Simoun: Synergizing Interactive Motion-appearance Understanding for Vision-based Reinforcement Learning

Yangru Huang\textsuperscript{1}, Peixi Peng\textsuperscript{2,4 *}, Yifan Zhao\textsuperscript{1}, Yunpeng Zhai\textsuperscript{1}, Haoran Xu\textsuperscript{3,4}, Yonghong Tian\textsuperscript{1,2,4 *}

\textsuperscript{1}School of Computer Science, Peking University
\textsuperscript{2}School of Electronic and Computer Engineering, Peking University
\textsuperscript{3}School of Intelligent Systems Engineering, Sun Yat-sen University
\textsuperscript{4}Peng Cheng Laboratory

yzhuang@stu.pku.edu.cn, \{ppeng, zhaojf, yzhai, yhtian\}@pku.edu.cn, xuhz@mail2.sysu.edu.cn

Abstract

Efficient motion and appearance modeling are critical for vision-based Reinforcement Learning (RL). However, existing methods struggle to reconcile motion and appearance information within the state representations learned from a single observation encoder. To address the problem, we present Synergizing Interactive Motion-appearance Understanding (Simoun), a unified framework for vision-based RL. Given consecutive observation frames, Simoun deliberately and interactively learns both motion and appearance features through a dual-path network architecture. The learning process collaborates with a structural interactive module, which explores the latent motion-appearance structures from the two network paths to leverage their complementarity. To promote sample efficiency, we further design a consistency-guided curiosity module to encourage the exploration of under-learned observations. During training, the curiosity module provides intrinsic rewards according to the consistency of environmental temporal dynamics, which are deduced from both motion and appearance network paths. Experiments conducted on DeepMind control suite and CARLA automatic driving benchmarks demonstrate the effectiveness of Simoun, where it performs favorably against state-of-the-art methods.

1. Introduction

Reinforcement learning (RL) from visual signals has achieved great success in recent years. Compared with learning from hand-crafted states, vision-based RL eliminates the arduous task of designing states with manual feature engineering. Therefore, it is beneficial for a variety of tasks such as video game playing [25, 40], robot manipulation [50], and autonomous navigation [4]. However, one of the major challenges of vision-based RL lies in its high dimensional observation, which is less interpretable and leads to low rewards and inefficient sampling [19, 20]. As a result, robust visual understanding is crucial to bridge the gap between vision and state-based RL in terms of performance and sample efficiency.

To make observation comprehensible for agents, recent works have recognized the importance of learning high-quality visual features [20, 5, 29, 49, 42]. From the perspective of decision-making, two kinds of features are essential for vision-based RL: the motion features, which closely relate to the actions performed by the agents, and the appearance features, which describe the contemporary environmental states. Despite their importance, few works
have tried to model these two kinds of features explicitly. Most methods tend to encode single or stacked multiple frames via a convolutional encoder, as shown in Fig. 1(a) and Fig. 1(b). The former neglects motion feature modeling, while the latter tightly entangles the motion and appearance information into a single network, causing information bias and optimization difficulty. To alleviate these issues, Shang et al. [33] propose to encode only appearance features in single frames and impose motion dynamics by latent vector differences (termed latent flow, see Fig. 1(c)). In addition, several world models learned directly from input images are also developed [11, 31, 3]. Nevertheless, the motion features of these methods are still computed from spatial latent space, which may lead to sub-optimal representation learning in complicated environments.

To overcome the limitations described above, we present Synergizing Interactive Motion-appearance Understanding (Simoun) for vision-based RL. The design principle of Simoun is to explicitly learn both motion and appearance features in the early encoding stage and then interactively fuse them at the late stage for accurate decision-making. As shown in Fig. 1(d), given consecutive environment frames, Simoun learns motion and appearance information by separate network paths. The motion path (colored in blue) explicitly models motion clues (such as the speed and direction of cars) from the residual frame of multiple neighboring input frames. The appearance path (colored in orange) models the environmental spatial structures and focuses on identifying patterns and objects (such as cars and traffic lights) from every single frame. Additionally, a structural interactive module further extracts latent motion-appearance structures reflected by the correlations of the dual-path features. It then modulates both paths with the computed structure masks. In this way, each path can take complementary information from the other during learning, and agents are able to better understand the spatial and temporal context of the environment. Finally, the latent vectors from the two paths are fused for decision-making.

