周报_20250303

目录

《Salience-Invariant Consistent Policy Learning for Generalization in Visual Reinforcement Learning》

Jingbo Sun
Institute of Automation, CASIA
Beijing, China
Pengcheng Laboratory
Shenzhen, China
School of Artificial Intelligence, UCAS
Beijing, China
sunjingbo2022@ia.ac.cn

Songjun Tu
Institute of Automation, CASIA
Beijing, China
Pengcheng Laboratory
Shenzhen, China
School of Artificial Intelligence, UCAS
Beijing, China
tusongjun2023@ia.ac.cn

Qichao Zhang
Institute of Automation, CASIA
Beijing, China
School of Artificial Intelligence, UCAS
Beijing, China
zhangqichao2014@ia.ac.cn

Ke Chen
Pengcheng Laboratory
Shenzhen, China
chenk02@pcl.ac.cn

Dongbin Zhao
Institute of Automation, CASIA
Beijing, China
School of Artificial Intelligence, UCAS
Beijing, China
dongbin.zhao@ia.ac.cn

在视觉强化学习（visual reinforcement learning，VRL）中，如何将所学策略推广到从未见过的新场景始终是一个关键挑战。因为训练环境的视觉观测往往与测试环境存在差异，导致智能体可能会过度拟合训练环境的特定视觉信息。例如，在未见过的环境中，一些无关任务的像素干扰会使智能体提取到与任务无关的特征，进而偏离训练时学到的最优行为，削弱了泛化能力。为了解决这一问题，本文提出了一种面向零样本（zero-shot）泛化的高效算法框架，称为 Salience-Invariant Consistent Policy Learning (SCPL)。该方法新引入了一个“价值一致性模块”（value consistency module）以及一个“动态模块”（dynamics module），以便更好地提取与任务相关的表征。具体来说，价值一致性模块在显著性（saliency）引导下，能保证智能体在原始观测与扰动后观测两种情形下都只关注到与任务紧密相关的像素；动态模块则利用图像增强后的数据来帮助编码器捕捉到与动力学和回报相关的特征。除此之外，本文还给出了理论分析，表明在策略一致性对于泛化十分重要。为此，本文提出了一个带有 KL 散度约束的“策略一致性模块”，让原图与扰动后图像在决策分布上保持一致，从而在新环境中也能延续训练时学到的良好策略。本文在 DMC-GB、Robotic Manipulation 以及 CARLA 三大基准测试上做了大量实验，结果表明，SCPL 在泛化性能上远超已有最先进方法。尤其是在 DMC 的 video hard 设置、Robotic 的 hard 设置以及 CARLA 基准上，平均表现分别提升了 14%、39% 和 69%。

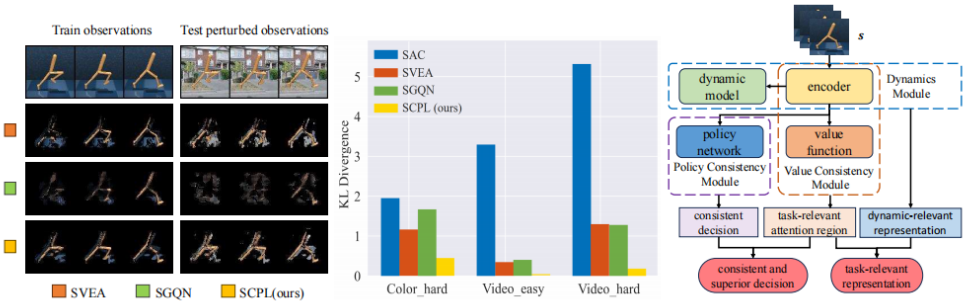

Figure 1: (Left) Saliency masked map of SVEA, SGQN, and SCPL (ours), which shows the attention regions of value functions on the DMC-GB benchmark. (Middle) The KL divergence of action distribution between training and test environments on DMC-GB, where our method holds the smallest KL divergence. (Right) Contribution overview of SCPL, which aims to improve visual generalization by achieving task-relevant representations and consistent and superior decisions.

