要がある．未知である将来の状態を事前に把握できれば，より良い行動を選択できるため，結果的に高い価値を得ることが可能である．そこで，本研究では深層学習に基づく未来画像生成技術を利用することで，未知である将来の状態を事前に予測する．事前に将来の状態を予測することで，より高い価値を得るための行動を選択することが可能となるため，早期に高い報酬が得られることが期待できる．

＜キーワード＞強化学習，時系列予測，深層学習

1 はじめに

試行錯誤しながらタスクを解くための適切な行動を学習する強化学習が精力的に取り組まれている．強化学習は，囲碁，将棋などのゲームのクリア，ロボットアームによる物体把持動作の獲得などのタスクにおいて有効性が確認されている．強化学習は事前に用意した学習データを与えて学習するのではなく，行動することで得た自身の経験から学習するため，人間が明確な正解を与えることが困難であるタスクを解くことができる可能性を持っている．

強化学習は観測した現在までの状態における価値を最大化するように学習する．価値とは将来にわたって獲得できる報酬の期待値であるため，先の状態を予測できれば，現在の状態のより高い価値を求めることができる．

そこで，本研究ではより高い現在の状態の価値を求めるために，価値を計算する際に先の状態を予測する時系列モデルを導入する．先の状態を予測するために，連続した画像の時系列データから次時刻以降に観測されるであろう未来の画像を予測する深層学習モデルを用いる．予測モデルにより先の状態を予測し，より高い現在の状態の価値を求めることで，早期に高い報酬が得られることが期待できる．

2 関連研究

2.1 強化学習に関する研究

深層学習の発展と共に強化学習に関して精力的に研究されている．Deep Q Network(DQN)[1] は強化学習手法の 1 つである Q 学習 [2] と深層学習ネットワークを組み合わせた手法である．Q 学習は，価値関数を用いて価値を最大化する行動を決める最適方策を学習する手法であり，DQN では価値関数をニューラルネットワークで近似する．DQN は価値関数により最適方策を求める価値ベースの手法であり，他にも価値関数を介することなく方策から直接的に最適方策を求める方策ベースの手法 [3, 4, 5] や，価値ベースと方策ベースの両方を取り入れた actor-critic[6, 7, 8] が提案されている．actor-critic は，深層学習ネットワークでモデル化した価値関数と方策の 2 つを学習する手法である．

actor-critic の発展手法として，soft-actor-critic (SAC)[9, 10, 11] が提案されている．SAC は，価値を推定する Q ネットワークと行動を決定する方策ネットワークの学習の際に，方策のエントロピーにより行動の探索力を評価し，価値と探索力の最大化を目的として学習する．これにより行動の探索力が向上し，学習が安定的となる．

深層学習手法の仕組みを導入することでより良い学習

を実現する強化学習手法も提案されている．Contrastive Unsupervised Reinforcement Learning(CURL)[12] と Data-regularized Q(DrQ)[13] は，対照学習を導入した強化学習手法である．CURL と DrQ は，環境の状態が画像である場合に，画像からの特徴表現能力を向上させるために対照学習を導入している．対照学習とは，自己教師あり学習のひとつであり，正解ラベルのないデータに対して特徴空間において似ているデータが同じような特徴量を持つように学習する手法である．

2.2 未来画像生成に関する研究

現在までに観測した画像群から，将来観測されるであろう画像を深層学習により生成する手法 [14, 15] が提案されている．提案されている手法の多くは時系列データに対応した深層学習モデルである Long Short Term Memory(LSTM)[16] に畳み込み処理を加えた畳み込み LSTM[17] がベースとなっている．未来画像生成手法の 1 つである PredNet[14] は，畳み込み LSTM を重ね合わせた各層で予測を行い，各層の予測誤差を次の層へ伝えることで高精度な未来画像を生成する手法である．