Although the dual-path design of Simoun promotes observation interpretability, low sample efficiency still exists due to limited data and sparse rewards. To alleviate this issue, a consistency-guided curiosity module is further designed. The idea is to adapt the learning process by concentrating more on under-learned observations, which can be deduced from the consistency of the motion and appearance paths. Intuitively, both the motion path and the appearance path describe the same observations. Hence, the dynamic information inferred from the following two sources should remain consistent: 1) latent motion vectors learned directly from the motion path and 2) differences between the appearance path latent vectors over multiple neighboring frames. If the opposite were true, then it indicates premature motion-appearance understanding, thus more exploration should be added. In this way, we build a strong correlation between reward discovery and state novelty. During training, the consistency-guided curiosity module provides intrinsic rewards in addition to extrinsic rewards from the environment, resulting in more efficient exploration.

Experiments on both CARLA and OpenAI DMControl environments show that the proposed method performs favorably against state-of-the-arts. Overall, the contributions of this paper are threefold:

1. We propose Simoun, a novel dual-path learning framework that explicitly and interactively learns both motion and appearance information from observations.

2. We design a structural interactive module to fully explore the complementarity of the two paths in Simoun and thereby further enhance visual understanding.

3. We devise a consistency-guided curiosity module to encourage the exploration of under-learned observations. The proposed module effectively increases sample efficiency by providing intrinsic rewards for the agents.

2. Related Works

Vision-based Reinforcement Learning To improve the performance and sample efficiency of vision-based RL, existing works can be roughly divided into three groups: 1) designing auxiliary loss/learning tasks [20, 14, 36, 49, 23], 2) employing various data augmentation techniques [21, 41, 2, 17, 24], and 3) modeling environment dynamics [12, 11, 13, 7, 27]. However, most existing works utilize a single-path network with multi-frame inputs, in which the motion and appearance information is tightly entangled without explicit separation. One exception is Flare [33], which also utilizes a single-path network to encode each frame individually and models motion information explicitly by taking latent vector differences. Although Flare achieves improved performance, its motion information is still computed from single-frame appearance features, causing insufficient temporal information extraction.

Dual-path Networks for Visual Modeling There is a rich literature of works on visual modeling with dual-path networks. One of the earliest works is the two-stream CNNs for action recognition [35], which utilizes a spatial stream with single-frame input and a temporal stream taking multi-frame optical flows. Thereafter, the concept of dual-path networks is heavily explored with various fusion strategies on different tasks [34, 6, 39, 43]. Our approach differs from existing ones in terms of task, architecture and learning mechanism. Compare with other visual tasks, vision-based RL needs to extract fine-grained motion details across different time steps, which poses a great challenge to RL models. To the best of our knowledge, Simoun is one of the first vision-based RL methods which aim to explicitly learn motion and appearance features by a dual-path architecture. Meanwhile, its structural interactive module is
also deliberately designed to extract motion and appearance clues residing in consecutive observation frames. By learning with RL-oriented objectives and curiosity-driven strategies, Simoun successfully imports the idea of dual-path networks into robust visual control. **Intrinsic Reward Exploration** One of the key elements of RL is the reward function, which aims to quantify the “goodness” of the agent’s decisions. However, the problem is that designing dense, well-defined extrinsic reward functions is difficult and unscaleable. One possible solution is to introduce an intrinsic reward function, which is calculated by the agents. Existing intrinsic reward exploration methods mainly focus on counting or predicting state novelty [1], prediction error [28], uncertainty [22], or environmental dynamics [32]. These methods are typically designed for the general state-based instead of vision-based RL. The works most relevant to us are CCFDM [26] and CCLF [37], which also formulate intrinsic rewards for visual-based RL. CCFDM utilizes forward dynamics and CCLF is based on a contrastive term. Differently, the consistency-guided curiosity module in Simoun exploits the dynamic consistency extracted from both motion and appearance features to achieve reliable intrinsic reward estimation. **3. Methodology**

