CF-009

Masked Modeling に基づく目標駆動型敵対的模倣学習
Goal-driven Adversarial Imitation Learning Based on Masked Modeling

五箇 亮太[1)]　前田 圭介[1)]　小川 貴弘[1)]　長谷山 美紀[1)]
Ryota Goka　Keisuke Maeda　Takahiro Ogawa　Miki Haseyama

1　はじめに

模倣学習は，専門家の行動データから方策を直接学習する手法であり，実世界の複雑なタスクにおける最適方策を獲得するための有望なアプローチの一つである．報酬関数を必要とする強化学習では，適切な報酬関数の設計はタスク依存性が高く専門的な知識を要するほか，報酬がスパースな環境や目標状態が曖昧な場合にはその設計自体が現実的に困難であるため，方策学習の実用性が大きく制限される．これに対し，模倣学習は明示的な報酬関数なしに学習可能であることから，ロボティクスや意思決定支援など多くの応用分野で注目されている．

近年，自然言語処理や画像処理の分野における Transformer アーキテクチャの成功を契機として，模倣学習を含む意思決定問題でも Decision Transformer [1] をはじめとする Transformer ベースの手法が数多く提案されている．特に，Bidirectional Encoder Representations from Transformers (BERT) [2] に代表される Masked Modeling に基づく自己教師あり学習は，状態・行動系列 (軌跡) の双方向時系列を考慮した高品質な潜在表現や目標駆動型の行動方策の獲得を可能とした [3–5]．しかしながら，自己教師ありの模倣学習手法では，専門家の軌跡の一部をマスクし，その欠損部分を教師信号として予測することを学習目標としているため，複数の有効な行動候補が存在する状況において，その多様性を反映した方策の学習が困難となる．そのため，モデルは平均的かつ曖昧な行動を出力する傾向にあるほか，従来の教師ありの模倣学習と同様に，学習時の状態分布と異なる状況下での頑健性が保証されず，推論時に専門家軌跡から逸脱する分布シフトの問題が顕在化しやすい [6]．さらに，Masked Modeling では，局所的な再構成精度には優れる一方で，系列全体の一貫性や因果的整合性を保証する設計にはなっておらず，モデルが系列全体として妥当な行動方策を獲得できるとは限らない．特に，長期的な依存関係や目標駆動型のタスクにおいては，整合的な軌跡の生成が困難となる．

そこで本研究では，Masked Modeling に基づく自己教師あり事前学習と敵対的模倣学習を統合した，新たな目標駆動型模倣学習手法を提案する．提案手法は，まず専門家の目標条件付き軌跡に対して Masked Modeling を適用し，欠損部分の予測を通じて各状態・行動の潜在表現を学習するとともに，再構成された軌跡と専門家の軌跡とを見分ける識別器を敵対的に学習させる．これにより，単なる局所的な復元ではなく，一貫性のある軌跡が生成されるような学習を可能とする．さらに，方策学習においても，敵対的模倣学習の枠組みを導入することで，エージェントが生成する目標条件付き軌跡が専門家の行動分布と整合するよう最適化を行う．事前学習では，マスク復元と敵対的識別によって高品質な潜在表現の獲得と系列整合性の向上を行い，方策学習では識別器を通じた分布整合性の保証を行う．以上より，二段階にわたる敵対的最適化を通じて，模倣精度と方策の汎用性を同時に高めることを目指す．提案手法は，Masked Modeling と敵対的学習を統合した新たな模倣学習フレームワークを提供し，多様な目標設定下でも汎用的かつ頑健な模倣学習が可能となる．

1)　北海道大学, Hokkaido University

2　関連研究

本章では，代表的な敵対的模倣学習手法の一つである Generative Adversarial Imitation Learning (GAIL) [7] について説明する．模倣学習の基本手法として広く用いられる Behavior Cloning [8] は，専門家の行動を教師あり学習で直接模倣するアプローチである．Behavior Cloning は，学習時と推論時で状態分布が乖離する分布シフトにより，未知の状態における誤差の蓄積が性能低下を招くという課題がある．また，この問題を回避するために提案された逆強化学習 [9] では，専門家の行動から報酬関数を推定し，強化学習によって方策を導出するため，高い計算コストと実装の複雑さが課題であった．これに対し，GAIL は，生成モデルである Generative Adversarial Network (GAN) [10] の枠組みに着想を得た模倣学習手法であり，専門家の軌跡分布と一致する方策を敵対的な訓練によって直接学習することを目的とする．GAIL では，状態・行動のペア (s,a) が専門家によるものか，学習中の方策によるものかを識別する識別器 $D : \mathcal{S} \times \mathcal{A} \to [0,1]$ と，その識別器を欺くことを目指す生成器の敵対的学習により方策 π の最適化が行われる．識別器 $D(s,a)$ は，(s,a) が学習者のものである確率を出力し，$1 - D(s,a)$ がそれを専門家によるものと判断する確率を表す．この枠組みにおける最適化は以下のミニマックス問題として定式化される．

