視覚強化学習における画像予測モデルを導入した顕著性誘導 Q ネットワーク

Integrating Image Prediction Model into Saliency-guided Q-Networks for Reinforcement Learning

○学　加藤　誉基 (中部大)　川出　拓誠 (中部大)　山内　悠嗣 (中部大)

Yoshiki KATO, Chubu University, tr24005-5173@sti.chubu.ac.jp
Takusei KAWADE, Chubu University, er21028-1764@sti.chubu.ac.jp
Yuji YAMAUCHI, Chubu University, yuu@isc.chubu.ac.jp

Reinforcement learning is an unsupervised learning method in which a type of agent learns optimal actions through interactions with the environment. It aims to improve the agent's decision-making by maximizing the value function. We proposed a framework that integrates an image prediction model into the value function, enabling long-term optimization by predicting future states. This study builds on previous research to enhance generalization performance in diverse environments. In this paper, we propose Saliency-Guided Q-Networks (SGQN), a reinforcement learning method that generates saliency maps based on value estimates and applies them to predicted images. By using saliency maps, important information in the image is emphasized while irrelevant information is suppressed leading to improved generalization performance. Experiments using the CartPole task demonstrate that our method achieves higher rewards at an earlier stage compared to SGQN without an image prediction model.

Key Words: reinforcement learning, image prediction, deep learning, saliency-guided Q-networks

1　はじめに

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，将来に亘って獲得できる報酬の期待値であるため，現在の状態から未知である先の状態の価値を推測している．西片等は，先の状態を予測できれば現在の状態より高い価値を求められるという発想から，先の状態を予測するモデルを価値関数に導入した [1]．これにより，将来の報酬を考慮した長期的な視点での最適化が実現した．

本稿では多様な環境における汎化性能を向上させるために先行研究 [1] を拡張する．本研究では，予測した画像に対して価値に基づく顕著性マップを作成し，価値の推定に利用する顕著性誘導 Q ネットワークによる強化学習を提案する．顕著性マップは，タスクを達成するために各画素の重要度を画像として表現したものである．顕著性マップが価値推定において画像中の重要な情報を強調し，不必要な情報の除去することが可能となるため，汎化能力の向上が期待できる．

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手法が精力的に研究されている．Deep Q Network(DQN)[2] は強化学習手法の 1 つである Q 学習 [3] と深層学習を組み合わせた手法である．Q 学習とは，価値関数により価値が最大とな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る．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 [4] や，価値ベースと方策ベースの両方を取り入れた actor-critic[5] 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)[6] 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するため，安定した学習が可能となる．

画像を入力とした視覚的強化学習に関する研究も盛んに行われている．Contrastive Unsupervised Reinforcement Learning (CURL)[7] は，画像から抽出する特徴量の表現能力を向上させるために対照学習を導入した強化学習手法である．Reinforcement Learning with Augmented Data(RAD)[8] と Data-regularized Q(DrQ)[9] は，データ拡張のみで CURL を上回った強化学習手法である．しかし，DrQ ではデータ拡張した画像を使用することでターゲットネットワークでは同じ状態であっても異なる価値が推定されるため，価値の推定が不安定になる問題を抱えている．この問題に対応するため，Stabilized Q-Value Estimation under Augmentation(SVEA)[10] は，ターゲットネットワークにデータ拡張前の画像を使用する方式に変更したことで安定した学習が可能となった．Soft Data Augmentation[11] は，自己教師あり学習の 1 つである Bootstrap Your Own Latent(BYOL)[12] に似たアプローチを採用し，拡張前後のデータから計算される特徴量の類似度が高くなるようにエンコーダを学習することで汎化性能を向上させた．

