DMR: Decomposed Multi-Modality Representations for Frames and Events Fusion in Visual Reinforcement Learning

Haoran Xu[1,2] Peixi Peng[2,3*] Guang Tan[1*] Yuan Li[4] Xinhai Xu[4] Yonghong Tian[2,3,5]
[1]School of Intelligent Systems Engineering, Shenzhen Campus of Sun Yat-sen University
[2]Peng Cheng Laboratory
[3]School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University
[4]Academy of Military Sciences [5]School of Computer Science, Peking University

Abstract

We explore visual reinforcement learning (RL) using two complementary visual modalities: frame-based RGB camera and event-based Dynamic Vision Sensor (DVS). Existing multi-modality visual RL methods often encounter challenges in effectively extracting task-relevant information from multiple modalities while suppressing the increased noise, only using indirect reward signals instead of pixel-level supervision. To tackle this, we propose a Decomposed Multi-Modality Representation (DMR) framework for visual RL. It explicitly decomposes the inputs into three distinct components: combined task-relevant features (co-features), RGB-specific noise, and DVS-specific noise. The co-features represent the full information from both modalities that is relevant to the RL task; the two noise components, each constrained by a data reconstruction loss to avoid information leak, are contrasted with the co-features to maximize their difference. Extensive experiments demonstrate that, by explicitly separating the different types of information, our approach achieves substantially improved policy performance compared to state-of-the-art approaches.

1. Introduction

Visual reinforcement learning (RL) is instrumental in training intelligent agents to make decisions by directly translating complex visual inputs into actions. It has found applications in various domains such as autonomous driving [47, 50], robot control [19, 52], and video games [31, 42]. While most visual RL methods use frame-based RGB cameras as their primary source of perception [19, 20, 45, 46], these methods face limitations in certain situations due to the camera's low dynamic range (70 dB) [27]. For example, in traffic scenarios with complex lighting conditions like night driving or tunnel traversal, the frame-based methods may experience substantial performance degradation [51].

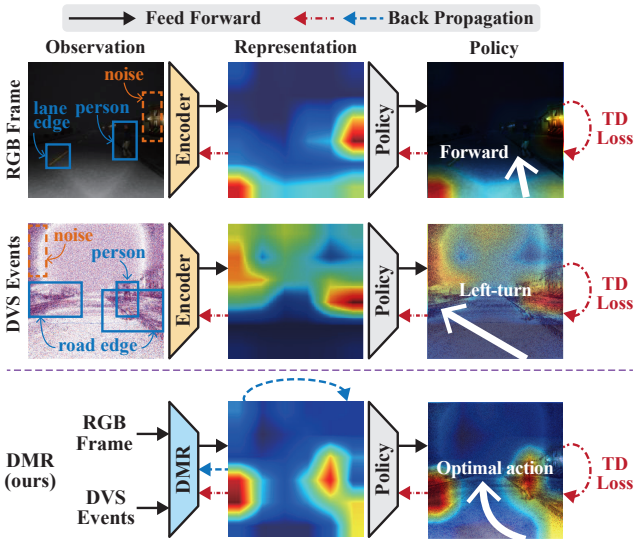


Figure 1. Several typical visual examples of frames and events based RL. (i) In the first row, insufficient ambient light causes RGB underexposure, leading to the overlooking of the front pedestrian and resulting in a forward policy aligned with the lane direction that could cause collisions. (ii) In the second row, the lack of texture in DVS causes the person and the background to blend, leading to a left-turn policy to avoid the highlighted area on the right. (iii) In contrast, our method (third row) can fully take advantage of RGB and DVS to extract task-relevant information and eliminate task-irrelevant and noisy information through joint TD and DMR learning, thereby obtaining an optimal evasion policy.

Recently, there has been a growing interest in using bio-inspired event cameras to address these challenges [12, 37, 39]. Event-based cameras are well suited to adverse visual conditions due to their rapid adaptation to changes in light intensity and a high dynamic range (120 dB) [27]. Their high sampling rate, typically in the order of KHz, al-

* Corresponding authors.

lows them to work in high-speed scenarios while exhibiting much less perceptual distortion than conventional cameras or LiDAR [34, 48]. Despite these advantages, event-based cameras operate by capturing asynchronous per-pixel brightness changes, called "events", rather than recording absolute brightness at a constant rate. As a result, they may miss crucial visual cues, especially from stationary or slow-moving objects, which could be captured by frame-based cameras. Given the distinct mechanisms of frame- and event-based cameras, it becomes essential to explore how to effectively leverage the strengths of both modalities.

