摘要

长期以来，人工智能致力于赋予机器类同于人类的视觉感知与决策能力，基于视觉的强化学习应运而生。该范式融合计算机视觉的感知能力与强化学习的策略优化能力，构建端到端的感知与决策流程，既规避了手工设计状态特征的繁杂任务，又具备较强的环境适应性。然而，不同于传统视觉任务依赖静态数据集，基于视觉的强化学习的观测源自智能体与环境的持续交互，数据分布随环境状态与策略不断演化，形成具有强时序性与策略依赖性的动态视觉场景。此外，相较于监督学习依赖显式标签进行训练，强化学习通过环境返回的观测与奖励等稀疏经验信号进行学习。这种在数据结构与监督机制上的根本差异，使得动态视觉场景下的强化学习面临数据获取、表征学习、模型构建与决策优化等多层次的全链条挑战。

为应对上述挑战，本文系统性地开展动态视觉场景下的强化学习方法研究，聚焦于“如何在动态视觉环境下实现自适应决策”这一研究问题，采用自底向上的分层求解思路，从数据层、表征层、模型层以及决策层四个关键维度递进展开研究，逐层构建具有适应性的视觉强化学习方法体系，为自动驾驶与智能机器人控制等现实应用场景中高效、稳健的智能系统部署提供关键支撑。具体而言，本文在数据层提升输入多样性，增强表征泛化能力；在表征层建模时空关联，提炼动态环境特征；在模型层准确模拟状态转移，增强策略可预测性；在决策层融合多模态信息，最终实现复杂动态视觉场景下的高效自适应决策。本文取得的主要创新成果包括：

针对数据层观测分布差异大的问题，提出面向数据分布多样化的频谱随机掩码增强方法。该方法突破传统空间域数据增强可能会扭曲学习信号的局限，从全新的频域视角对观测数据展开增广操作。一方面，深入分析数据频谱特性后，设计出频谱随机掩码方法。此方法以全新的方式变换数据，在不破坏本质特征的情况下增加数据多样性，为模型学习提供更丰富信息。另一方面，基于增广不变性原理对价值函数和策略函数进行正则化处理，让模型面对不同分布数据时，能稳定学习价值估计与决策策略，增强鲁棒性和泛化能力。在机器人控制环境 DMControl-GB 中进行的实验表明，该方法对颜色或视频攻击具鲁棒性。在背景视频变化剧烈场景中，相较于国际前沿方法，该方法在 Walker-walk 任务的回合奖励提升 41.85%，证实其泛化性能得到提升。

针对表征层时空关联难建模的问题，提出面向时变空间表征增强的运动与外观协同交互方法。该方法突破现有混合纠缠运动外观的信息建模方式，首次在强化学习框架下凭借双路径架构精准刻画动作驱动的运动特征与观测关联的外观表征之间的交互。一方面，借助单独的网络路径分别学习运动和外观信息。运动路径通过相邻输入帧的帧差建模，外观路径聚焦单帧环境空间结构建模，并通过注意力引导的结构交互模块

获取运动与外观的互补信息，以此帮助智能体全面理解环境时空上下文。另一方面，针对样本效率低的问题，引入运动与外观一致性引导的好奇心模块，激励智能体探索学习不充分的观测。在机器人控制环境 DMControl 中的实验显示，该方法在 100K 步时回合奖励比国际前沿方法提升 7.63%，样本效率显著提高。

针对模型层模态状态转移不一致的问题，提出面向多模态环境模型精准转移的求同存异分解建模方法。该方法突破传统仅关注整体模态状态转移的建模方式, 创造性地将模态之间的共性和差异与环境动态建模过程深度融合。一方面，分解跨模态共性与差异：共性挖掘利用不同模态跨时间相互预测特性，使学习的共同信息契合环境动态；差异处理对不同模态不一致特征施加正交约束，防止同一模态特征过度正则。另一方面，构建模态感知驱动的动态建模机制：利用一致性特征预测跨模态未来状态，捕获共同场景动态；将不一致内容联合一致性特征送入不同模态的预测分支，推导各自的完整演变路径；同时，引入奖励预测函数过滤任务无关信息。在自动驾驶环境 CARLA 中的实验表明，该方法在正常和极端天气下平均累积奖励比国际前沿方法提升 11.1%，为全面且细致地理解复杂多变的环境提供了强大助力。