Saliency-uided Q-networks(SGQN)[13] は，観測画像に対して価値に基づく顕著性マップを作成し，価値の推定に利用する顕著性誘導 Q ネットワークを導入した手法である．顕著性マップは，Q ネットワークの出力に対する画像の各ピクセルの寄与度を表し，Q ネットワークの勾配情報に基づいて作成される．観測画像から生成された顕著性マップと，データ拡張によって背景を変更した画像から生成された顕著性マップとの誤差を用いてエンコーダとデコーダを学習することで，価値推定における重要な特徴量をエンコーダが抽出できる．これにより，顕著性マップが価値推定において重要な情報を強調し，不必要な情報を除去が可能になるため，エージェントは最も重要な情報に基づいて最適な行動を選択できる．Mask Distractions[14] は，観測画像に対してタスクに無関係な情報をフィルタリングする Masker ネットワークを導入することで汎化性能を向上させた．Masker ネットワークの学習では，価値関数の損失を用いて最適化され，エージェントの意思決定がタスクに関連する情報に基づくように誘導される．

2.2　予測画像生成に関する研究

観測した画像群から，次時刻以降に観測されるであろう画像を深層学習により予測する手法 [15, 16] 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)[17] に畳み込み処理を加えた畳み込み LSTM[18] がベースとなっている．予測画像の生成手法の 1 つである PredNet[15] は，畳み込み LSTM を重ね合わせた各層で次時刻以降の画像を予測し，各層の予測誤差を次の層へ伝搬させることで高精度な予測画像を生成する．

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予測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，高精度な予測画像を生成する手法 [19] が提案されている．また，Finn らは，ロボットハンド

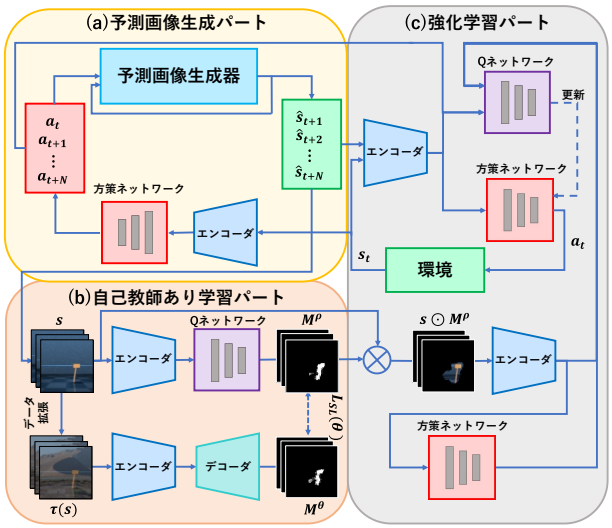

Fig.1 Flow of the proposed method.

が物体を押す，または引き寄せるといった動作を行う動画像，ロボットハンドの姿勢と動作の種類を入力とし，畳み込み LSTM により予測画像を生成する手法 Convolutional Dynamic Neural Advection(CDNA)[20] を提案している．LSTM ベースの手法 [15, 19, 20] では，長期的なフレーム間の関係を考慮した予測の処理が困難であり，勾配消失の問題や計算コストが課題となっている．これらの課題を解決するために，近年では Transformer[21] を用いた手法が注目されている．Transformer をベースとした手法 [22, 23] では，Multi-head Self-Attention(MHSA) を用いることで，過去の全てのフレームとの関係を考慮できるため，時系列の長さに関係なく，情報の劣化を抑えながら重要なフレーム間の関連性を学習できる．

3 提案手法

3.1 提案手法の概要

本稿では，視覚強化学習における画像予測モデルを導入した顕著性誘導 Q ネットワークを提案する．提案手法の特長は以下の通りである．

- 画像予測モデルと顕著性誘導 Q ネットワークの導入
  価値推定時に先の状態を予測した画像を考慮することで，現在の状態より高い価値を推定することが期待できる．さらに，本研究では予測した画像に対して価値に基づく顕著性マップを作成し，価値の推定に利用する顕著性誘導 Q ネットワークを導入する．これにより，価値推定において予測した画像の重要な情報を強調し，不必要な情報の除去することが可能となり汎化能力の向上が期待できる．
- 評価時における計算量
  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する．従って，計算コストが高い予測画像および顕著性マップの生成処理を行わない．そのため，評価時の計算量はベースとなる SAC[6] と同等となる．