The integration of frame- and event-based cameras has been explored for tasks like object detection [26] and depth estimation [11]. However, in vision-based RL, where entire observations are mapped to decisions only using temporal-difference (TD) loss [46], without pixel-level [40, 49] or instance-level supervision [26, 38], simply aggregating frames and events can result in increased noise and task-irrelevant information. This phenomenon results in noise injection in the latent state space and leads to reduced RL performance. To address this, we categorize the information from frames and events into three distinct types: 1) Combined task-relevant feature, referred to as *co-feature*; 2) RGB-specific noise and task-irrelevant feature, or simply *RGB noise*; and 3) DVS-specific noise and task-irrelevant feature, or *DVS noise*. The co-features represent the full information from both modalities that is essential for the RL task, while the noise represents unwanted information that may negatively impact the RL process. As shown in Fig. 1, combining frames and events helps to extract important regions, including the pedestrian and road edges. These regions are difficult to identify precisely using either modality alone. It is notable that these three parts are all latent and only the rewards collected by interacting with environments are available as external guidance during learning, which is consistent with the standard RL pipeline.

To learn the three types of information, we propose a novel three-branch representation learning framework. The framework comprises parallel branches that independently encode RGB noise and DVS noise, with a third branch merging RGB frames and DVS events at the input level to extract the co-features. In the framework, three types of constraints are designed. Firstly, the co-features are learned under the guidance of RL-related loss [13], so that the co-features are useful for the RL task. Secondly, a contrastive loss is designed to increase the distance between the noise and co-features. Thirdly, to ensure information completeness, the co-features and noise features are used to reconstruct the raw event and frame observations. In summary, the contributions of this paper are three-fold:

- We present a novel approach to fuse RGB frames and DVS events in vision-based RL, highlighting the concept of decomposed representation learning. This approach is a pioneering effort in handling the RL task through the fusion of frame- and event-based modalities.
- We devise a new three-branch learning framework that effectively separates task-relevant information from noise. This filtering process mitigates noise injection in the latent state space, proving to be highly beneficial for downstream policy learning.
- We conduct comprehensive experiments using the proposed Carla benchmark. The results verify the efficacy of our method in various traffic scenarios and adverse weather conditions.

2. Related work

Vision-based RL. In vision-based reinforcement learning (RL), the agent typically requires low-dimensional abstract representations of visual observations to expedite the decision-learning process. This process, termed as *state abstraction*, can be accomplished using four main approaches: (i) Reconstruction-based techniques [39, 44, 45]; (ii) Reward and transition dynamics prediction [13]; (iii) Contrastive-based representations [1, 8, 23, 24, 32, 45]; (iv) Bisimulation-based techniques [3, 16, 47]. Conventional vision-based RL mainly focuses on frame-based RGB cameras, which are vulnerable to lighting anomalies and motion blur. These issues can be mitigated by integrating event-based neuromorphic cameras.

Event-based Neuromorphic Sensors. DVS, a prevalent form of neuromorphic sensor, detects local pixel-level intensity changes without global exposure. Its high dynamic range (exceeding 120 dB) enables quick adaptation to diverse lighting conditions [10]. DVS's asynchronous event streams allow it to capture high-speed motion with a high temporal resolution [4]. Event-based sensors find applications in object detection [12, 27, 28], image reconstruction [33], semantic segmentation [37], and odometry [17, 53]. DVS's potential in policy learning is currently under active exploration [2, 39, 41]. Walters *et al*. [41] took full advantage of the high-frequency characters of events to realize continuous RL. Andersen *et al*. [2] designed an event-based autonomous navigation control framework to detect gates in a racing track. Vemprala *et al*. [39] proposed an event variational autoencoder that directly learns representations of asynchronous event streams for RL, rather than pre-processing the events over a time period as an image-like 3D tensor, such as voxel grid [26, 33]. Hence, it may require huge GPU memory consumption during training when the scenario involves a large number of events.

Multi-Modality Learning of RGB and DVS. Since DVS only captures per-pixel brightness changes, it may miss crucial visual cues. Recent research on combining frame- and event-based modalities has gained momentum [11, 36, 38, 51]. Notably, RENet [51] extracted multi-scale temporal cues from events and calibrated them with frames in a coarse-to-fine manner. EFNet [36] proposed cross-

modality channel-wise attention at multiple levels of the network to adaptively fuse frames and events. FPNet [38] combined features from frames and events using a multi-scale pyramid network to minimize information loss during fusing. Except for the above multi-modality learning with RGB and DVS, other traditional modality combinations, such as RGB&Depth [25] and RGB&Lidar [6], have been explored. Despite advances in existing multi-modality RL methods [5, 22, 30], most of them face challenges in effectively removing modality-specific noise and task-irrelevant information using TD loss, leading to suboptimal decisions based on a noisy feature map. When considering RGB and DVS, their distinct advantages in imaging principles and complementary nature become more apparent, which motivates us to explore a new decomposition approach to address the RL problem. We show that explicitly modelling task-irrelevant noises and task-relevant information is necessary for RGB and DVS to enhance RL performance.