SCPL 概述

基于 SAC 进行改进。

新增三个模块：价值一致性模块、动态模型模块、策略一致性模块。

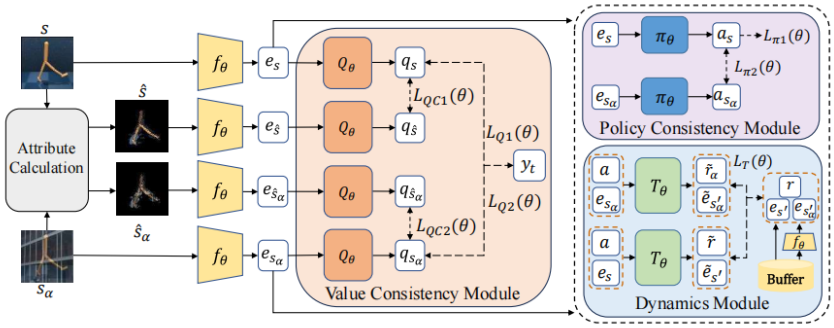

Figure 2: Overview of SCPL. The value consistency module is trained using the original and augmented observations s and s_α, along with their saliency attribute maps $\hat{s}$ and $\hat{s}_\alpha$. The dynamics module aids the encoder f_θ in acquiring task-relevant representations, while the policy consistency module introduces a constraint to maintain consistency in action distributions.

价值一致性模块

目标：确保编码器和价值函数在原始观察（s）和扰动观察（s_a）中均关注任务相关区域，并保持注意力一致性。

关键设计：

✧ **显著性图生成：**

使用梯度法（Guided Backpropagation）生成显著性图 s^，标记 Q 网络对输入图像的高响应区域。

1. 计算 Q 值对输入图像的梯度 $M(Q,s,a)=\partial Q(s,a)/\partial s$。
2. 保留梯度值最高的前 1−ρ 分位数像素（如 ρ=0.9 表示保留前 10%的高响应区域）。
3. 生成显著性图 $s\hat{}=s\odot M_\rho(Q,s,a)$，其中⊙为逐元素乘法。

在机器人抓取任务中，显著性图会高亮机械臂和待抓取物体，忽略背景纹理。

✧ **损失函数设计：**

1. Q 值损失：分别在原始和增强数据上计算均方误差，确保 Q 值预测准确。
2. 显著性一致性损失：强制 Q 值在原始图与显著性图、干扰图与干扰显著性图上一致。

例如：若原图中机械臂区域的 Q 值为高，干扰图中对应区域的 Q 值也应保持高位。

3. 总损失：综合 Q 值损失和显著性一致性损失，通过权重 λ 平衡二者重要性。

作用：

通过显著性图显式引导注意力，避免智能体被无关像素干扰。

增强数据训练迫使编码器提取扰动不变的特征。

动态模型

目标：辅助编码器学习与环境动态相关的鲁棒表征（如物体运动轨迹、状态转移规律）。

关键设计：

- **动态模型**：预测下一状态的特征和奖励。
  - 输入：当前状态特征 $f_\theta(s)$和动作 a。
  - 输出：预测的下一状态特征$\hat{e}_{s'}$和奖励$\tilde{r}$。
- **动态损失**：最小化预测特征和真实特征的均方误差，以及预测奖励和真实奖励的误差。

作用：

- 迫使编码器理解环境动态（如“机械臂移动后，目标物体的位置变化”），而非静态纹理。
- 结合增强数据训练，提升对干扰观察的鲁棒性。

策略一致性模块

目标：强制策略在原始和干扰观察下输出相似的动作分布。

关键设计：

1. **策略损失**：基于 SAC 框架优化策略，平衡探索与利用。
2. **KL 散度约束：**
   - 计算原始观察特征 e_s 和干扰观察特征 e_{sa} 下策略动作分布的差异。
   - 最小化 KL 散度，强制策略在两种观察下动作一致。
   - **示例**：在自动驾驶中，无论晴天或雨天，智能体对“刹车”动作的概率应接近。