図 1 に提案手法の流れを示す．提案手法は，図 1 (a) の予測画像生成パート，図 1 (b) の自己教師あり学習パート，図 1 (c) の強化学習パートの 3 つで構成される．まず，図 1 (a) に示すように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 を学習済みの予測画像生成器に入力し，1 フレーム先の予測画像 $\hat{s}_{t+1}$ を生成する．その後，生成した予測画像 $\hat{s}_{t+1}$ 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 フレーム先の予測画像 $\hat{s}_{t+2}$ を生成する．これを繰り返すことで N フレーム先の予測画像 $\hat{s}_{t+N}$ を生成する．次に，図 1 (b) に示すように時刻 $t+N$ までの画像から価値に基づく顕著性マップ生成する．この顕著性マップは閾値処理によりバイナリ化され，重要な領域を特定する顕著性マスクを生成する．生成した顕著性マスクと画像を掛け合わせることで，時刻 $t+N$ までのマスクした画像を生成する．そして，図 1 (c) に示すように，時刻 $t+N$ までの予測画像とマスクした画像をエンコーダを介して Q ネットワークに入力し，その得られる価値を用いて Q ネットワークと方策ネットワークの重みを更新する．

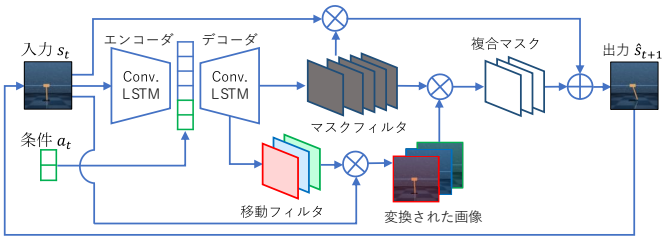

Fig.2 Overview of the CDNA.

3.2 予測画像生成

予測画像生成器には Convolutional Dynamic Neural Advection (CDNA)[20] 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 フレーム先の予測画像を生成する条件付き予測画像生成器である．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，観測されたオブジェクトの動きの変化を予測する移動フィルタを生成する．その後，入力された画像に移動フィルタを適用し，1 フレーム先の予測画像を生成する．

図 2 に CDNA の構成を示す．CDNA を構成するエンコーダ，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，得られた特徴量と時刻 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 で観測されたオブジェクトの動きの変化を捉える移動フィルタと，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 フレーム先の予測画像 $\hat{s}_{t+1}$ を生成する．CDNA は，生成した予測画像 $\hat{s}_{t+1}$ 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される．

3.3 Saliency-guided Q-networks(SGQN)

強化学習には画像ベースの強化学習手法である Saliency-guided Q-networks(SGQN)[13] を採用する．SGQN は，actor-critic の手法である SAC[6] をベースとしており，観測画像から顕著性マップを作成し，価値の推定に顕著性誘導 Q ネットワークを導入することで汎化性能を向上させた手法である．顕著性誘導 Q ネットワークの学習では，価値推定を通じて行動選択を最適化すると同時に，顕著性マップを用いて価値推定における重要な領域を学習する．

図 3 に予測画像を用いた SGQN の顕著性誘導 Q ネットワークの流れを示す．まず，観測した画像 s_t をエンコーダに通して特徴ベクトルを抽出し，Q ネットワークによって価値を推定する．同時に，Q ネットワークの出力に基づき勾配を計算し，Guided Backpropagation(GBP)[24] より，価値推定に寄与する顕著性を計算する．GBP では，入力画像の画素に関する価値の勾配を示したもので順伝播，逆伝播の勾配計算時に ReLU 関数を使用し，負の勾配を 0 に置き換えることで価値推定時に使用した特徴量を得ることができる．これにより，Q 値の計算に強く影響を与える画素を表す顕著性マップ M_t を生成する．この顕著性マップ M_t は閾値 ρ によりバイナリ化され，重要な領域を表現する顕著性マスク M_t^ρ として使用される．