また，多くの手法ではネットワークの入力に動画像だけではなく，他の情報を追加したマルチモーダルモデルにより高精度な未来画像の生成を実現している．PredNet に，観測した画像群とその画像の動き情報を入力することで高精度な未来画像を生成する手法 [18] が提案されている．Finn らは，ロボットハンドが物体を押す，または引き寄せる動作をしている動画像，ロボットハンドの姿勢と動作の種類の 3 つを入力とし，畳み込み LSTM によりロボットハンドの未来画像を生成する手法 Convolutional Dynamic Neural Advection(CDNA)[19] を提案している．

本研究と同様に，予測モデルを強化学習に導入した手法が幾つか提案されている．WORLD MODEL[20] は，予測モデルである回帰型ニューラルネットワーク (Recurrent Neural Network： RNN) により状態の時間的な特徴をとらえ，行動の決定に利用する．また，Simulated Policy Learning(SimPLe)[21] は，畳み込み LSTM をベースにした予測モデルを環境のシミュレータとして用いる手法である．シミュレータが環境そのものを学習することで，観測した状態には含まれない情報を理解する狙いがある．

2.3 提案手法の概要

本研究では，時系列予測モデルを導入した価値関数に基づく強化学習の手法を提案する．提案手法の特長は以下の通りである．

- 価値の推定に未来画像を考慮
  価値推定時に先の状態を予測した未来画像を考慮することで，現在の状態より高い価値を推定できる．また，生成した未来画像に対して画像認識技術を適用することが可能となり，様々なタスクに応用することができる．
- 学習時のみ未来画像生成器を使用
  評価時は方策ネットワークを持つエージェントのみを利用し，計算コストが高い未来画像生成の処理を行わない．そのため，評価時の計算量は従来法と同じとなる．

3 提案手法

3.1 提案手法の流れ

図 1 に提案手法の流れを示す．提案手法は，強化学習パートと未来画像生成パートの 2 つで構成される．予測モデルである未来画像生成器には学習済みのモデルを使用し，価値を推定する Q ネットワークと行動を決定する方策ネットワークを学習する．

まず，環境から観測した時刻 t における画像 s_t をエンコーダを介して方策ネットワークに入力し，行動 a_t を決定する．次に画像 s_t と行動 a_t を学習済みの未来画像生成器に入力し，1 フレーム先の未来画像 $\hat{s}_{t+1}$ を生成する．その後，生成した未来画像 $\hat{s}_{t+1}$ を方策ネットワークに与えたときの時刻 $t+1$ の行動 a_{t+1} を出力し，さらに 1 フレーム先の未来画像 $\hat{s}_{t+2}$ を生成する．これを繰り返すことで N フレーム先の未来画像 $\hat{s}_{t+N}$ を生成する．

最後に生成した時刻 $t+N$ までの未来画像をエンコーダを介して Q ネットワークに入力し，先の状態の価値 $Q(\hat{s}_{t+N}, a_{t+N})$ を求める．また，Q ネットワークと方策ネットワークの重みは求めた価値により更新する．

3.2 強化学習

強化学習手法には，Contrastive Unsupervised Reinforcement Learning(CURL)[12] を採用する．CURL は，actor-critic の手法である SAC をベースとし，対照学習と呼ばれるアプローチを導入することで特徴の表