针对决策层模态异构优势协同不足的问题，提出面向多模态价值估计与策略学习优势强化的情境感知 Actor-Critic 方法。该方法突破传统“融合后行动”范式来进行多模态强化学习，独辟蹊径地提出一种“辨别后决策”的多模态自适应架构，从价值估计和策略学习的角度评估每个模态的个体效应及其集体相互作用。一方面，针对多模态价值冲突，引入双重一致价值估计方法，结合全局与局部模态价值估计，明确建模和协调不同模态的动作价值。同时施加自监督约束，缩小全局与局部模态的价值差距，提升估计稳定性。另一方面，针对单模态策略占主导引发的模态失衡问题，设计目标导向的渐进式策略学习方法，借多行动者机制保留各传感器独立决策，从各模态行动者与全局模态行动者采取的动作中筛选最优动作，逐步蒸馏至最终策略。在自动驾驶环境 CARLA 中的实验表明，该方法在极端天气下回合奖励比国际前沿方法提升 24.28%，凸显了其在动态场景下的自适应能力。

综上所述，本文围绕动态视觉场景下的自适应决策展开了深入探索，从数据、表征、模型以及决策这四个关键层面分别切入，通过频谱随机掩码方法、运动外观协同交互方法、求同存异分解建模方法以及情境感知价值估计与策略学习方法，改善数据分布、提炼时空表征，构建环境模型以及实现自适应价值估计与策略学习，并应用在自动驾驶任务和机器人控制任务上，有效提升了累积奖励与样本效率，为动态视觉场景下的自适应决策提供了切实可行的解决方案。

关键词：基于视觉的强化学习，数据增广，多模态学习，价值评估，策略学习

Research on Reinforcement Learning Methods in Dynamic Visual Scenarios

Yangru Huang (Computer Application Technology)
Supervised by Prof.Yonghong Tian

ABSTRACT

Artificial Intelligence research has long pursued human-like visual perception and decision-making in machines. Visual Reinforcement Learning (VRL) bridges this gap by integrating Reinforcement Learning (RL) with Computer Vision (CV), enabling end-to-end perception-to-action pipelines that bypass manual feature engineering and adapt across diverse environments. However, unlike traditional CV tasks using relatively stable static datasets, VRL data emerges from agent-environment interactions, producing dynamic scenarios with evolving, policy-dependent data. Additionally, compared to supervised learning guided by explicit supervision signals, RL depends on environmental feedback such as observations and rewards for training. These fundamental differences in data structure and supervision mechanisms pose multi-level, full-chain challenges in data acquisition, representation learning, model construction, and decision optimization within dynamic visual environments.

To address these challenges, this thesis systematically investigates reinforcement learning in dynamic visual scenarios, focusing on the core question: how to achieve adaptive decision-making in dynamic visual environments. A bottom-up, layered strategy is adopted, advancing from the data, representation, model, to the decision level, progressively building a VRL framework with adaptive capabilities. This work aims to provide key support for robust and efficient deployment of intelligent systems in real-world applications such as autonomous driving and robotic control. Specifically, this thesis enhances input diversity at the data level to improve representation generalization; models spatiotemporal correlations at the representation level to extract dynamic environmental features; accurately simulates state transitions at the model level to boost policy predictability; and integrates multimodal information at the decision level to enable efficient and adaptive decision-making in complex dynamic visual environments.

To address the significant observational distribution shift at the data level, this thesis proposes a Spectral Random Masking (SRM) method. Breaking through the limitations of spatial-domain data augmentation that may distort learning signals, this approach augments

observation from a novel frequency domain perspective. First, by conducting in-depth spectral analysis of data characteristics, we design a spectral random masking mechanism. This method can enhance diversity while preserving intrinsic features, thereby providing richer information for model learning. Second, leveraging the principle of augmentation invariance, we impose regularization constraints on both value and policy functions. This ensures stable learning of effective value estimations and decision policies across varying data distributions, significantly improving model robustness and generalization capabilities. Experimental results on the robotic control benchmark DMControl-GB demonstrate the method's resilience against color perturbations and video attacks. Particularly in scenarios with drastic background variations, it achieves a 41.85% improvement in episodic return for the Walker-walk task compared to state-of-the-art methods, validating its enhanced generalization performance.