次に元画像から生成された顕著性マスク M_t^ρ と，データ拡張によって背景を変更した画像 $\tau(s_t)$ から生成された顕著性マップ M_t^θ とのバイナリクロスエントロピー誤差 $L_{SL}\theta_t$ を用いてエンコーダとデコーダを学習することで，価値推定における重要な特徴量をエンコーダが抽出できる．

観測した画像 s_t から計算された価値 $Q(s,a)$ と観測した画像 s_t に顕著性マスク M_t^ρ を掛け合わせたマスク画像 $s_t \odot M_t^\rho$ から計算された価値 $Q(s \odot M^\rho, a)$ との平均二乗誤差を損失として Q ネットワークを学習する．SGQN における Q ネットワークの損失関数を式 (1) に示す．

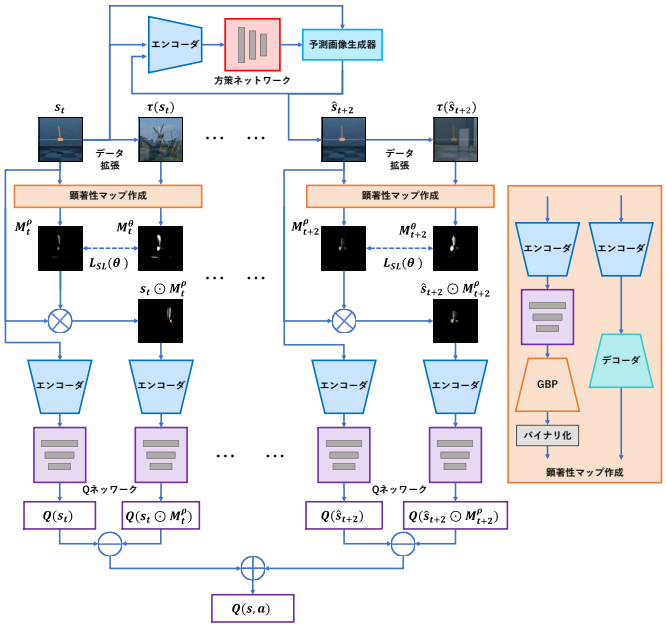

Fig.3 Flow of self-supervised learning in the SGQN.

$$L_{SQ} = L_1 + \lambda L_2 \tag{1}$$

$$L_1 = r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \tag{2}$$

$$L_2 = Q(s_t, a_t) - Q(s_t \odot M_t^{\rho}, a_t) \tag{3}$$

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，λ は各項の価値の重みを調整するハイパーパラメータを表す．式 (2) に示す SAC の損失は，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 $Q(s_t, a_t)$ の差分であり，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，式 (3) に示す顕著性誘導 Q ネットワークの損失は，価値推定において不要な情報を除去しつつ，重要な領域に着目した価値に基づくように学習できる．

3.4 画像予測モデルを導入した顕著性誘導 Q ネットワーク

画像予測モデルを導入した強化学習 [1] の損失関数は，式 (2) に予測した先の状態の価値 $Q(\hat{s}_{t+n}, a_{t+n})$ を加えた式 (4) により定義される．

$$L_3 = r_t + \frac{1}{2}\gamma\{Q(s_{t+1}, a_{t+1}) + \frac{1}{N-1}\sum_{n=2}^{N} Q(\hat{s}_{t+n}, a_{t+n})\} - Q(s_t, a_t) \tag{4}$$

ここで，$\hat{s}_{t+n}$ は予測画像生成器で生成した n 時刻先の予測画像である．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．