作用：

- 避免智能体因视觉干扰（如雨天模糊）而采取不一致动作（如突然转向）。
- 缩小训练与测试环境的策略差异，提升泛化能力。

算法流程

1. 数据采样：从经验池中采样原始数据，生成干扰数据（如随机卷积、叠加动态背景）。

2. 更新价值模块：通过最小化 Q 值损失和显著性一致性损失，优化编码器和价值网络。

3. 更新动态模块：通过预测下一状态和奖励，优化编码器和动态模型。

4. 更新策略模块：通过策略损失和 KL 散度约束，优化策略网络。

Algorithm 1 SCPL (changes to SAC in blue)

Parameter: $\mathcal{B}$: replay buffer, N_A: dynamics module update frequency, τ: data augmentation function, α: learning rate, λ: value consistency coefficient, β: policy consistency coefficient.

for each iteration **do**
  Sample a transition:
  $a, s^{'} \sim \pi_\phi(\cdot|s), P(\cdot|s,a)$
  Add transition to replay buffer:
  $\mathcal{B} \leftarrow \mathcal{B} \cup \{(s,a,\mathcal{R}(s,a),s')\}$
  Sample a batch of transition:
  $\{s,a,r,s^{'}\} \sim \mathcal{B}$
  Generate augmented data:
  $s_\alpha \leftarrow \tau(s)$
  Update value consistency module:
  $\{\theta,\zeta\} \leftarrow \{\theta,\zeta\} -$
  $\alpha\nabla_{\{\theta,\zeta\}}(L_{Q1}(\theta,\zeta)+L_{Q2}(\theta,\zeta)+\lambda L_{QC1}(\theta,\zeta)+\lambda L_{QC2}(\theta,\zeta))$
  Update dynamics module:
  $\{\theta,\psi\} \leftarrow \{\theta,\psi\} - \alpha\nabla_{\{\theta,\psi\}}L_T(\theta,\psi)$
  Update policy consistency module:
  $\phi \leftarrow \phi - \alpha\nabla_\phi(L_{\pi o}(\phi)+\beta L_{\pi c}(\phi))$
end for

网络架构：
编码器：11 层卷积网络，输入为连续 3 帧图像（84×84），捕获时序信息。
价值网络 、策略网络：全连接网络，分别输出 Q 值和动作分布。
动态网络：全连接网络，预测下一状态特征和奖励。

在得出以上结论的关键还有一个理论支撑，下面介绍一下。
核心结论：策略一致性（KL 散度）与泛化性能直接相关。
推导过程：

1. 定义策略在训练环境（πo ）和测试环境（πp）的 KL 散度。
2. 证明累积奖励差异的上界与 KL 散度成正比：

$$\eta(\pi_o) - \eta(\pi_p) \leq C \cdot D_{\mathrm{KL}}^{\max}(\pi_o, \pi_p)$$

   其中 C 为常数，η 为累积奖励。
   意义：KL 散度越小，策略在测试环境中的性能下降越小，泛化能力越强。

实验部分

一、实验设置

1. **基准任务：**

- DMC-GB（DeepMind Control-Generalization Benchmark）：包含 5 项控制任务（如 Walker Walk、Cartpole Swing-up），测试环境分为：
  - Color Hard：改变物体颜色和背景色调。
  - Video Easy：添加简单动态背景（如缓慢移动的云）。
  - Video Hard：覆盖复杂动态背景（如快速闪烁的视频）。

- 机器人操作（Robotic Manipulation）：包含 3 项任务（Reach、Push、Peg in Box），测试环境分为：
    - Test Easy：更换桌面纹理和背景颜色。
    - Test Hard：叠加复杂图像干扰（如动态图案）。
- CARLA 自动驾驶：在多种天气（晴天、雨天、日落）下测试车辆行驶距离和安全性。