3. Preliminaries

Soft Actor-Critic (SAC). The RL problem is normally formulated as a Markov Decision Process (MDP), denoted as a tuple $\mathcal{M} =< \mathcal{O}, \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma >$, where $\mathcal{O}, \mathcal{S}, \mathcal{A}$ are visual observation space, state space, and action space, respectively. As per the convention [19, 31], we define the agent's interaction process in an MDP as follows: (i) the agent perceives the visual observation o_t and stacks consecutive observations $\{o_{t-2}, o_{t-1}, o_t\}$ into the current state s_t; (ii) the agent then selects an action $a_t \in \mathcal{A}$ based on a stochastic policy $\pi(a_t|s_t)$; (iii) the agent receives reward feedback $r_{t+1} \sim \mathcal{R}(s_t, a_t)$, and transitions to the next state $s_{t+1} \sim \mathcal{P}(s_{t+1}|s_t, a_t)$. The goal of this formulation is to find an optimal policy π^* that maximizes the expected cumulative reward across the entire rollout of MDPs.

SAC [14, 15] is a widely-used RL algorithm, which incorporates an α-discounted maximum entropy, denoted as $H(\cdot)$, to ensure diverse action exploration. Formally, the objective of SAC is defined as:

$$J(\pi) = \sum_t \mathbb{E}_{s_t, a_t \sim \pi} \left[\mathcal{R}(s_t, a_t) + \alpha \mathcal{H}(\pi(a_t|s_t)) \right]. \quad (1)$$

During the interaction with the environment, the action-value Q is estimated by minimizing the soft Bellman error:

$$\mathcal{L}_Q = \mathbb{E}_{s_t, a_t \sim \pi} \left[Q(s_t, a_t) - (r_t + \lambda V(s_{t+1})) \right]. \quad (2)$$

Additionally, the state-value V can be approximated by sampling an action from the current policy:

$$V(s_{t+1}) = \mathbb{E}_{a_{t+1} \sim \pi} \left[\tilde{Q}(s_{t+1}, a_{t+1}) - \alpha \log \pi(a_{t+1}|s_{t+1}) \right], \quad (3)$$

where the weights in $\tilde{Q}$ are computed as an exponentially moving average of the weights in Q. Thereby, the policy is optimized by decreasing the divergence between the exponential of Q function and the policy π:

$$\mathcal{L}_\pi = \mathbb{E}_{a_t \sim \pi} \left[\alpha \log \pi(a_t|s_t) - \tilde{Q}(s_t, a_t) \right]. \quad (4)$$

DeepMDP. DeepMDP [13] initially learns to encode the high-dimensional observation into a compact and continuous representation, and then acquires a policy under it. DeepMDP extracts a parameterized latent MDP $\bar{\mathcal{M}}$ for the original MDP $\mathcal{M}$. Let $\Phi : \mathcal{S} \to \bar{\mathcal{S}}$ be an state abstract function. We denote by $(\bar{\mathcal{M}}, \Phi)$ the latent space model of MDP $\mathcal{M}$. $\bar{\mathcal{M}}$ contains transition model $\bar{\mathcal{P}}_{\theta_p}$ and reward model $\bar{\mathcal{R}}_{\theta_r}$, parameterized by θ_p and θ_r, respectively. To obtain the DeepMDP, the reward difference and the 2-Wasserstein metric W_2 are minimized through:

$$\mathcal{L}_\mathcal{P} = \mathbb{E}_{s_t, a_t \sim \pi} W_2(\Phi\mathcal{P}(\cdot|s_t, a_t) - \bar{\mathcal{P}}_{\theta_p}(\cdot|\Phi(s_t), a_t)), \quad (5)$$

$$\mathcal{L}_\mathcal{R} = \mathbb{E}_{s_t, a_t \sim \pi} \left\| \mathcal{R}(s_t, a_t) - \bar{\mathcal{R}}_{\theta_r}(\Phi(s_t), a_t) \right\|. \quad (6)$$

where the shorthand notation $\Phi\mathcal{P}(\cdot|s_t, a_t)$ denotes the original state samples s_{t+1} over the distribution $\mathcal{P}(\cdot|s_t, a_t)$ and then embedding s_{t+1} via Φ. In this paper, we keep DeepMDP as the basic RL method, but our method can be easily applied to other RL frameworks, as demonstrated in the supplementary materials.

4. Methodology

4.1. Multi-Modality Visual RL Problem

Our RL pipeline comprises two main components, multi-modality representation learning and policy learning. Specifically, the agent obtains a multi-modality perception $\{o^{D_i}\}$, where $i \in \{1, 2, \cdots, d\}$, from the joint observation space $\mathcal{O} = \Pi\mathcal{O}^{D_i}$. Here, $\mathcal{O}$ is defined as the Cartesian product of d sub observation spaces $\mathcal{O}^{D_i}$, with each $\mathcal{O}^{D_i}$ representing the observation space of the modality D_i. The joint state s_t is formed by concatenating several consecutive visual observations from multiple modalities, namely $\cup_i^d \{o_{t-2}^{D_i}, o_{t-1}^{D_i}, o_t^{D_i}\}$. The agent learns to encode the original high-dimensional state s_t into a compact representation z_t for the subsequent policy learning.