本研究では多様な環境における汎化性能を向上させるため，式 (4) の損失に予測した画像から顕著性マップを作成して価値の推定に利用する式 (3) の損失を導入する．提案手法の損失関数は，式 (6) により定義される．

$$\begin{aligned} L_4 = {} & \frac{1}{2}\{Q(s_t, a_t) - Q(s_t \odot M_t^{\rho}, a_t)\} \\ & + \frac{1}{4}\{Q(s_{t+1}, a_{t+1}) - Q(s_t \odot M_{t+1}^{\rho}, a_{t+1})\} \\ & + \sum_{n=2}^{N-1} \frac{1}{2^{n+1}}\{Q(\hat{s}_{t+n}, a_{t+n}) - Q(\hat{s}_{t+n} \odot M_{t+n}^{\rho}, a_{t+n})\} \\ & + \frac{1}{2^N}\{Q(\hat{s}_{t+N}, a_{t+N}) - Q(\hat{s}_{t+N} \odot M_{t+N}^{\rho}, a_{t+N})\} \end{aligned} \tag{5}$$

$$L_{PQ} = L_3 + \beta L_4 \tag{6}$$

ここで，β は各項の価値の重みを調整するハイパーパラメータである．損失関数の予測した価値に顕著性マップを用いることで，タスクを達成するために必要な領域を重視する効果が得られる．また，本手法は学習時のみ予測画像および顕著性マップを生成し，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．そのため，計算量は従来法と同等となる．

(a)学習時と同じ評価用背景　(b)Video easyレベルの評価用背景　(c)Video hardレベルの評価用背景

Fig.4 Example of cart pole task images.

4 評価実験

4.1 実験概要

提案手法の有効性を確認するためにカートポールタスクにより評価する．比較する手法は以下の通りである．

- SAC[6]
- 画像予測モデルを導入した強化学習 (式 (4) を採用)
- SGQN[13]
- 提案手法 (式 (6) を採用)

カートポールタスクでは，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，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．評価に用いる難易度の異なる画像の例を図 4 に示す．図 4(a) は学習時と同様の背景，図 4(b) は学習時の環境と異なる動画で構成されている Video easy レベルの背景，図 4(c) は学習時の環境と異なる動画に加え，地面や影も異なる動画で構成されている Video hard レベルの背景を示す．200,000 ステップを 1 回の学習とし，学習を 3 回した報酬の平均と標準偏差を比較する．

4.2 実験結果

図 5 にカートポールタスクにおける実験結果を示す．図 5(a) より，提案手法と画像予測モデルを導入した強化学習は SAC より早期に若干高い報酬を獲得できている．一方，図 5(b) と (c) より，どの評価用背景においても，20,000 ステップ辺りから SGQN は SAC よりも高い報酬を獲得していることが確認できる．

また，Video easy レベル，Video hard レベルにおいて，SGQN と比較して提案手法は早期に高い報酬を獲得している．これは，予測画像に対して価値に基づく顕著性マップを作成し，価値の推定に利用したことで，価値推定において予測した画像中の重要な情報を顕著性マップが強調し，不必要な情報が除去されたため，より正確な価値を推定できたと考えられる．顕著性誘導 Q ネットワークを導入した提案手法は将来の状態の予測において必要な情報のみを取り入れた意思決定が可能となる．これにより，将来の報酬に必要な情報のみを見据えた視点での最適化が促進される．

各手法における価値推定時の顕著性マップを比較する．顕著性マップの可視化には GBP を使用する．図 6 に顕著性マップを可視化した画像を示す．なお，6 フレームまでを画像予測モデルに入力し，7 及び 8 フレーム目の画像を予測している．図 6 より，従来法よりも，提案手法の方がカートポールの領域を注視していることが分かる．そのため，価値推定において重要な情報を強調し，不必要な情報を除去できた可能性が高いといえる．

5 結論

本稿では視覚強化学習における画像予測モデルを導入した顕著性誘導 Q ネットワークを提案した．提案手法では予測した画像に対して価値に基づく顕著性マップを作成し，価値の推定に利用することにより，価値推定において重要な情報を強調し，不必要な情報を除去したことで，異なる環境で高い報酬の獲得を実現した．今後は異なるタスクでの検証を行う予定である．