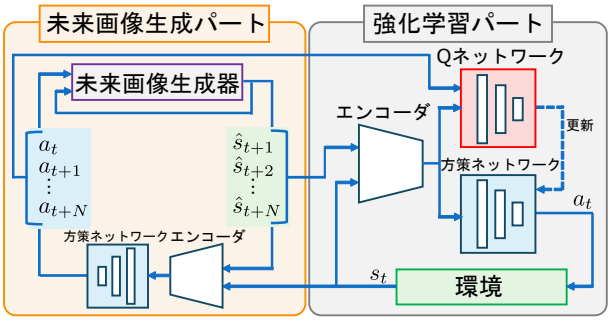

図 1　提案手法の流れ

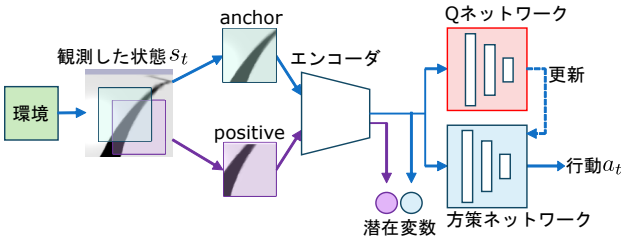

図 2　CURL における対照学習の流れ

現能力を向上させた手法である．図 2 に CURL の対照学習部分の流れを示す．対照学習では観測した画像からランダムに異なる 2 つの領域をトリミングし，anchor と positive の 2 つのデータに拡張する．そして，anchor と positive の間の類似度が高くなるよう，それぞれの潜在変数を用いてエンコーダを学習し，Q ネットワークとエージェントで利用する．

従来法における Q ネットワークの損失関数を式 (1) に示す．

$$L = r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \tag{1}$$

ここで，r_t は時刻 t において獲得する報酬，γ は報酬の削減率である．報酬 r_t 及び次時刻の価値 $Q(s_{t+1}, a_{t+1})$ の和と現在の価値 $Q(s_t, a_t)$ の差分を損失とし，現在の価値が次時刻の価値に近づくように学習する．

一方，提案手法の損失関数は式 (1) の予測した先の状態の価値を加えた式 (2) により定義される．

$$L = r_t + \frac{1}{2}\gamma\{Q(s_{t+1}, a_{t+1}) + \frac{1}{N-1}\sum_{n=2}^{N} Q(\hat{s}_{t+n}, a_{t+n})\} - Q(s_t, a_t) \tag{2}$$

ここで $\hat{s}_{t+n}$ は未来画像生成器で生成した n 時刻先の未来画像である．損失関数に予測した先の状態の価値を含めることで，次時刻の価値が明示的に表現され，現在の価値と次時刻の価値の差分を正確に求められる．

また，本手法は学習時のみ未来画像生成器を導入し，評価時は方策ネットワークを持つエージェントのみを用いる．これにより，大きなモデルである未来画像生成器による処理時間の増加を考慮することなくエージェントを利用することができる．

3.3　未来画像生成

未来画像生成器には Convolutional Dynamic Neural Advection(CDNA)[19] を採用する．CDNA は連続した数フレームの画像群と，それらの画像群から観測されるオブジェクトの動きや姿勢などを条件として加え，1 フレーム先の未来画像を生成する条件付き未来画像生成器である．CDNA は畳み込み LSTM をベースにしており，入力された画像とその画像の条件から，画像に映っているオブジェクトの動きの変化を捉える予測フィルタを生成する．その後，入力された画像に予測フィルタを適用し，1 フレーム先の未来画像を生成する．

図 3 に CDNA の構成を示す．CDNA はエンコーダ・デコーダ型のネットワーク構成であり，エンコーダ，デコーダは畳み込み LSTM をベースとする．まず，エンコーダに時刻 t の画像 s_t を入力して得た特徴量と時刻 t の条件 a_t を結合する．次に結合した特徴量をデコーダに入力し，画像 s_t に映っているオブジェクトの動きの変化を捉えるフィルタ (移動フィルタ) と，オブジェクトの位置を示すマスクフィルタを生成する．その後，画像 s_t に移動フィルタを適用し，マスクフィルタとかけ合わせた複合マスクを生成する．最後に，画像 s_t に複合マスクを適用することで 1 フレーム先の未来画像 $\hat{s}_{t+1}$ を生成する．CDNA は，生成した未来画像 $\hat{s}_{t+1}$ と時刻 $t+1$ の実際の画像 s_{t+1} との平均二乗誤差を損失として学習する．

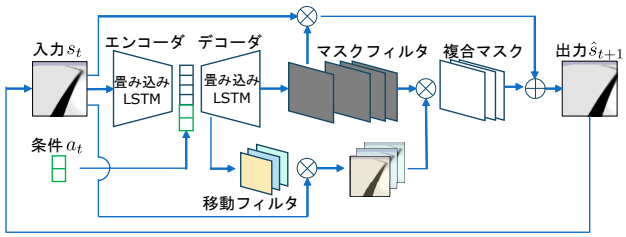

図 3　CDNA の構成

CDNA により生成した未来画像を確認する．ライントレースシミュレータ[1]により走行する車が観測した画像を使用し，条件にその車の左右のタイヤの制御値を加える．CDNA は LSTM をベースにしたネットワーク