2. **对比方法：**

- SAC：基础强化学习算法，无数据增强。
- SVEA：通过数据增强更新价值函数。
- SGQN：结合显著性引导注意力对齐。
- MaDi：通过掩码网络过滤任务无关区域。
- CNSN：基于归一化增强泛化能力。

3. **评价指标：**

- 平均回报（Episode Return）：智能体在测试环境中的累积奖励。
- KL 散度：训练与测试环境策略动作分布的差异（越小越好）。
- 注意力区域准确率（ACC/AUC/F1）：衡量智能体是否聚焦任务相关像素。

Table 1: DMC-GB Generalization Performance

Setting	Task	SAC[9]	SVEA[11]	SIM[39]	TLDA[43]	PIE-G[44]	SGQN[2]	CG2A[23]	MaDi[7]	CNSN[20]	**SCPL**
Color hard	Walker stand	423 ± 155	942 ± 26	940 ± 2	947 ± 26	941 ± 35	948 ± 25	**972 ± 23**	–	942 ± 19	960 ± 11
	Walker walk	255 ± 61	760 ± 145	803 ± 33	823 ± 58	884 ± 20	810 ± 43	902 ± 46	–	815 ± 65	**939 ± 19**
	Cartpole	615 ± 29	837 ± 23	841 ± 13	760 ± 60	749 ± 46	806 ± 6	856 ± 40	–	679 ± 35	**857 ± 12**
	Ball in cup	391 ± 245	961 ± 7	953 ± 7	932 ± 32	964 ± 7	887 ± 10	**972 ± 10**	–	894 ± 78	966 ± 9
	Finger spin	373 ± 70	**977 ± 5**	960 ± 6	–	–	899 ± 27	928 ± 43	–	–	929 ± 24
	Average	411	895	899	865	884	870	926	–	833	**930(+1%)**
Video easy	Walker stand	351 ± 245	961 ± 8	963 ± 5	**973 ± 6**	957 ± 12	955 ± 9	968 ± 6	967 ± 3	967 ± 6	968 ± 8
	Walker walk	228 ± 65	819 ± 71	861 ± 33	873 ± 34	870 ± 22	910 ± 24	918 ± 20	895 ± 24	842 ± 58	**941 ± 9**
	Cartpole	359 ± 80	782 ± 27	770 ± 13	671 ± 57	597 ± 61	761 ± 28	788 ± 24	**848 ± 6**	752 ± 26	814 ± 21
	Ball in cup	338 ± 201	871 ± 106	820 ± 135	887 ± 58	922 ± 20	950 ± 24	**963 ± 28**	807 ± 144	913 ± 45	**963 ± 10**
	Finger spin	260 ± 98	808 ± 33	815 ± 38	744 ± 18	837 ± 107	956 ± 26	912 ± 69	679 ± 17	–	**963 ± 8**
	Average	300	848	845	830	837	906	909	839	869	**930(+2%)**
Video hard	Walker stand	225 ± 58	747 ± 43	827 ± 24	602 ± 51	852 ± 56	851 ± 24	895 ± 35	920 ± 14	871 ± 23	**953 ± 15**
	Walker walk	104 ± 18	385 ± 63	459 ± 67	271 ± 55	600 ± 28	739 ± 21	687 ± 18	504 ± 33	480 ± 46	**818 ± 32**
	Cartpole	174 ± 24	401 ± 38	367 ± 47	286 ± 47	401 ± 21	544 ± 43	472 ± 24	619 ± 24	417 ± 31	**675 ± 3**
	Ball in cup	196 ± 82	498 ± 147	287 ± 39	257 ± 57	786 ± 47	782 ± 57	806 ± 44	758 ± 135	691 ± 72	**924 ± 7**
	Finger spin	26 ± 21	307 ± 24	362 ± 9	241 ± 29	762 ± 59	822 ± 24	819 ± 38	358 ± 25	–	**897 ± 22**
	Average	145	467	460	331	680	747	736	632	615	**853(+14%)**