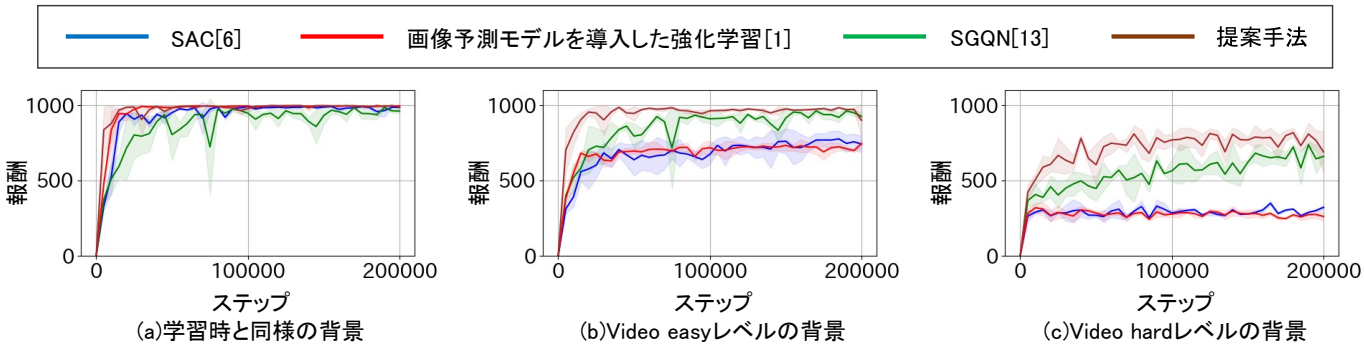

Fig.5 Rewards at each step in the line cartpole task. (a) Rewards obtained in the same background as in training. (b)-(c) Rewards in the evaluation background.

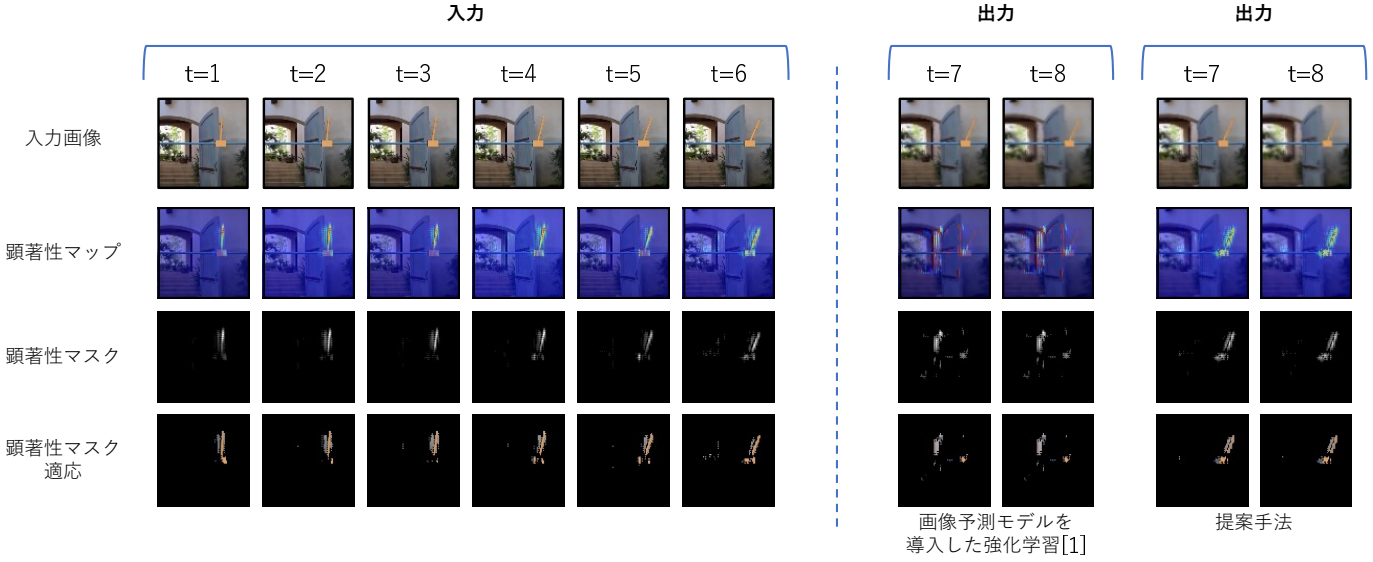

Fig.6 Comparison of saliency maps in the Cartpole task for each method in the Video hard level.