We start by formalizing the task of vision-based RL in Sec. 3.1 and then delineate Simoun in detail. The general idea of Simoun is to 1) explicitly model the motion/appearance dynamics (and their correlations) of the agent’s operating ambiance and 2) capitalize on the learned dynamics to achieve efficient exploration. As shown in Fig. 2, our framework consists of two parts. First, given inputs in the form of consecutive frames, it interactively models the motion and appearance features using a dual-path architecture with a set of targeted objectives and a structural interactive module (Sec. 3.2). Then the decision-making strategy is learned with adaptive curiosity assignment steered by the proposed consistency-guided curiosity module (Sec. 3.3). The overall learning objective and inference process of Simoun are finally summarized in Sec. 3.4.

**3.1. Problem formulation**

Vision-based RL can be formulated as a Partially Observable Markov Decision Process (POMDP) \( \mathcal{M} = \langle \mathcal{O}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle \), where \( \mathcal{O} \) denotes the observation space containing pixel frames \( o_t \) at different time step \( t \) and \( \mathcal{A} \) denotes the action space. At each \( t \), the agent chooses an action \( a_t \in \mathcal{A} \). \( \mathcal{P}(o_{t+1}|o_t, a_t) \) is the observation transition, \( \mathcal{R}(o_t, a_t) \) is the reward function, and \( \gamma \in [0, 1) \) is the discount factor. The goal is to identify an optimal policy \( \pi \) that maximizes the expected cumulative reward based on the visual observations rather than the complete environment state:

\[
J(\pi) = \sum_t \mathbb{E}(o_t, a_t) \left[ R(o_t, a_t) \right].
\]

During training, \( \pi \) is used to interact with the environment, and the related data are stored in a replay buffer \( B \).

**3.2. Dual-path Interactive Modeling**

In vision-based RL, motion and appearance clues are both vital for the agent to perform accurate decision-
making. Simoun adopts a dual-path architecture to interactively learn motion and appearance features. Specifically, it consists of three components: the motion path, the appearance path, and the structural interactive module that serves as a switchboard between the motion and appearance structures learned from the two paths.

**Motion Path** Motion information is crucial for the visual system to understand the dynamics of the surrounding environment. The motion path aims to extract the motion-related features (such as the velocity of moving objects) from the changes between consecutive frames. Given a tuple of three adjacent observations $[\alpha_{t-2}, \alpha_{t-1}, \alpha_t]$ sampled from the replay buffer $B$, the input of the motion path is the residual of adjacent frames $[(\alpha_{t-1} - \alpha_{t-2}), (\alpha_{t} - \alpha_{t-1})]$, concatenated along channel dimensions. A convolutional encoder $E^n$ is then used to learn a lower-dimensional motion representation. Specifically, the encoder includes four convolution layers with $3 \times 3$ kernel size and ReLU non-linearity. Denote the feature map of the last convolution layer as $F^n_t \in \mathbb{R}^{c \times h \times w}$, a fully connected (FC) layer with layer normalization (LN) is used to reduce the dimension of $F^n_t$ to get the motion feature vector $f^n_t$.

To further impel the motion path to catch abundant environment temporal structures, a motion-aware constraint is deliberately specified. Specifically, given $f^n_t$ and corresponding action $a_t$ at timestep $t$, an action-conditioned two-FC-layer transition model $G$ is used to obtain a motion-action joint representation of the current timestep. A latent-space transition loss is then defined as:

$$\mathcal{L}_{\text{trans}} = \| G(f^n_t, a_t) - f^n_{t+1} \|_2^2,$$

which encourages $f^n_t$ to model the motion trends over time by predicting the future temporal dynamics $f^n_{t+1}$ of the next time step.

**Appearance Path** The appearance path is designed to extract the visual appearance representations of the object and scene from individual observation frames. The input of the appearance path is a single frame $\alpha_t$ from the replay buffer $B$. Similar to the motion path, the appearance path also adopts a four-layer encoder $E^a$ with the same architecture (except for the number of input channels of the first layer) to get feature map $F^a_t \in \mathbb{R}^{c \times h \times w}$, followed by FC and LN layers to get the appearance representation $f^a_t$.