Since RL directly learns policy from entire observations without the guidance of pixel-level supervision, the heterogeneous modalities pose challenges in extracting task-relevant features that are crucial for policy. Therefore, we propose Decomposed Multi-modality Representations (DMR) framework for RL, as shown in Fig. 2. This framework mainly integrates two modalities, RGB frames and DVS events, i.e., $D_i \in \{\text{rgb}, \text{dvs}\}$ and $d = 2$.

4.2. Event Processing in DMR

DVS can capture independent pixel-level changes in light intensity, resulting in the generation of asynchronous event

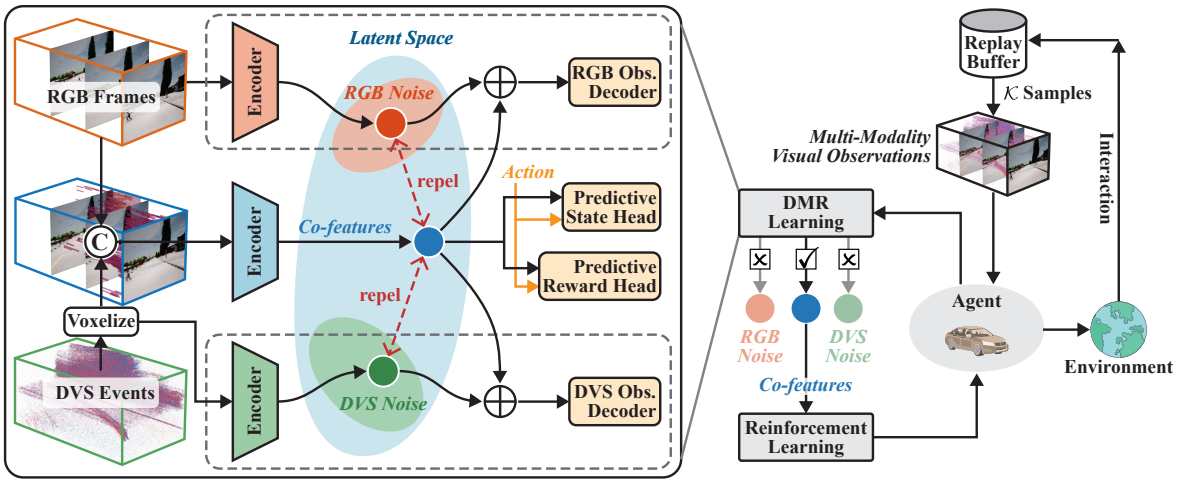


Figure 2. The proposed multi-modality learning framework DMR. We explicitly decompose the input into co-features and modality-specific noises. The co-features are extracted under the guidance of two task-relevant predictive heads, while the two noise components are contrasted with the co-features to maximize their distance. In addition, the completeness of information is ensured by imposing a reconstruction constraint on the decoder of each modality. The downstream task decisions are learned by using co-features through RL algorithms. Notice that during testing, DMR retains only the intermediate co-feature branch.

streams. An event e_i in the stream is defined as a four-attribute tuple (x_i, y_i, t_i, p_i), which is triggered when the logarithmic intensity of the pixel (x_i, y_i) at timestamp t_i exceeds the pre-defined threshold $\pm$Q. This process can be described as:

$$L(x_i, y_i, t_i) - L(x_i, y_i, t_i - \Delta t) = p_i \cdot \mathrm{Q}, \tag{7}$$

where Δt is the sampling rate of DVS, the polarity $p_i \in \{1, -1\}$ is determined based on the intensity change, with 1 representing an increment and -1 representing a decrement.

When processing asynchronous DVS events, it is a common practice to convert events within a fixed-length temporal window into a fixed-size tensor representation, referred to as a voxel grid [26, 33]. To synchronize events with the low sampling rate of RGB frames, we partition the incoming events within the fixed time interval of RGB frames. The events occurring between the pair-wise frames (o_{t-1}^{rgb} and o_t^{rgb}) are discretized into a spatio-tempral voxel grid $\mathcal{E}_t$ with B temporal bins (as depicted in Fig. 2). Each element in the voxel grid has three dimensions, two-dimensional location (x_l, y_m), and temporal dimension (t_n). Formally:

$$\mathcal{E}_t(x_l, y_m, t_n) = \sum_{x_l, y_m = x_i, y_i} p_i \max(1 - |t_n - t_i^*|), \tag{8}$$

where t_i^* is the normalized event timestamp, which is defined as $t_i^* = \frac{B-1}{\Delta d}(t_i - t_0)$; Δd is the time interval between adjacent RGB frames and t_0 is the first event timestamp within the interval.

In this paper, we set $B = 5$, and set RGB sampling rate to 20Hz, that is, $\Delta d = 0.05$s. Therefore, the observation o_t^{dvs} of the DVS camera at each time t is preprocessed to form a voxel grid, which is then fed into DMR.