[1] https://github.com/nplan/gym-line-follower

であるため，入力された時系列データが多くなるほどネットワークに情報が蓄積され，より高精度な未来画像の生成が可能となる．そのため，CDNA では入力した最初の数フレームでは未来画像の生成は行わず，情報の蓄積のみを行う．ここでは，予測までのフレームの事前入力を $t = 1 \sim 4$ まで行い，$t = 5, 6$ の未来画像を生成する．$t = 6$ の未来画像は，CDNA 自身が生成した $t = 5$ の未来画像をネットワークに入力する外挿により生成する．

図 4 に CDNA により生成した未来画像を示す．図 4 は車がカーブを走行するシーンであり，差分画像より，ラインの位置を正しく予測できていることがわかる．

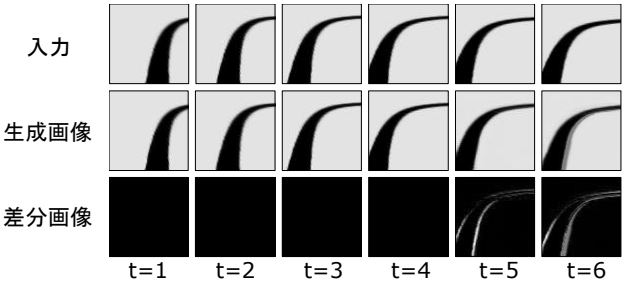

図 4　生成した未来画像

4　評価実験

提案手法の有効性を確認するために評価実験を行う．ライントレースタスクとカートポールタスクの 2 つのタスクで実験し，従来法である CURL と比較する．

4.1　ライントレースタスク

ライントレースタスクの環境を図 5(a) に示す．白色の背景に黒色のラインのコースをランダムに生成し，車がラインに沿って走行することを学習させる．車の前方にはカメラが搭載されており，図 5(b) のような画像を撮影し環境の状態とする．車についた左右のそれぞれのタイヤの制御値 $[-1.0, 1.0]$ をエージェントの行動とし，ラインから逸れないように走行した距離が報酬となる．ライントレースタスクでは，100,000 ステップを 1 回の学習として 5 回学習した平均を従来法と比較する．評価に用いるコースを 3 種類用意し，学習データと同じ条件で生成した評価用コースを normal，normal よりも単純なコースを easy，normal よりも急なカーブを増やしたコースを hard とする．図 6 にそれぞれの評価用コースの例を示す．

図 8(a) に学習中の報酬の遷移，図 8(b)，(c)，(d) に

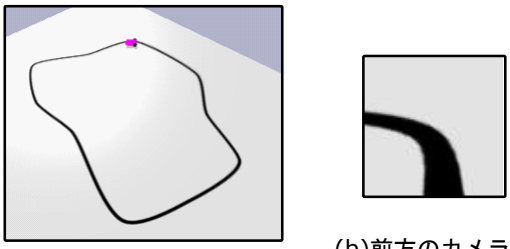

図 5　ライントレースタスクの環境

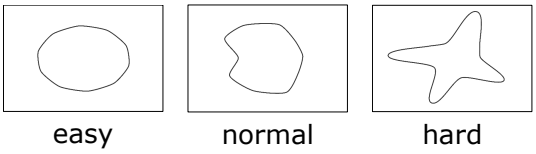

図 6　評価用コースの例

評価コース easy，normal，hard の走行した結果を示す．図 8(a) より 20,000 ステップ辺りから提案手法の方が高い報酬を獲得できていることがわかる．また，図 8(b)，(c) よりどの評価用コースにおいても提案手法の方が高い報酬を獲得していることが確認できる．一方，評価用コース hard は学習データよりも複雑なコースであるため完走することは困難であるが，図 8(d) より提案手法の方が高い報酬を獲得していることが確認できる．これは，提案手法では走行が難しい急なカーブ以外を従来法よりも走行できているためである．