To explicitly make appearance representations discriminative to different scenes, we take inspiration from CURL [20] and adopt an unsupervised contrastive loss between similar and dissimilar sample pairs. Given observation frame $\alpha_t$, we consider $\bar{\alpha}_t$ (obtained from $\alpha_t$ by data augmentation) as the positive sample, while samples coming from different observation frames are regarded as negatives. The contrastive loss is then defined as:

$$\mathcal{L}_{\text{con}} = -\log \frac{\exp(f^a_t \top f^a_{\bar{\alpha}})}{\exp(f^a_t \top f^a_{\bar{\alpha}}) + \sum_{i \neq \bar{\alpha}} \exp(f^a_t \top f^a_i)}.$$  (3)

where $f^a_t$ and $f^a_{\bar{\alpha}}$ are appearance representations of $\alpha_t$ and $\bar{\alpha}_t$, respectively. $j$ is the sample index in a training batch with $K$ samples. Note that the augmented data is only used for representation learning and is inaccessible during decision-making. This is to avoid unstable training, which is easily caused by the augmented data dominating the evaluation of the Q-target.

**Structural Interactive Module** Instead of simply concatenating the features from motion and appearance paths as state representations, interactively communicating the learned motion-appearance structures between both paths can lead to a more robust visual understanding. Intuitively, as shown in Fig. 3 (left), given consecutive observation frames, the motion structure can be revealed by inter-frame appearance variation. Meanwhile, the appearance structure can be obtained by spatial pixel correlation. Therefore, the structural interactive module embraces a single and inter-frame relation discovering mechanism for efficient motion-appearance structure mining and propagation.

Specifically, as shown in Fig. 3 (right), the inputs of the structural interactive module are motion feature map $F^n_t$ and appearance feature maps $F^a_{t-2}$, $F^a_{t-1}$ and $F^a_t$. To extract motion structures, an inter-frame attention mechanism is designed. In particular, $F^a_{t-1}$ is treated as the query, and $F^a_{t-2}$ and $F^a_t$ are used as two keys. A inter-frame attention map $X$ can be obtained as:

$$X = \sigma(\sigma(F^a_{t-2} \top \bar{F}^a_{t-1}) + \sigma(F^a_{t-1} \top \bar{F}^a_{t-1})), \quad (4)$$

where $\bar{F}^a_t \in \mathbb{R}^{c \times h \times w}$ denotes a new feature map generated by feeding $F^a_t$ to a convolution layer for reducing the spatial complexity ($F^a_{t-1}$ and $F^a_{t-2}$ can be obtained similarly). $\sigma$ denotes the Softmax function. Combining the outputs of two initial Softmax functions might not yield values sum-
Consistency-guided Curiosity Module

Figure 4. The idea of consistency-guided curiosity module.

The soft indexing of $\tilde{F}_{t-1}^{m}$ by $\tilde{F}_{t-2}^{m}$ or $\tilde{F}_{t}^{a}$ is performed at the spatial dimensions, resulting in a soft attention map of dimensions $hw \times hw$. Then a spatial gating operation is applied on $X$ using weights calculated from both $\tilde{F}_{t}^{m}$ and $\tilde{F}_{t}^{m}$ to consolidate appearance structures and obtain two motion-appearance structure masks:

$$X^{m} = X \cdot \tilde{F}_{t}^{m}, \quad A^{a} = X \cdot \tilde{F}_{t}^{a},$$

where $\tilde{F}_{t}^{m}$ is a feature map of $F_{t}^{m}$ obtained with another convolution layer. Finally, $F_{t}^{m}$ and $F_{t}^{a}$ are enhanced by modulating with their corresponding motion-appearance structure masks:

$$F_{t}^{m} = F_{t}^{m} + \beta \cdot X^{m}, \quad F_{t}^{a} = F_{t}^{a} + \beta \cdot X^{a},$$

where $\beta$ is a learnable parameter initialized as zero [48], and the enhanced $F_{t}^{m}$ and $F_{t}^{a}$ instead of the original ones are further used to calculate $f_{t}^{m}$ and $f_{t}^{a}$.

After acquiring the structure-enhanced motion and appearance features, they are finally fused as $f_{t} = [f_{t}^{m}, f_{t}^{a}]$, where $f_{t}$ is used as the final state representation.