4.3. Representation Learning in DMR

Let z_t^i denote the representation for the original observation o_t^i of modality $i \in \{\mathrm{rgb}, \mathrm{dvs}\}$. The representations z_t^i may differ significantly for the two modalities even when they yield similar policies, because of the different working principles of RGB and DVS cameras. We decompose z_t^i into co-features z_t^{c} and modality-specific noises h_t^i as:

$$z_t^i = z_t^{\mathrm{c}} \oplus h_t^i. \tag{9}$$

To achieve this, DMR comprises three branches, as depicted in Fig. 2. The upper and lower branches take RGB frames and DVS events as inputs, respectively. The data then pass through their respective encoders, denoted as $\Phi_{\theta_{\mathrm{rgb}}}$ and $\Phi_{\theta_{\mathrm{dvs}}}$, to generate modality-specific noise (h_t^{rgb}, h_t^{dvs}). The intermediate branch takes the concatenation of RGB and DVS as input. Its output, co-features z_t^c, are generated by the intermediate encoder parameterized as Φ_{θ_c}.

To ensure the completeness of information, we employ reconstruction decoders, denoted as $\mathcal{D}_{\theta_i}$, to ensure that the respective original observations o_t^i can be recovered:

$$\mathcal{L}_{\mathcal{D}} = \sum_{i \in \{\mathrm{rgb}, \mathrm{dvs}\}} \left\| \mathcal{D}_{\theta_i}(z_t^{\mathrm{c}} + h_t^i) - o_t^i \right\|_2, \tag{10}$$

where $t \in \mathcal{K}$ and $\mathcal{K}$ is the set of sample indices in a training batch that are from different time steps in different MDPs.

While ensuring the completeness of z_t^i, we utilize the task-relevant predictive heads to guide the extraction of the co-features z^{c}. Here, we incorporate the tractable reward

Algorithm 1 Pseudocode for DMR Learning

Initialize the replay buffer $\mathcal{B}$ with random episodes.
while *Not converged* **do**
  `// Representation Learning`
  Collect multi-modality visual sequences randomly $\{(o_t^{\text{rgb}}, o_t^{\text{dvs}})\}_{t\in\mathcal{K}} \sim \mathcal{B}$.
  Obtain decomposed representations $z_t^{\text{c}}, h_t^{\text{rgb}}, h_t^{\text{dvs}}$ via $\Phi_{\theta_{\text{c}}}, \Phi_{\theta_{\text{rgb}}}, \Phi_{\theta_{\text{dvs}}}$.
  Perform completeness constraint on $z_t^{\text{rgb}}, z_t^{\text{dvs}}$ for each modality via Eqs. (9) and (10).
  Extract co-features z_t^{c} via Eqs. (11) and (12).
  Distinguish noise from co-features z_t^{c} via Eq. (13).
  `// Reinforcement Learning`
  Estimate action, state-value via Eqs. (14) and (15).
  Establish z_t^{c}-driven policies via Eq. (16).
  `// Environment Interaction`
  Execute $a_t \sim \pi_\phi(a_t|z_t^{\text{c}})$, receive $r_t \sim \mathcal{R}(s_t, a_t)$.
  Observe $o_{t+1}^{\text{rgb}}, o_{t+1}^{\text{dvs}}$, and $s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)$.
  Add experience (s_t, a_t, r_t, s_{t+1}) to the replay buffer.
end while
return $\Phi_{\theta_{\text{c}}}, \pi_\phi$

and state head from DeepMDP [13] into the predictive head. It is worth noting that DMR can be seamlessly plugged into various multi-modality visual RL methods, while providing notable enhancements in performance. Thus, we have:

$$\mathcal{L}_{\mathcal{P}} = \left\| \bar{\mathcal{P}}_{\theta_p}(z_t^{\text{c}}, a_t) - z_{t+1}^{\text{c}} \right\|, \tag{11}$$

$$\mathcal{L}_{\mathcal{R}} = \left\| \bar{\mathcal{R}}_{\theta_r}(z_t^{\text{c}}, a_t) - r_{t+1} \right\|, \tag{12}$$

where $\bar{\mathcal{P}}_{\theta_p}$ and $\bar{\mathcal{R}}_{\theta_r}$ are state and reward predictive heads, respectively. These auxiliary models share the same structure except that the output of $\bar{\mathcal{R}}_{\theta_r}$ is a one-dimension scalar.

Finally, the noise should exhibit clear dissimilarity from the co-features. In other words, there should be minimal overlap between h_t^i and z_t^{c}. To achieve this distinction, we design the following contrastive constraint:

$$\mathcal{L}_{\mathcal{C}} = -\log \frac{f(z_t^{\text{c}}, \tilde{z}_t^{\text{c}})}{f(z_t^{\text{c}}, \tilde{z}_t^{\text{c}}) + \sum\limits_{i\in\{\text{rgb},\text{dvs}\}} \sum\limits_{k\in\mathcal{K}} f(z_k^{\text{c}}, \tilde{h}_k^i)}, \tag{13}$$

where $\tilde{z}_t^{\text{c}}$ and $\tilde{h}_t^i$ indicate the moving-averaged target values [7] of z_t^{c} and h_t^i, respectively, and the function $f(a, b) = \exp(\langle a, b\rangle/\tau)$ measures the similarity between a and b using the dot product $\langle a, b\rangle$ and the temperature parameter τ.