4.2　カートポールタスク

カートポールタスクの環境を図 7 に示す．左右に移動するカートに設置したポールを倒さないようにカートを制御することを学習させる．カートの左右方向の制御値 $[-1.0, 1.0]$ をエージェントの行動とし，図 7 のような画像を環境の状態とする．カートの位置とポールの角度にしたがって報酬が与えられる．

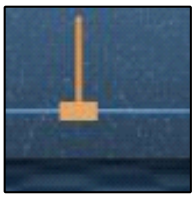

図 7　カートポールタスクの環境

図 9 に学習中の報酬の遷移を示す．図 9 より 140,000 ステップ辺りから提案手法の方が高い報酬を獲得できていることがわかる．

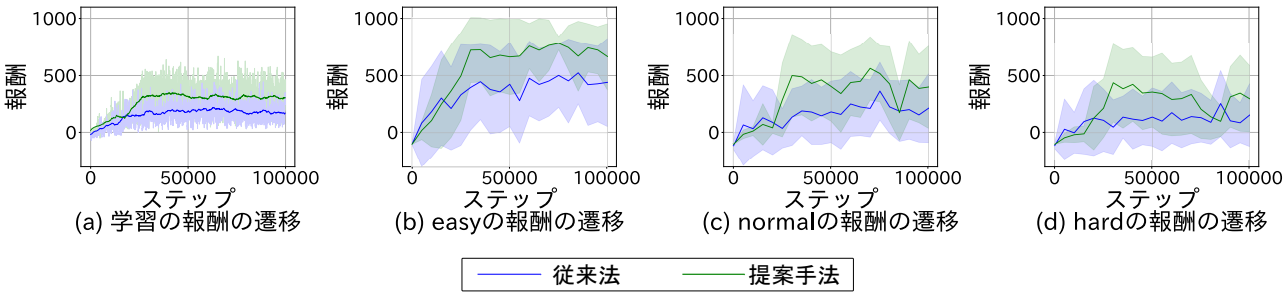

図 8　ライントレースタスクの学習中の報酬と各コースにおける報酬の遷移

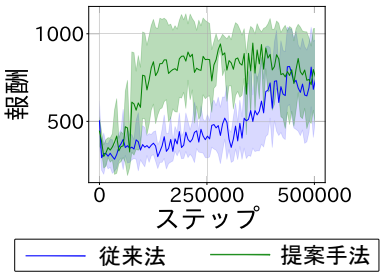

図 9　カートポールタスクの学習中の報酬の遷移

5　考察

4 章の評価実験により，先の状態の価値を考慮した提案手法が高い報酬を得ることを確認した．ここでは，追加した未来画像生成器に関して考察する．

5.1　事前入力数を変更した実験

未来画像生成器の事前入力画像数を変更した際の報酬を確認する．これにより，未来画像生成器の予測の精度が強化学習にどの程度影響があるかを把握することができる．事前入力画像数の異なる未来画像生成器を用いて提案手法によりライントレースタスクを学習する．事前入力画像数を増やすことで未来画像生成器に情報が蓄積され，生成する未来画像の精度の向上が期待できる．事前入力画像数が 2，4，9 の 3 種類の未来画像生成器を用いた提案手法と従来法の報酬を比較する．

実験結果を図 10 に示す．どの未来画像生成器においても，従来法を上回る報酬を獲得していることが確認できる．また，提案手法に事前入力画像数を 4 とした時の報酬が最も高いことが確認できる．表 1 に未来画像生成器の事前入力画像数ごとの生成画像と実際の画像との平均二乗誤差 (MSE) を示す．表 1 から事前入力画像数 4 のときの MSE が最も小さいことがわかる．未来画像生成器に蓄積する情報として与える事前入力画像数は，多いと高い報酬を獲得できるが，多すぎると獲得できる報酬が少なくなることがわかる．

表 1　事前入力画像数における MSE の比較 (1×10^{-3})