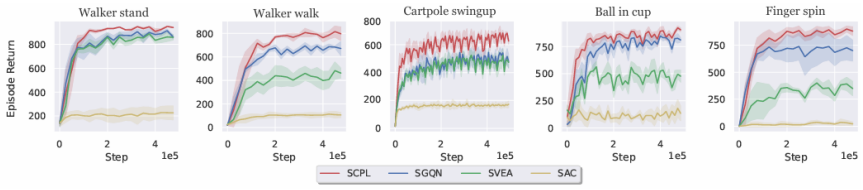

Figure 4: The performance of SAC, SVEA, SGQN, and SCPL in *Video hard* setting. SCPL (red line) shows better generalization.

Video Hard 环境：

SCPL 在 Walker Stand、Cartpole 等任务中平均回报比最佳基线（SGQN）提升 14%。

例如，Walker Walk 任务中，SCPL 回报为 818±32，而 SGQN 为 739±21。

Color Hard 环境：

SCPL 在 Ball in Cup 任务中回报达 966±9，比 SVEA（961±7）进一步提升。

结论：SCPL 在复杂动态背景干扰下表现最优，尤其在 Video Hard 环境下优势显著。

Table 2: Metrics for attention region of RL agents

	SAC	SVEA	SGQN	**SCPL**
ACC	0.889	0.926	0.932	**0.942**
AUC	0.811	0.833	0.862	**0.908**
F1	0.341	0.462	0.463	**0.566**

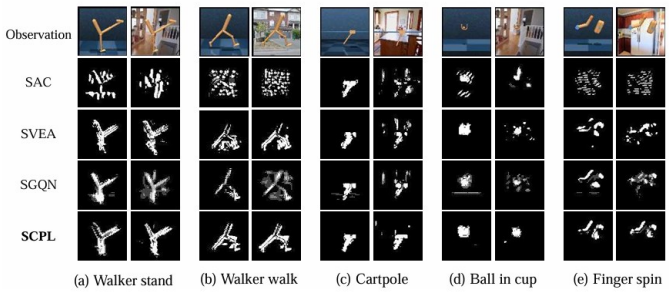

Figure 5: Saliency attribute maps for SAC, SVEA, SGQN, and SCPL in *Training* and *Video hard* setting. In observations of each task, the first column is the original observation, and the second column is the perturbed observation.

可以看到 SCPL 能在原图与扰动图中都关注到正确的任务相关像素

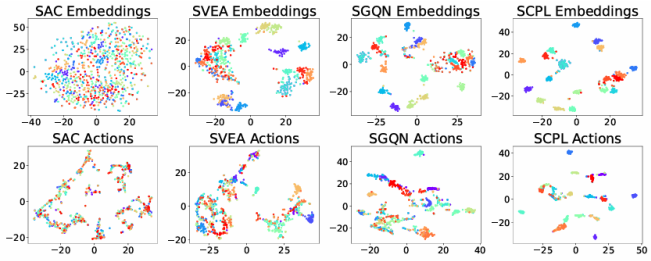

Figure 6: t-SNE maps of embeddings and actions learned with SVEA, SGQN, and SCPL for 20 motion situations, generated by randomly selecting 40 backgrounds from *Video hard*. Different motion situations are represented by different colors, and dots represent representations or actions.

SCPL 的特征表示（embeddings）在相同动作不同背景下高度聚集，不同动作明显分离。

基线方法（如 SVEA）的特征分布混乱，不同背景干扰导致类别重叠。

说明：SCPL 的编码器提取了背景无关的通用特征。

Table 3: Ablation study of three significant components in SCPL

Benchmark	Environment	SAC	SAC + dynamics module	SAC + value consistency	SAC + value + policy consistency	**SCPL**
Video hard	Walker stand	225 ± 58	630 ± 26	918 ± 29	949 ± 9	**953 ± 15**
	Walker walk	104 ± 18	336 ± 14	673 ± 9	812 ± 20	**818 ± 32**
	Cartpole	174 ± 24	351 ± 15	567 ± 64	624 ± 52	**675 ± 3**
	Ball in cup	196 ± 82	405 ± 29	805 ± 67	909 ± 12	**924 ± 7**
	Finger spin	26 ± 21	292 ± 6	703 ± 15	802 ± 6	**897 ± 22**
	Average	145	403(**+178%**)	733(**+405%**)	819(**+465%**)	**853**(**+488%**)