4.4. Reinforcement Learning based on DMR

With the full sensory input decomposed, we can proceed to develop policies for the downstream task using the extracted co-features. These co-features are isolated from irrelevant information, enabling them to more effectively support the objectives of downstream control.

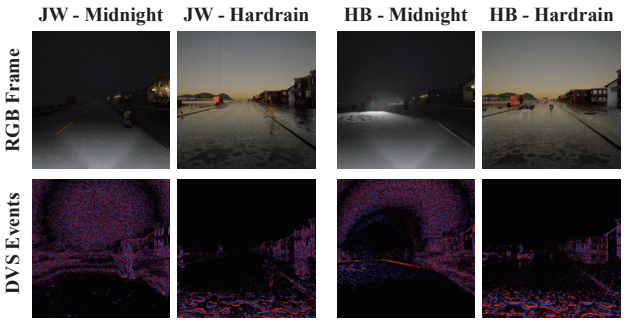


Figure 3. Illustration of the Carla autopilot benchmark.

We modify the baseline RL algorithm SAC [14, 15] to align with our co-features-driven policy learning approach. In this process, we estimate the action-value Q and state-value V by utilizing the Bellman equation and the co-features z_t^{c} generated from the encoder $\Phi_{\theta_{\text{c}}}$:

$$\mathcal{L}_Q = \mathop{\mathbb{E}}_{t\in\mathcal{K}} \left[Q(z_t^{\text{c}}, a_t) - (r_t + \lambda V(z_{t+1}^{\text{c}})) \right], \tag{14}$$

$$V(s_{t+1}) = \mathop{\mathbb{E}}_{t\in\mathcal{K}} \left[\tilde{Q}(z_{t+1}^{\text{c}}, a_{t+1}) - \alpha \log \pi(a_{t+1}|z_{t+1}^{\text{c}}) \right]. \tag{15}$$

The policy π_ϕ can be derived from:

$$\mathcal{L}_\pi = \mathop{\mathbb{E}}_{t\in\mathcal{K}} \left[\alpha \log \pi_\phi(a_t|z_t^{\text{c}}) - \tilde{Q}(z_t^{\text{c}}, a_t) \right]. \tag{16}$$

The full training pipeline of DMR is provided in Algorithm 1 and the framework in Fig. 2. Since the policy optimization is driven only by the co-features, the auxiliary encoders of the two noise branches can be omitted during the testing phase. This means that DMR retains only the encoder $\Phi_{\theta_{\text{c}}}$ and the policy learning network π_ϕ during test, thereby allowing efficient mapping from high-dimensional multi-modality observations to visuomotor policies.

5. Experiments

5.1. Experimental Setup

We mainly focus on the autonomous driving environments in experiments. The environments contain numerous task-irrelevant objects, and the sensors are sensitive to changing weather conditions, providing comprehensive and realistic scenarios to evaluate our method. Since RL involves trial-and-error interactions with the environment, most RL methods test the algorithms in simulators [19–21, 29, 46]. Hence, we adopt the widely-used Carla [6, 18, 43, 50] to establish our new Carla benchmark. Carla supports a rich set of scenarios with varying lighting and weather conditions. More importantly, it is one of the few simulators that allows generation of asynchronous events and RGB frames simultaneously.

As shown in Fig. 3, our Carla benchmark features two traffic scenarios: the HighBeam (HB) scenario, where an

Scenario (Weather)	Metrics	Single-modality Policies			Multi-modality Policies				
		RGB	DVS-F	DVS	TransFuser	EFNet	FPNet	RENet	DMR
JW (Midnight)	Distance	144±70	163±92	190±101	111±79	84±41	106±96	189±107	**230±77**
	Reward	102±67	130±86	136±93	77±64	62±37	84±82	158±101	**194±73**
JW (Hardrain)	Distance	113±83	115±69	87±44	123±54	125±66	47±34	50±40	**146±58**
	Reward	83±72	96±65	52±48	84±52	89±68	23±31	6±30	**111±57**
HB (Midnight)	Distance	80±60	51±63	109±76	97±81	87±67	116±68	106±39	**117±68**
	Reward	58±56	29±59	71±74	46±69	63±63	**85±62**	68±42	71±72
HB (Hardrain)	Distance	91±61	51±20	70±32	122±61	114±65	106±47	125±62	**150±51**
	Reward	70±63	30±24	49±31	85±57	69±68	64±46	73±59	**112±51**

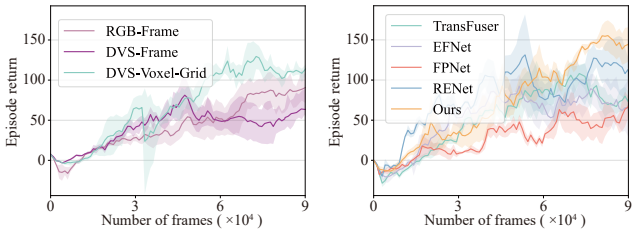
Table 1. Testing performance comparison with SOTA methods under the proposed Carla benchmark. (The best single-modality policies are highlighted in gray background, and the best results in both single- and multi-modality policies are shown in **bold**.)