$$\min_{\pi} \max_{D} \ \mathbb{E}_{(s,a)\sim\pi}\left[\log D(s,a)\right] + \mathbb{E}_{(s,a)\sim\pi_E}\left[\log\left(1 - D(s,a)\right)\right] \tag{1}$$

ここで，π_E は専門家の方策である．この目的関数は，GAN と同様に，方策関数である占有尺度 (occupancy measure) を専門家の分布に近づけるように学習を行う構造をもっている．識別器 D は，(s,a) が専門家からのものであれば $D(s,a) \to 0$，学習方策によるものであれば $D(s,a) \to 1$ となるように学習される．一方で方策 π は，識別器による判別を困難にし，すなわちすべての (s,a) に対して $D(s,a) \approx 0.5$ となるように更新される．実際の最適化では，固定した方策 π の下で識別器 D を更新し，次に固定した識別器 D の下で方策 π を更新するという交互最適化が行われる．このとき，識別器の出力を即時報酬 $r(s,a)$ とみなすことができるため，$r(s,a) = -\log D(s,a)$ と定義することができる．この報酬を用いて強化学習の枠組みで方策 π を最適化することで，エージェントは識別器の出力を低減させ (すなわち，専門家らしい行動を強化し)，結果として専門家の軌跡分布を模倣する方策が得られる．

Copyright © 2025 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

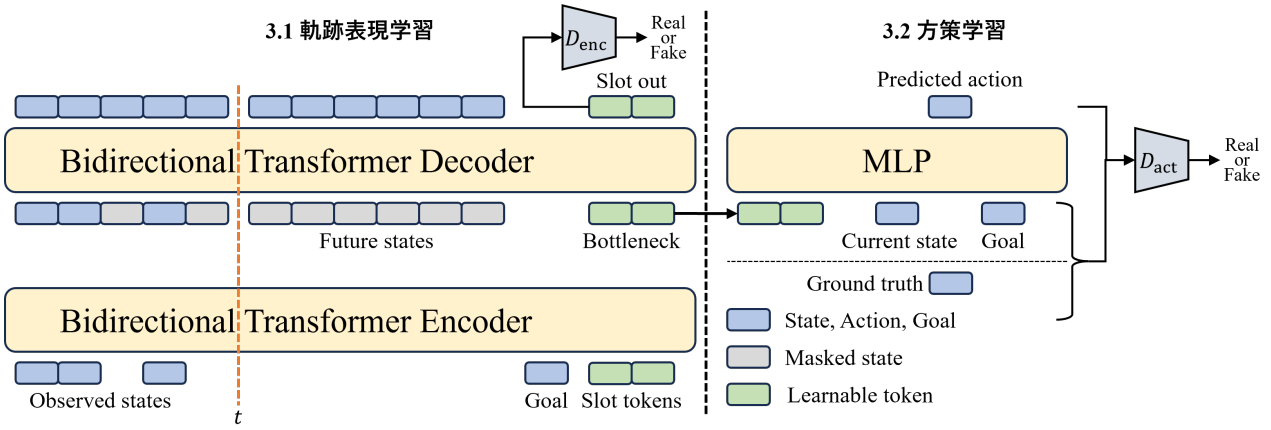

図 1: 提案手法の概要図.