验证了各模块的贡献。

移除动态模块：DMC-GB Video Hard 任务平均回报从 853 下降至 819（降幅 4%）。

移除策略一致性模块：回报进一步下降至 733（降幅 14%）。

仅使用价值一致性模块：回报为 733，说明动态和策略模块缺一不可。

结论：三个模块协同作用，动态模块对鲁棒性贡献最大（17%），策略模块次

之（12%）。

Table 4: Performance comparison on Robotic Manipulation

Setting	Task	SAC	SVEA	SGQN	**SCPL(ours)**
Train	Reach	1.5 ± 6.7	33.6 ± 0.6	33.6 ± 0.7	**33.8 ± 0.3**
	Push	−25.3 ± 13.4	10.8 ± 7.0	18.8 ± 6.4	**19.2 ± 6.1**
	Peg	−12.6 ± 13.2	152.6 ± 21.1	179.8 ± 23.1	**194.6 ± 12.1**
	Average	−12.1	65.7	77.4	**82.6(+7%)**
Test easy	Reach	−22.7 ± 6.4	32.2 ± 1.0	28.2 ± 5.9	**33.3 ± 0.3**
	Push	−23.6 ± 11.2	2.9 ± 10.4	−12.6 ± 12.6	**6.4 ± 6.5**
	Peg	−33.6 ± 20.2	110.4 ± 44.3	94.6 ± 12.0	**181.0 ± 14.0**
	Average	−26.6	48.5	36.7	**73.6(+52%)**
Test hard	Reach	−19.9 ± 4.8	27.8 ± 1.2	18.9 ± 4.6	**31.9 ± 2.0**
	Push	−24.2 ± 11.0	**-1.7 ± 13.5**	−17.7 ± 11.3	−3.1 ± 5.1
	Peg	−25.1 ± 5.2	114.0 ± 43.6	124.8 ± 28.8	**166.4 ± 15.0**
	Average	−23.1	46.7	42.0	**65.1(+39%)**

Test Hard 环境：

SCPL 在 Peg in Box 任务中回报为 166.4±15.0，比 SGQN（124.8±38.8）提升 33%。

平均性能提升 39%，远超其他方法。

原因：SCPL 通过显著性引导，在复杂桌面干扰下仍能准确定位目标物体。

Table 5: Performance comparison on CARLA

Setting	SAC	SVEA	SGQN	**SCPL(ours)**
Train	472 ± 110	297 ± 14	614 ± 41	**643 ± 87**
Wet noon	468 ± 68	353 ± 112	473 ± 187	**564 ± 123**
Hard rain noon	306 ± 114	268 ± 89	406 ± 63	**442 ± 199**
Wet sunset	23 ± 16	125 ± 36	39 ± 17	**271 ± 28**
Soft rain sunset	45 ± 25	22 ± 5	59 ± 44	**243 ± 29**
Mid rain sunset	44 ± 24	42 ± 31	63 ± 46	**242 ± 11**
Test Average	177	162	208	**352(+69%)**

CARLA 自动驾驶任务表现

雨天环境（Hard Rain Noon）：SCPL 平均行驶距离为 442±199 米，比 SGQN（406±63 米）提升 9%。

日落环境（Mid Rain Sunset）：SCPL 行驶距离达 242±11 米，而其他方法（如 SVEA）仅 42±31 米。

结论：SCPL 在低光照和雨天干扰下仍能稳定导航，平均性能提升 69%。