参考文献

[1] 西片等, “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”, 動的画像処理実利用化ワークショップ, 2023.

[2] V. Mnih *et al.*, “Playing atari with deep reinforcement learning”, *Neural Information Processing Systems in conjunction with Deep Learning Workshop*, 2013.

[3] C. J. Watkins *et al.*, “Q-learning”, *Machine learning*, vol. 8, pp. 279–292, 1992.

[4] J. Schulman *et al.*, “Proximal policy optimization algorithms”, *CoRR*, vol. abs/1707.06347, 2017.

[5] V. Konda *et al.*, “Actor-critic algorithms”, *NeurIPS*, vol. 12, 1999.

[6] T. Haarnoja *et al.*, “Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor”, *ICML*, vol. 80, 2018.

[7] M. Laskin *et al.*, “Curl: Contrastive unsupervised representations for reinforcement learning”, *International Conference on Machine Learning*, 2020.

[8] M. Laskin *et al.*, “Reinforcement learning with augmented data”, *NeurIPS*, vol. 33, pp. 19 884–19 895, 2020.

[9] D. Yarats *et al.*, “Image augmentation is all you need: Regularizing deep reinforcement learning from pixels”, *ICLR*, 2021.

[10] N. Hansen *et al.*, “Stabilizing deep q-learning with convnets and vision transformers under data augmentation”, *NeurIPS*, vol. 34, pp. 3680–3693, 2021.

[11] N. Hansen *et al.*, “Generalization in reinforcement learning by soft data augmentation”, pp. 13 611–13 617, 2021.

[12] J.-B. Grill *et al.*, “Bootstrap your own latent-a new approach to self-supervised learning”, *NeurIPS*, vol. 33, pp. 21 271–21 284, 2020.

[13] D. Bertoin *et al.*, “Look where you look! saliency-guided q-networks for generalization in visual reinforcement learning”, *NeurIPS*, vol. 35, pp. 30 693–30 706, 2022.

[14] B. Grooten *et al.*, “Madi: Learning to mask distractions for generalization in visual deep reinforcement learning”, *CoRR*, vol. abs/2312.15339, 2023.

[15] W. Lotter *et al.*, “Deep predictive coding networks for video prediction and unsupervised learning”, *ICLR*, 2017.

[16] X. Liang *et al.*, “Dual motion gan for future-flow embedded video prediction”, *ICCV*, 2017.

[17] S. Hochreiter *et al.*, “Long short-term memory”, *Neural computation*, vol. 9, pp. 1735–1780, 1997.

[18] S. Xingjian *et al.*, “Convolutional lstm network: A machine learning approach for precipitation nowcasting”, *NeurIPS*, 2015.

[19] 西片等, “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.

[20] C. Finn *et al.*, “Unsupervised learning for physical interaction through video prediction”, *NeurIPS*, vol. 29, pp. 64–72, 2016.

[21] A. Vaswani *et al.*, “Attention is all you need”, *NeurIPS*, vol. 30, 2017.

[22] X. Ye *et al.*, “Vptr: Efficient transformers for video prediction”, *ICPR*. IEEE, 2022, pp. 3492–3499.

[23] J. Zhu *et al.*, “Video frame prediction from a single image and events”, vol. 38, no. 7, pp. 7748–7756, 2024.

[24] J. T. Springenberg *et al.*, “Striving for simplicity: The all convolutional net”, *CoRR*, vol. abs/1412.6806, 2014.

[25] Y. Tassa *et al.*, “Deepmind control suite”, *CoRR*, vol. abs/1801.00690, 2018.