ego-vehicle experiences varying lighting conditions while encountering a cyclist, and the JayWalk (JW) scenario, where the ego-vehicle encounters both stationary and moving pedestrian obstacles intermittently. Moreover, the benchmark includes extreme weather conditions (Midnight and Hardrain) that can cause RGB camera failure or excessive noise with DVS cameras. For multi-modality observations, we focus on the fusion of RGB frames (RGB for short) and DVS voxel grids (DVS). In addition, we introduce the frame-based DVS events, termed DVS-F [26], as a type of observation to show the effectiveness of DVS voxelization. In the benchmark, the ego-vehicle's objective follows the common setup as in [47], aiming to drive as far as possible without collisions within 500 steps. All experiments are trained across 3 random seeds and 20 evaluation rollouts per seed, yielding mean and standard deviation of the metrics of episode reward and distance.

Our benchmark and code are available online[1].

5.2. Performance Comparison

We compare DMR with both single- and multi-modality algorithms. For the single-modality baselines, we maintain DeepMDP as the baseline RL algorithm, employing the three types of perception input introduced earlier, that is RGB, DVS, and DVS-F. Since there is no previous RL algorithm that combines RGB frames and DVS events, for the multi-modality baselines, we compare DMR against state-of-the-art (SOTA) multi-modality fusion methods, including TransFuser [6], EFNet [36], FPNet [38], and RENet [51]. To ensure a fair comparison, we adopt the same 4-layer CNN structure [46] and parameter initialization for each modality's encoder in these SOTA methods, while keeping DeepMDP as the baseline task predictive head. Further details on comparisons and parameter settings are provided in the supplementary materials. Evaluation results after 100K training steps on the Carla benchmark are presented in Tab. 1. The episode reward curves during the training phase in the JW-Midnight scenario are depicted in Fig. 4, demonstrating the superiority of our approach.



(a) Single-modality policies. (b) Multi-modality policies.

Figure 4. Training performance under the JW-Midnight scenario.

5.2.1 Single-modality Policies

In the Midnight scenario, abnormal exposure can lead to the failure of RGB, resulting in the lowest performance result. DVS is capable of detecting useful events in extremely low-light conditions because of its high dynamic range, resulting in the highest performance in Midnight. However, in the Hardrain weather conditions, raindrops cause undesirable changes in pixel-level illumination, resulting in excessive noise from DVS. Consequently, under HB-Hardrain, RGB performs the best. Moreover, under JW-Hardrain, there is only a slight difference between RGB and the best-performing DVS-F, which is caused by a slight deviation due to the instability of RL sampling. Remarkably, DMR outperforms all the single-modality methods by a substantial margin in terms of reward and distance metrics. In summary, except for the HB-midnight scenario, where our method offers limited improvement, our method significantly surpasses single-modality methods, underscoring the advantages of multi-modality fusion.

[1] https://github.com/kyoran/DMR

Models	Metrics	Distance	Reward
M1	1 Branch	185±55	145±54
M2	2 Branch	194±85	141±78
M3	+Repel	214±74	181±77
M4	+Rec (ours)	**230±77**	**194±73**

Table 2. Effect of components in DMR.

5.2.2 Multi-modality Policies

Since DVS outperforms DVS-F in the majority of cases, we utilize RGB and DVS as the perceptual inputs for the multi-modality experiments. The learned policies of SOTA multi-modality methods often fail to match the performance achieved by single-modality methods. This could be attributed to the common adoption of multi-scale and attention mechanisms in current state-of-the-art multi-modality methods. These approaches often mix task-relevant information with accumulated noise, complicating the extraction of information crucial for downstream tasks. In contrast, our method offers a solution by explicitly eliminating noise and providing refined co-features for the RL. Compared to alternative multi-modality RL methods, our approach obviates the need for constructing intricate and resource-intensive fusion networks, while still attaining advantages in sample efficiency and learning performance.

5.3. Ablation Study

5.3.1 Effect of DMR Components

To assess the impact of each component, we incrementally incorporate individual components, resulting in a series of models labeled M1 to M4, shown in Tab. 2. Specifically, M1 utilizes solely the co-feature branch for input fusion. In M2, the upper and lower branches are employed, and the features from both branches are concatenated to feed policy learning. M3 and M4 represent the three-branch variants, integrating explicit feature decomposition. In M3, the co-features and modality-specific noises are repelled using the contrastive constraint as specified by Eq. (13). M4 builds upon M3 by incorporating the reconstruction decoder depicted in Eq. (10). DeepMDP is retained as our task predictive head and baseline RL algorithm.