To better capture reward-related features from the fused state representation, a reward function $R$ is further introduced to predict a numerical reward value to each state-action pair. $R$ has a similar architecture with the transition model $G$ except the output dimension is set to one. A reward loss $L_{r}$ is then defined at the mean squared error between the predicted and actual reward:

$$L_{r} = (R(f_{t}, a_{t}) - r_{t+1}^{*})^2,$$

where $r_{t+1}^{*}$ is the actual external reward value at the next time step, which is returned by the environment.

3.3. Consistency-guided Curiosity Module

Given state representation $f_{t}$ from the dual-path model, we adapt SAC [9, 10] as the base RL algorithm following previous methods [47, 5], which aims to maximize the expected cumulative reward to find an optimal policy by approximating the action-value function $Q$ and a stochastic policy $\pi$ based on a $\alpha$-discounted maximum entropy $H(\cdot)$:

$$J(\pi) = \sum_{t} E(\omega_{t}, a_{t}) \sim \pi[r(\omega_{t}, a_{t}) + \alpha H(\pi(\cdot|\omega_{t}))].$$

The action-value function $Q$ are learned by minimizing the soft Bellman error:

$$L_{Q} = E(\omega_{t}, a_{t}) (Q(\omega_{t}, a_{t}) - (r_{t} + \lambda V(\omega_{t+1}))^2, \quad (9)$$

and the soft state value $V$ can be estimated by sampling an action under the current policy:

$$V(\omega_{t+1}) = E_{a_{t+1} \sim Q}[Q(\omega_{t+1}, a_{t+1}) - \alpha \log \pi(a_{t+1}|\omega_{t+1})], \quad (10)$$

where $Q$ denotes the exponential moving average of the parameters of $Q$. The policy is optimized by decreasing the difference between the exponential of the soft-Q function and the policy:

$$L_{\pi} = \mathbb{E}_{\omega_{t} \sim \pi} \left[ \alpha \log \pi(a_{t}|\omega_{t}) - Q(\omega_{t}, a_{t}) \right].$$

Although SAC algorithm introduces the entropy term to encourage exploration, it still heavily relies on carefully engineered environmental extrinsic rewards and suffers from low sample efficiency. Instead of extrinsic rewards, intrinsic curiosity can be a powerful concept to endow an agent with an automated mechanism to continuously explore its environment in the absence of task information. During training, in addition to the extrinsic rewards $r^{*}$ obtained from the environment, the curiosity module also provides intrinsic rewards $r^{i}$. The learning objective SAC in Eq. 8 is therefore extended as:

$$J(\pi) = \sum_{t} E(\omega_{t}, a_{t}) \sim \pi[r^{*}(\omega_{t}, a_{t}) + \alpha H(\pi(\cdot|\omega_{t})) + r^{i}(\omega_{t}, a_{t})].$$

To promote proper intrinsic rewards $r^{i}$, we propose to leverage the consistency of motion and appearance features from the dual-path model. Intuitively, as shown in Fig. 4, the environmental temporal dynamics can be obtained from two sources: 1) directly learned from the motion path, and 2) through the variations of the spatial features learned from the spatial path, similar to the latent flow [33]. Optimally, these two sources should be consistent with each other. That is, the motion feature $f_{t}^{m}$ should be similar to the difference between the spatial features $f_{t-2}^{a}$, $f_{t-1}^{a}$ and $f_{t}^{a}$. The idea of the consistency-guided curiosity module is then to encourage the agent to explore when the temporal dynamics produced by the two sources are inconsistent. To this end, the intrinsic reward $r^{i}$ is defined as:

$$r^{i} = Ce^{-d^{2}} d(|f_{t}^{m}|, |f_{t-1}^{a} - f_{t-2}^{a}| + |f_{t}^{a} - f_{t-1}^{a}|) \frac{\max_{r^{i}}}{\min_{r^{i}}},$$

where $C$ is temperature weight, $\lambda$ is an exponential decay weight, $d$ is the L2 distance function, $t$ is the environment.
Algorithm 1 Inference procedure of Simoun