3 提案手法

3.1 軌跡表現学習

本章では，Masked Modeling に基づく自己教師あり表現学習と敵対的模倣学習を統合した，目標条件付き模倣学習手法について説明する．提案手法の概要を図 1 に示す．本手法では，まず，専門家の軌跡から得られる高次の潜在表現の獲得を目的とした，Masked Autoencoder (MAE) [11] に基づく自己教師あり学習を行う．具体的には，入力された状態軌跡 $s = \{s_1, s_2, \ldots, s_T\}$ の一部をランダムにマスクし，観測された状態軌跡 $\tilde{s}$ および目標状態 g を線形射影することで得られた埋め込みを Slot token とともに Transformer encoder に入力する．Slot token は，観測された状態軌跡の意味的な構造を抽出するために導入され，Transformer encoder を通じて観測された状態の中から重要な情報と結びつくように学習される．また，軌跡中の情報を効率的に圧縮・抽出するために，Random-causal masking と呼ばれるマスク戦略を採用する．これは，過去の状態履歴に対してはランダムにマスクを適用し，将来の状態に対しては因果性を保った決定論的にマスクを適用することで，不完全な状態履歴に基づいた予測を可能とし，かつ長期的整合性を考慮した潜在表現の獲得を可能とする．Encoder は，観測され非マスク状態および Slot token を入力とし，その出力として系列全体の要約を含む Bottleneck 表現を獲得する．Decoder には，Bottleneck とマスクを含む元の状態軌跡 s が与えられ，欠損部分に対応する状態の再構成が行われる．この復元タスクを通じて，モデルは観測履歴と目標情報に基づき，将来の状態軌跡を再構成する能力を獲得する．さらに，軌跡全体の一貫性を保持するために，本手法では Decoder より得られた Slot out を入力とする識別器 D_{enc} を構築し，復元された軌跡に基づく Slot 表現と専門家の Slot 表現とを識別する判別タスクを通じて，Slot 表現の系列レベルでの構造的整合性と意味的正当性を保証するように学習を行う．

軌跡表現の学習において，生成器 (MAE) の損失関数 $\mathcal{L}^{\text{trl}}_{\text{gen}}$ を以下のように定義する．

$$\mathcal{L}^{\text{trl}}_{\text{gen}} = \mathcal{L}^{\text{trl}}_{\text{rec}} + \lambda_{\text{trl}} \cdot \mathcal{L}^{\text{trl}}_{\text{adv}} \tag{2}$$

ここで，$\mathcal{L}^{\text{trl}}_{\text{rec}}$ はマスクされた状態の再構成損失で，$\mathcal{L}^{\text{trl}}_{\text{adv}}$ は識別器を騙すことを目的とした敵対的損失であり，λ_{trl} はその重みである．各損失は，それぞれ次式により算出される．

$$\mathcal{L}^{\text{trl}}_{\text{rec}} = \frac{1}{|\mathcal{M}|} \sum t \in \mathcal{M} |\hat{s}_t - s_t|^2 \tag{3}$$

$$\mathcal{L}^{\text{trl}}_{\text{adv}} = -D_{\text{enc}}(z_{\text{fake}}) \tag{4}$$

ここで，$\mathcal{M}$ はマスクされた状態の集合であり，$\mathcal{L}^{\text{trl}}_{\text{rec}}$ はマスクされた状態 s^{mask}_t と予測された状態 $\hat{s}_t$ との平均二乗誤差により定義される．また，識別器 D_{enc} は，Slot 表現が復元軌跡に由来する場合には出力が低く，専門家軌跡に由来する場合には出力が高くなるように学習される．本研究では，この識別器に対して Slicing Adversarial Network (SAN) [12] に基づく Hinge loss および Wasserstein loss を組み合わせた損失を適用する．以上により，自己教師ありが学習に基づくマスクされた状態の復元と Slot 表現の識別を敵対的に学習することで，本手法は局所的な状態の再構成のみならず，軌跡全体として整合性の取れた潜在表現の獲得を実現している．これにより，後続の方策学習において，より構造化された軌跡表現を活用できるようになる．

3.2 方策学習

本節では，目標状態条件下における方策学習について説明を行う．提案手法では，MAE の Encoder より獲得される系列全体の Bottlenck 表現を，過去の観測履歴の要約情報として用いる．この Slot 表現に加え，直近の観測状態と目標状態を入力とすることで，現在の状況に応じた柔軟な行動選択を実現する方策モデルを構築する．本手法の方策学習は，自己教師あり学習と敵対的模倣学習を統合的に最適化する構成となっている．まず，Slot 表現は過去の状態軌跡の文脈的特徴を抽出した潜在変数として，履歴に基づく長期的依存性を捉えた行動選択を可能とする．本手法では，この Slot 表現に加え，直近の状態と目標状態を結合して入力することで，文脈と目標の両方を考慮した行動生成を実現する．さらに，生成された行動と状態および目標状態を入力とする識別器を導入し，専門家の行動軌跡と生成行動を識別する敵対的学習を行う．方策ネットワークは，識別器を騙すような行動を生成することで，専門家らしい振る舞いの模倣を促す．これにより，行動分布の整合性と目標適応性を兼ね備えた方策の獲得を可能とする．

本研究における方策学習の生成器の損失は以下のよう

Copyright © 2025 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

表 1: 提案手法および比較手法における Normalized scores (5 つのランダムシード値における平均 ± 標準偏差).