To address the challenge of modeling spatiotemporal correlations at the representation level, this thesis proposes a Synergizing Interactive Motion-appearance Understanding (Simoun) method. Overcoming the limitations of existing approaches that lack explicit dynamic modeling mechanisms, this work pioneers a dual-path network architecture within RL frameworks to characterize interactions between action-induced motion features and observation-related appearance representations. First, the motion pathway analyzes frame differences across consecutive observation sequences to capture temporal dynamics, while the appearance pathway focuses on modeling environmental spatial structures. A novel structural interaction module is designed to enable bidirectional information exchange between pathways, allowing agents to holistically understand spatio-temporal contexts. Second, to mitigate low sample efficiency, we introduce a curiosity-driven exploration module guided by motion-appearance consistency constraints, which actively incentivizes exploration of under-learned observational patterns. Experimental evaluations on the DMControl robotic manipulation benchmark demonstrate the method's effectiveness, achieving a 7.63% improvement in episodic return compared to state-of-the-art baselines at 100K training steps, with particularly notable gains in sample efficiency.

To address the inconsistency in modality-specific state transitions at the modeling level, this thesis proposes a Dissected Dynamics Modeling (DDM) method. Breaking through traditional approaches that solely focus on holistic modality modeling, this work innovatively integrates the discovery of cross-modal commonalities and differences into environmental dynamics learning through deep fusion. First, explicit decomposition of cross-modal consensus and discrepancies is achieved: Commonality mining leverages cross-temporal predictability across modalities to ensure learned shared representations align with environmental dynamics, while difference processing applies orthogonal constraints between modality-specific inconsis-

tent features to prevent over-regularization of coherent and incoherent characteristics within individual modalities. Second, a modality-aware dynamic modeling mechanism is constructed: Consensus features drive cross-modal future state prediction to capture shared scene dynamics, while discrepancy components combined with consensus features feed into modality-specific prediction heads to derive complete modality-aware evolution trajectories. A reward-aware predictive function is further introduced to eliminate task-irrelevant information. Experiments in the CARLA autonomous driving simulator demonstrate the method's superiority, achieving an 11.1% improvement in average cumulative rewards over state-of-the-art approaches across normal and extreme weather conditions, providing robust support for comprehensive environmental understanding in complex scenarios.

To address modality-specific value-policy dominance conflicts at the decision-making level, this thesis proposes a Context-aware Dynamic Actor-Critic (ConDAC) framework. Departing from conventional "Fuse-and-Act" paradigms for multimodal RL, this work pioneers a "Discern-and-Decide" architecture that evaluates both individual modality contributions and collective interactions through value/policy perspectives. First, confronting multimodal value conflicts, we develop a dual-consistent value estimation mechanism that synergizes global value estimation with local modality-customized value modeling. This enables explicit coordination of cross-modal action valuations, reinforced by self-supervised constraints between global and local value estimates to minimize divergence and enhance stability. Second, addressing single-modality policy superiority, we devise a progressive policy distillation framework featuring a multi-actor system. This preserves modality-specific decision autonomy through dedicated actors while implementing optimal action selection from both individual and fused global policies via distillation. Experimental validation in the CARLA autonomous driving simulator demonstrates remarkable adaptability, achieving a 24.28% increase in episodic return over state-of-the-art methods under extreme weather conditions, with particularly enhanced performance in dynamic collision avoidance scenarios.

In conclusion, this thesis systematically investigates adaptive decision-making in dynamic visual scenarios through four pivotal dimensions: data, representation, modeling, and decision-making. By developing (1) the Spectral Random Masking method for data distribution generalization, (2) the Synergizing Interactive Motion-appearance Understanding framework for spatiotemporal representation learning, (3) the Dissected Dynamics Modeling approach for environment dynamics modeling, and (4) the adaptive multimodal Context-aware Dynamic Actor-Critic architecture for decision coordination, we holistically address critical challenges in VRL. Validated on autonomous driving and robotic manipulation tasks, these innovations

collectively achieve significant improvements in cumulative rewards and sample efficiency, establishing a comprehensive technical framework for robust adaptive decision-making in dynamically evolving visual environments.

KEY WORDS: Visual Reinforcement Learning, Data Augmentation, Multimodal Learning, Value Estimation, Policy Optimization