1: for each environment step $t$ do
2:   Collect observation frames $o_{t-2}$, $o_{t-1}$ and $o_t$.
3:   Extract motion and appearance feature maps:
4:   $F^m_t = \mathcal{E}^m([o_{t-1} - o_{t-2}, (o_t - o_{t-1})])$,
5:   $F^a_j = \mathcal{E}^a(o_j), j = t - 2, \ldots, t$
6:   Structural Interactive Modeling:
7:   $X = \sigma(\gamma(F^m_{t-2} F^a_{t-1}) + \sigma(F^a - F^m_{t-1}))$
8:   $F^m_t = F^m_t + \beta \cdot (X \cdot F^m_t)$
9:   $F^a_t = F^a_t + \beta \cdot (X \cdot F^a_t)$
10: Extract low-dimensional features:
11:   $f^m_t = \text{LayerNorm}(\text{FC}(F^m_t))$
12:   $f^a_t = \text{LayerNorm}(\text{FC}(F^a_t))$
13: Take action based on the fused features:
14:   $a_t = [f^m_t, f^a_t]$
15:   $o_{t+1} \sim P(o_{t+1} | o_t, a_t)$
16: Environment state transition:
17: end for

The representation learned by Simoun and other observation encoding paradigms from an information bias perspective to gain in-depth understanding of them.

Specifically, we leverage a recent approach [18] that can quantify static and dynamic information learned by any spatial-temporal model. The approach estimates the amount of static vs. dynamic bias based on the mutual information between sampled input sequence pairs. Then it calculates the percentage of units (channel dimensions) of the model feature layer that encodes several pre-defined information factors (static, dynamic, joint, and residual). The quantifying results of Simoun and other models are illustrated in Fig. 5. It is immediately evident that previous methods tend to encode the static factor more than the dynamic, indicating a strong bias toward appearance information. It is also clear that the two paths of Simoun learn corresponding static and dynamic factors as expected, with the motion path having more dynamic units than the appearance path. By fusing the two paths together, Simoun enjoys abundant dynamic and static information, meanwhile having minimal residual units that do not involve any dynamic or static factor. As will be shown in the experiment section, such an abundant and informative representation significantly improves the decision-making process.

5. Experiments

In this section, we explore how Simoun can improve vision-based RL in terms of sample efficiency and performance gains. Two benchmarks are used for evaluation: DeepMind Control Suite (DMControl) [38] for continuous control and CARLA [4] for autonomous driving.

Experimental Settings: Simoun is implemented on the basis of the SAC algorithm [10]. For DMControl, to avoid the potential effects of different hyperparameters, we follow the previous training setup of DrQ [44] and choose...
six commonly adopted tasks: Walker-walk, Finger-spin, Cartpole-swingup, Reacher-easy, Cartpole-run and Ball in cup-catch. To evaluate sample efficiency, the performance at 100k and 500k environment steps are reported during the training stage. For CARLA, we mostly follow the setup of DBC [47], where the goal is to travel as far as possible on Highway 8 of Town 4 in 1000 time steps without any collisions with 20 moving cars. For observation acquisition, we horizontally concatenate the images from three cameras on the roof of vehicles to get 84 × 252 images. Random convolution [21] is adopted for data augmentation in Eq. 3. All experiments are trained across 5 random seeds to report the mean and standard deviation of the rewards. More details can be found in the Supplementary Material.

Methods Compared: We extensively compare Simoun with a variety of methods including the SAC [10] baseline, explicit motion modeling approach (Flare [33]), auxiliary loss-based approaches ( CURL [20], MLR [46]), data augmentation approaches (DrQ [44], SVEA [15]), dynamic modeling based approaches (Dreamer [11], SPR [30], PlayVirtual [45], DeepMDP [8]), and curiosity based approach (CCLF [37]).

5.1. Performance Comparisons

Results on DMControl Table 1 presents the experimental results on the DMControl benchmark. It can be observed that Simoun achieves considerable performance gains compared to other state-of-the-art methods. In particular, at 100k steps, significant performance improvement can already be reached by Simoun, which indicates improved sample efficiency.