事前入力画像数	MSE
2	8.26
4	7.05
9	7.45

5.2　予測するフレーム数を変更した実験

未来画像生成器で予測するフレーム数を変更して報酬を比較する．これにより，どの程度まで先の状態の価値を推定することが適切であるかを実験的に確認することができる．予測するフレーム数の異なる未来画像生成器を用いて，提案手法によりライントレースタスクを学習する．予測するフレーム数が 1，2，4，9 の 4 種類の未来画像生成器を用いた提案手法と従来法の報酬を比較する．

実験結果を図 11 に示す．予測フレーム数が 1，2 では従来法よりも高い報酬を獲得しているが，予測フレーム数が 4，9 では従来法と同じ程度の報酬しか獲得できていない．この結果から，提案手法において，4 ステップ以上の先の状態の価値を推定することが有効ではないことがわかる．表 2 に未来画像生成器の予測フレーム数ごとの MSE を示す．予測するフレーム数の増加に従い，MSE の値も増加している．予測するフレーム数が 2 以上の場合は，外挿により自身が生成した未来画像を用いて次のフレームを予測するため，生成する未来画像の精度は低下する．そのため，精度の低い未来画像を用いて先の価値を推定することになり，学習が進まないと考えられる．

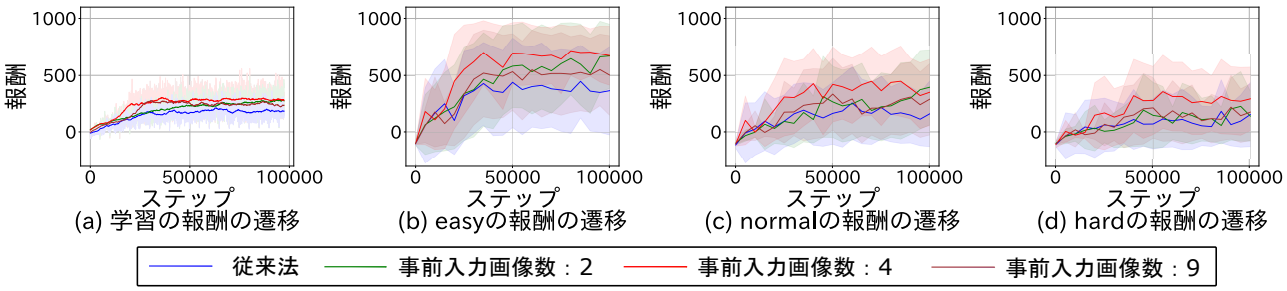

図 10　事前入力画像数を変更したライントレースタスクにおける報酬の遷移

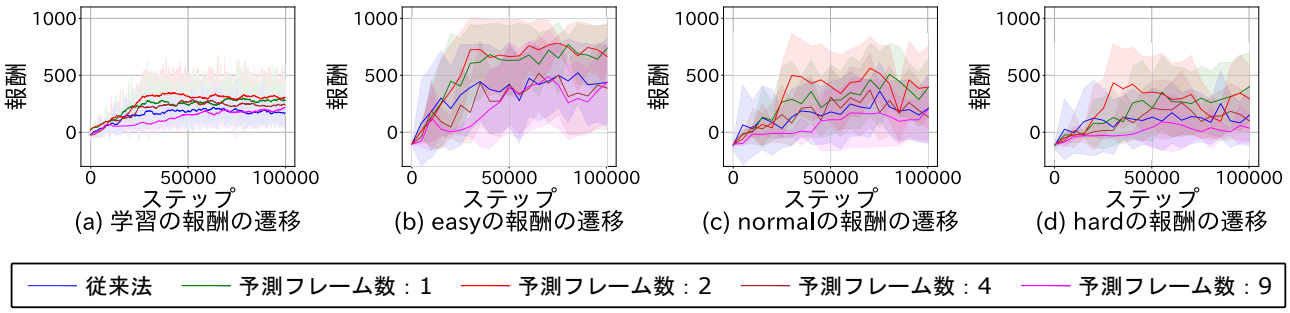

図 11　予測フレーム数を変更したライントレースタスクにおける報酬の遷移

表 2　予測フレーム数における MSE の比較 (1×10^{-3})

予測フレーム数	MSE
1	7.05
2	12.79
4	36.36
9	68.29

6　終わりに

本研究では時系列予測モデルを導入した価値関数に基づく強化学習手法を提案した．今後は価値の推定時のみではなく，行動決定時にも予測した先の状態を用いる手法について検討する予定である．

参考文献

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, “Playing atari with deep reinforcement learning”, *NeurIPS*, 2013.