Dataset	BC	CQL	IQL	DT	WGCSL	CGIQL	DWSL	RvS-G	GCPC	Ours
umaze	$63.4_{\pm 9.4}$	$88.2_{\pm 2.3}$	$\mathbf{92.8}_{\pm 3.4}$	$55.6_{\pm 6.3}$	$90.8_{\pm 2.8}$	$\underline{91.6}_{\pm 4.0}$	$71.2_{\pm 4.2}$	$70.4_{\pm 4.0}$	$71.2_{\pm 1.3}$	$68.4_{\pm 7.8}$
umaze-diverse	$63.4_{\pm 4.4}$	$47.4_{\pm 2.0}$	$71.2_{\pm 7.0}$	$53.4_{\pm 8.6}$	$55.6_{\pm 15.7}$	$\mathbf{88.8}_{\pm 2.2}$	$\underline{74.6}_{\pm 2.8}$	$66.2_{\pm 5.6}$	$71.2_{\pm 6.6}$	$70.4_{\pm 5.4}$
medium-play	$0.6_{\pm 0.5}$	$72.8_{\pm 5.7}$	$75.8_{\pm 1.3}$	0	$63.2_{\pm 13.7}$	$\mathbf{82.6}_{\pm 5.4}$	$\underline{77.6}_{\pm 3.0}$	$71.8_{\pm 4.7}$	$70.8_{\pm 3.4}$	$75.9_{\pm 5.3}$
medium-diverse	$0.4_{\pm 0.5}$	$70.8_{\pm 10.3}$	$\mathbf{76.6}_{\pm 4.2}$	0	$46.0_{\pm 12.6}$	$76.2_{\pm 6.3}$	$74.8_{\pm 9.3}$	$72.0_{\pm 3.7}$	$72.2_{\pm 3.4}$	$\underline{76.4}_{\pm 3.4}$
large-play	0	$36.4_{\pm 10.3}$	$50.0_{\pm 9.7}$	0	$0.6_{\pm 1.3}$	$40.0_{\pm 16.2}$	$15.2_{\pm 7.7}$	$35.6_{\pm 7.6}$	$\underline{78.2}_{\pm 3.2}$	$\mathbf{83.0}_{\pm 4.8}$
large-diverse	0	$36.0_{\pm 8.3}$	$52.6_{\pm 5.9}$	0	$2.4_{\pm 4.3}$	$29.8_{\pm 6.8}$	$19.0_{\pm 2.8}$	$25.2_{\pm 4.8}$	$\underline{80.6}_{\pm 3.9}$	$\mathbf{88.0}_{\pm 4.7}$
Average	21.3	58.6	69.8	18.2	43.1	68.2	53.7	56.9	<u>74.0</u>	**77.0**

に定義される．

$$\mathcal{L}_{\text{gen}}^{\text{pl}} = \|\hat{a}_t - a_t\|_2^2 - \lambda_{\text{pl}} \cdot D_{\text{act}}(s_{1:t}, \hat{a}_t, g) \tag{5}$$

ここで，$\mathcal{L}_{\text{gen}}^{\text{pl}}$ は予測された行動と専門家の行動との間の平均二乗誤差による再構成損失，および識別器に対する敵対的損失の加重和で表現される．また，λ_{pl} は識別器に対する敵対的損失の重みである．識別器 D_{act} については，状態系列 $s_{1:t}$，生成行動 $\hat{a}_t$，および目標状態 g を入力とし，その入力が専門家の行動軌跡に由来するかどうかを判別するように学習される．本研究では，前節と同様に SAN に基づく損失関数を識別器に適用することで，学習の安定性と識別性能の向上を図る．以上により，生成側の方策ネットワークは，識別器を欺くように行動を生成することで，より専門家らしい振る舞いを獲得することが促される．提案手法は，Slot 表現を活用した行動生成と状態・行動・目標に基づく敵対的模倣学習を統合することで，目標に即した多様で一貫性のある行動の学習を可能としており，未知の目標や分布外の状態においても柔軟に対応できる方策の獲得を目指している．