Results on CARLA The results of the CARLA benchmark are reported in Table 2. It is clear that Simoun outperforms all other methods on the episode reward. Additionally, the average driving distance is farther than other methods by a large margin and the average collision intensity is also smaller. Although the driving smoothness of Simoun is slightly decreased due to increased steer and brake, this small cost has led to considerable overall reward gain to break through the current status quo.

5.2. Ablation Study

Effectiveness of the Dual-path Design To demonstrate the effectiveness of the dual-path design in Simoun, we compare it with three single-path methods (depicted in Fig. 1): individual frame encoding, stacked frames encoding, latent flow encoding, and our dual-path encoding. For a fair comparison, the motion and appearance losses of Simoun are adopted for all four models to eliminate the affection of loss difference, and the interactive and curiosity modules are also removed from the dual-path model. The results are shown in Fig. 6 (left). Several observations can be made: 1) The low performance of individual frame encoding (black line) indicates the importance of modeling motion information. 2) By considering the motion across frames, stacked frame encoding (green line) performs much better than individual frame encoding. 3) The latent flow encoding (blue line) improves over stacked frame encoding
with explicit motion modeling. However, there is limited room for further improvement due to the preliminary technique used for motion extraction. 4) The proposed dual-path model (red line) remarkably outperforms the other three methods by modeling motion and appearance explicitly. Interestingly, the results in Fig. 6 (left) echo perfectly with the information bias degrees illustrated and discussed in Fig. 5 and Sec. 4, which indicates potentially deeper connections between motion-appearance information bias and the performance of decision-making.

**Effectiveness of Each Components in Simoun** To investigate the contribution of each component in Simoun, we first evaluate the performance of each individual path, then test the dual-path model by gradually adding the structural interactive module and the consistency-guided curiosity module. It can be found in Fig. 6 (middle) that the individual path gives relatively low performance when trained separately, with the motion path performing better than the appearance path. When adopting the dual-path model, both the structural interactive module and the consistency-guided curiosity module can further improve performance, which demonstrates their effectiveness. To better understand which visual clues does Simoun concentrate, we visualize the motion-appearance structure masks (X^m and X^a in Eq. 5) of the two paths. As can be observed from Fig. 7, the motion path tends to focus on the moving trajectory of other vehicles (positions where the vehicles passed by), while the appearance path focuses strongly on spatial positions where the other vehicles exist.

**5.3. Further Discussions**

Does the performance gain of the dual-path model come from the increased network parameters? To answer this question, we compare the dual-path model with a double-channeled stacked frame model, which has nearly 2× more parameters. From Fig. 6 (right) we can observe that increasing model parameters indeed improves performance. However, the dual-path model still outperforms the double-channeled stacked frames model using only half of its parameters. This proves the effectiveness of the dual-path model mainly comes from its motion-appearance modeling paradigm, rather than increased network capacity.

Does Simoun improves domain generalization? To evaluate the domain generalization ability of Simoun, we select the “catch” task on the “ball in cup” scenario as the source domain and test Simoun on DMControl Generalization Benchmark [16] with four different environment domains (color easy, color hard, video easy, and video hard). Fig. 8 shows that Simoun performs on par with other methods on color-shifted domains. However, the performance is much better on video-shifted domains, where the background is also moving. This is attributed to the specialized modeling of motion information by Simoun, which drives the agent to pay more attention to reward-related motions rather than the irrelevant dynamic background.
6. Conclusion

We have proposed Simoun, a unified framework for vision-based RL with a dual-path network for motion and appearance understanding. The design of Simoun demonstrates the effectiveness of motion-appearance structural interaction, and further shows the benefits of consistency-guided intrinsic curiosity. Empirical results suggest that the proposed method has advantages in terms of sample efficiency, performance gains, and generalization ability. By analyzing Simoun from an information bias perspective, we build a connection between motion-appearance information bias and vision-based RL performance. We hope this connection can further inspire more efficient model designs for vision-based RL tasks.

7. Acknowledgement

The study was funded by the Key-Area Research and Development Program of Guangdong Province with contract No. 2020B0101380001, as well as the National Natural Science Foundation of China under contracts No. 62027804, No. 61825101, No. 62088102 and No. 62202010. Computing support was provided by Pengcheng Cloudbrain.

References