[2] C. J. Watkins, and P. Dayan, “Q-learning”, *Machine learning*, vol. 8, no. 3, pp. 279–292, 1992.

[3] R. J. Williams, “Simple statistical gradient-following algorithms for connectionist reinforcement learning”, *Machine learning*, vol. 8, no. 3, pp. 229–256, 1992.

[4] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, “Trust region policy optimization”, *Proc. of the 32nd Int’l Conf. on Machine Learning*, vol. 37, pp. 1889–1897, 2015.

[5] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms”, *arXiv preprint arXiv:1707.06347*, 2017.

[6] V. Konda, and J. Tsitsiklis, “Actor-critic algorithms”, *Advances in Neural Information Processing Systems*, vol. 12. MIT Press, 1999.

[7] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, “Asynchronous methods for deep reinforcement learning”, *Proc. of The 33rd Int’l Conf. on Machine Learning*, vol. 48, pp. 1928–1937, 2016.

[8] S. Fujimoto, H. van Hoof, and D. Meger, “Addressing function approximation error in actor-critic methods”, *Proc. of the 35th Int’l Conf. on Machine Learning*, vol. 80, pp. 1587–1596, 2018.

[9] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, “Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor”, *Proc. of the 35th Int’l Conf. on Machine Learning*, vol. 80, pp. 1861–1870, 2018.

[10] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, and S. Levine, “Soft actor-critic algorithms and applications.” *CoRR*, vol. abs/1812.05905, 2018.

[11] T. Haarnoja, A. Zhou, S. Ha, J. Tan, G. Tucker, and S. Levine, “Learning to walk via deep reinforcement learning.” *CoRR*, vol. abs/1812.11103, 2018.

[12] M. Laskin, A. Srinivas, and P. Abbeel, “Curl: Contrastive unsupervised representations for reinforce-

ment learning", *International Conference on Machine Learning*, pp. 5639–5650, 2020.

[13] D. Yarats, I. Kostrikov, and R. Fergus, "Image augmentation is all you need: Regularizing deep reinforcement learning from pixels", *International Conference on Learning Representations*, 2021.

[14] W. Lotter, G. Kreiman, and D. Cox, "Deep predictive coding networks for video prediction and unsupervised learning", *International Conference on Learning Representations*, 2017.

[15] X. Liang, L. Lee, W. Dai, and E. P. Xing, "Dual motion gan for future-flow embedded video prediction", *Proc. of the IEEE Int'l Conf. on Computer Vision (ICCV)*, Oct. 2017.

[16] S. Hochreiter, and J. Schmidhuber, "Long short-term memory", *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting", *Advances in neural information processing systems*, pp. 802–810, 2015.

[18] 西片 智広，山内 悠嗣，"動き情報を加えた PredNet による未来画像生成の高精度化"，画像センシングシンポジウム，2021.

[19] C. Finn, I. Goodfellow, and S. Levine, "Unsupervised learning for physical interaction through video prediction", *Advances in neural information processing systems*, vol. 29, pp. 64–72, 2016.

[20] D. Ha, and J. Schmidhuber, "Recurrent world models facilitate policy evolution", *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[21] L. Kaiser, M. Babaeizadeh, P. Mios, B. Osiski, R. H. Campbell, K. Czechowski, D. Erhan, C. Finn, P. Kozakowski, S. Levine, A. Mohiuddin, R. Sepassi, G. Tucker, and H. Michalewski, "Model based reinforcement learning for atari", *International Conference on Learning Representations*, 2020.