4 実験

4.1 実験設定

本章では，提案手法の有効性を検証するために実施した実験について説明する．対象とするタスクは，オフライン模倣学習ベンチマークである D4RL [13] より提供される難易度の異なる複数の経路探索課題から構成される AntMaze タスクを用いた．Antmaze-umaze のほか，umaze-diverse，medium-play，medium-diverse，large-play，large-diverse といった異なる難易度および軌跡長で構成される 6 つの設定を対象とし，提案手法の頑健性を検証した．実験においては，学習率は 1×10^{-3}，バッチサイズは 1024，観測される状態履歴は 10，予測行動長は 70 とした．各試行は 5 つのランダムシードで繰り返し，最終的な評価は 100 エピソードのロールアウトに基づく Normalized Score の平均と標準偏差で評価した．マスク復元損失および敵対的損失の重み係数は，それぞれ $\lambda_{\text{trl}} = 0.01$，$\lambda_{\text{pl}} = 0.01$ とした．識別器はそれぞれ 2 層の全結合 (MLP) 層で構成され，活性化関数には ReLU を採用した．

また，以下の代表的なベースライン手法と比較することで，提案手法の有効性を検討した．

- Behavior Cloning (BC) [8]: 教師あり学習により専門家の行動軌跡を直接模倣する基本手法.
- Conservative Q-Learning (CQL) [14]: オフライン強化学習における Q 値の過大推定問題を抑制する手法.
- Implicit Q-Learning (IQL) [15]: 行動価値を間接的に最適化する戦略的なオフライン強化学習手法.
- Decision Transformer (DT) [1]: 軌跡をトークン系列として扱う Transformer ベースのオフライン強化学習手法.
- Weighted Goal-conditioned Supervised Learning (WGCSL) [16]: 目標状態への到達度に基づき重みを割り当てる目標条件付き模倣学習手法.
- Goal-conditioned IQL (CGIQL): 目標条件付きで IQL を拡張したオフライン強化学習手法.
- Distance Weighted Supervised Learning (DWSL) [17]: 状態軌跡予測との整合性に着目した目標付き模倣学習手法.
- RL via Surpervised Learning (RvS-G) [18]: 報酬付き系列生成に基づく条件付き模倣学習手法.
- Goal-Conditioned Predictive Coding (GCPC) [4]: 予測符号化に基づく目標付き自己教師あり模倣学習手法.

4.2 実験結果

表 1 の結果より，umaze では，IQL をはじめとする既存のオフライン強化学習手法が，提案手法よりも高い Normalized Score であることが確認された．この結果の要因として，Antmaze-umaze は短距離のゴール到達を要求する小規模な迷路で構成されているため，軌道の多様性が極めて限定的である．そのため，行動選択肢が実質的にほぼ一意に決定される状況が多く，模倣学習においても極めて単純な制御戦略で高い成功率が達成可能なタスクであると考えられる．CQL および IQL は，訓練データから直接価値推定を行うため，環境内の単一目標の達成行動に最適化されやすい構造を持つことに起因すると考えられる．提案手法は，Masked Modeling による自己教師あり学習と敵対的訓練による分布整合を同時に行う設計上，局所的な単純模倣に特化することなく軌道全体の構造や表現一般化を重視した学習を行う．そのため，極めて単純な目標の達成のみを求められるタスクでは，表現学習の恩恵が相対的に現れにくく，単純な方策最適化型手法に劣後する傾向にあったと考えられる．加えて，Masked Modeling タスクでは入力情報の一部を欠損させるため，学習中に完全情報を前提とする価値関数最適化手法に比べ，情報利用効率がやや低下する側面も存在する．そのため，単純な行動パターンを持つタスクにおいては逆に不利に働いた可能性も示唆される．

Copyright © 2025 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

一方で，large-play や large-diverse のようなより長距離かつ複雑な軌道遷移が求められるタスクにおいては，提案手法が他の手法を大きく上回る性能を示している．これらのタスクでは，目標状態への到達に至る経路の自由度が高く，状態遷移の長期的依存関係や軌跡全体としての整合性が学習の成否に大きく影響する．このような環境においては，Slot 表現による系列構造の抽出や，Masked Modeling を通じた予測的表現の獲得が効果的に機能し，行動生成において柔軟性と一貫性の両立を実現していると考えられる．また，識別器による敵対的損失を通じて，行動の分布整合性が促進されることで，専門家の軌跡に近い構造的な行動パターンが模倣され，複雑なゴール到達行動の再現性が高まったものと推察される．以上の結果から，提案手法は単純な模倣にとどまらず，目標指向の複雑なタスクに対しても汎用的に適応可能な模倣方策を学習可能であることが示唆される．

5 まとめ

本研究では，Masked Modeling を用いた自己教師あり表現学習と敵対的模倣学習を統合した新たな目標思考型模倣学習手法を提案した．本手法は，Masked Autoencoder を用いて状態軌跡から文脈的・構造的な表現を獲得し，それを用いて行動生成を行うとともに，生成行動の専門家らしさを識別器によって評価・最適化することで，系列レベルで一貫性と柔軟性を備えた方策の学習を可能とした．また実験により，提案手法は，単純な軌道パターンにおいては一部の方策最適化手法に劣るものの，より複雑かつ長期的な依存性を含むタスクにおいては顕著な性能向上を示した．

謝辞

本研究の一部は，JSPS 科研費 JP23K21676，JP23K11211，および JP25KJ0520 の助成により行われた．

参考文献

[1] L. Chen, K. Lu, A. Rajeswaran *et al.*, “Decision transformer: Reinforcement learning via sequence modeling,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 15 084–15 097, 2021.

[2] J. Devlin, M.-W. Chang, K. Lee *et al.*, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186.

[3] M. Carroll, O. Paradise, J. Lin *et al.*, “Uni [mask]: Unified inference in sequential decision problems,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 35 365–35 378, 2022.

[4] Z. Zeng, C. Zhang, S. Wang *et al.*, “Goal-conditioned predictive coding for offline reinforcement learning,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 25 528–25 548, 2023.

[5] K. Wen, Y. Hu, Y. Mu *et al.*, “M ^3 pc: Test-time model predictive control for pretrained masked trajectory model,” *arXiv preprint arXiv:2412.05675*, 2024.

[6] S. Ross, G. Gordon, and D. Bagnell, “A reduction of imitation learning and structured prediction to no-regret online learning,” in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2011, pp. 627–635.

[7] J. Ho and S. Ermon, “Generative adversarial imitation learning,” *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[8] D. A. Pomerleau, “Efficient training of artificial neural networks for autonomous navigation,” *Neural computation*, vol. 3, no. 1, pp. 88–97, 1991.

[9] B. D. Ziebart, A. L. Maas, J. A. Bagnell *et al.*, “Maximum entropy inverse reinforcement learning.” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2008, pp. 1433–1438.

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza *et al.*, “Generative adversarial nets,” *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[11] K. He, X. Chen, S. Xie *et al.*, “Masked autoencoders are scalable vision learners,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.

[12] Y. Takida, M. Imaizumi, T. Shibuya *et al.*, “San: Inducing metrizability of gan with discriminative normalized linear layer,” in *Proceedings of International Conference on Learning Representations*, 2024, pp. 1—34.

[13] J. Fu, A. Kumar, O. Nachum *et al.*, “D4rl: Datasets for deep data-driven reinforcement learning,” *arXiv preprint arXiv:2004.07219*, 2020.

[14] A. Kumar, A. Zhou, G. Tucker *et al.*, “Conservative q-learning for offline reinforcement learning,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 1179–1191, 2020.

[15] I. Kostrikov, A. Nair, and S. Levine, “Offline reinforcement learning with implicit q-learning,” *arXiv preprint arXiv:2110.06169*, 2021.

[16] R. Yang, Y. Lu, W. Li *et al.*, “Rethinking goal-conditioned supervised learning and its connection to offline rl,” *arXiv preprint arXiv:2202.04478*, 2022.

[17] J. Hejna, J. Gao, and D. Sadigh, “Distance weighted supervised learning for offline interaction data,” in *Proceedings of International Conference on Machine Learning*, 2023, pp. 12 882–12 906.

[18] S. Emmons, B. Eysenbach, I. Kostrikov *et al.*, “Rvs: What is essential for offline rl via supervised learning?” *arXiv preprint arXiv:2112.10751*, 2021.

Copyright © 2025 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.