Tab. 2 presents the performance of each model. It can be seen that M2 slightly improves on M1 in terms of distance while having little effect on reward. This is possibly because of the uncertainty in replay buffer sampling during RL training. In addition, with the introduction of three branches and contrastive constraints (M3), there is a significant improvement in both distance and reward. Finally, with the incorporation of the reconstruction decoder (M4), reward and distance further improve, indicating the necessity of the information completeness constraint. Furthermore, we analyze the CAMs across different modality

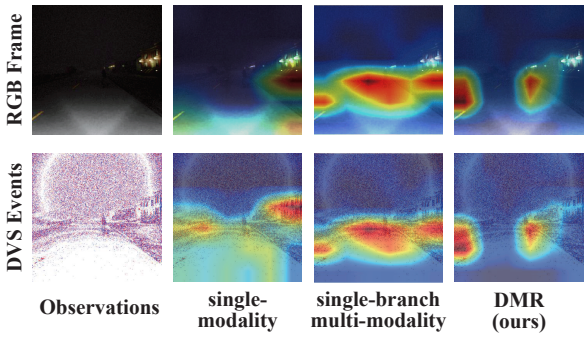
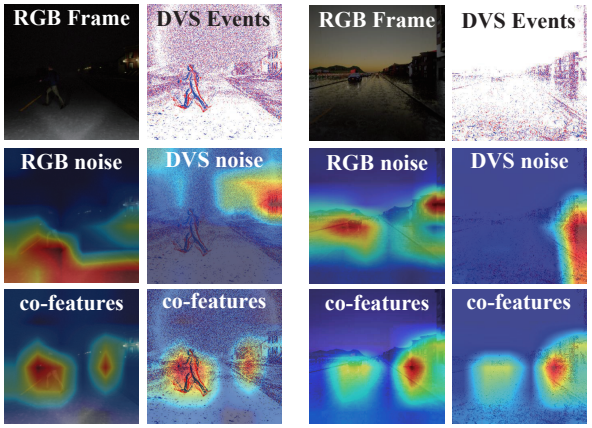


Figure 5. CAMs under different modality RL configurations in the JW-Midnight scenario.



(a) JW-Midnight: an extremely low-light condition.

(b) HB-Hardrain: a rapidly-changing illumination condition.

Figure 6. CAMs of DMR under different illumination conditions.

configurations, including single-modality models that take either RGB frames or DVS events as input (second column), a basic multi-modality model (M1) that takes both inputs (third column), and DMR (fourth column). From Fig. 5, it can be seen that the CAMs for single-modality models primarily highlight the front road and the adjacent buildings, activating an unnecessarily broad space. The simple multi-modality model without using decomposition and contrastive constraints generates a more focused area, but still contains task-irrelevant regions. In contrast, our method DMR effectively captures pertinent areas for RL while eliminating irrelevant regions. These areas precisely cover the pedestrians on the road and the left roadside, which are crucial cues for driving decision making.

5.3.2 Analysis of Modality-Specific Noise

We analyze the encoding capabilities of capturing both task-irrelevant noise and task-relevant features. Fig. 6 depicts the original observations and corresponding CAMs of DMR. In the extremely low-light condition (JW-Midnight), DVS can capture the front pedestrian while RGB camera suffers

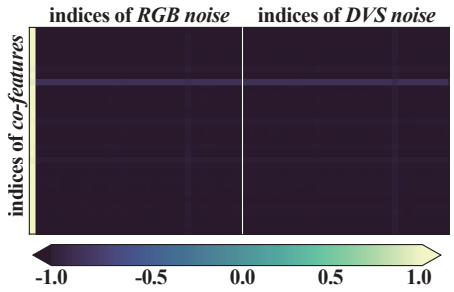


Figure 7. A similarity matrix example at the 100K'th training step.

from exposure failure. It can be seen that RGB noise highlights the high beam region on the road, while DVS noise is activated across a broader region, with the highest activation on the building. We can also see that the co-features attentively grasp the pedestrian and the right roadside simultaneously. In the rapidly-changing illumination condition (HB-Hardrain), DVS generates excessive event noise, while RGB can capture rich texture information. Notably, RGB noise mainly highlights brighter regions such as the front road and nearby vehicles and buildings, while DVS noise is prominent around puddles and splashing water. We observe that the co-features distinctly focus on the front cyclist, left vehicle, and right building, which are crucial for driving decision-making. CAMs of SOTA methods are provided in the supplementary materials.

Fig. 7 demonstrates the encoder's behavior in decomposition and discrimination through quantitative analysis. We present the similarity matrix between co-features and the modality-specific noises from RGB frames and DVS events, obtained from a training batch of 32 samples at the 100K'th training step. The color bar at the bottom represents the similarity ranging from low to high. The similarity is quantified by the dot product with a temperature parameter $\tau = 0.1$ as shown in Eq. (13). Notably, 32 samples share the same property as the set $\mathcal{K}$, where these samples are obtained from different time steps in different MDPs. Each row in the similarity matrix depicts similarities between the co-feature and itself, RGB noise, and DVS noise. We can see that the co-features exhibit strong coherence among themselves, while their similarity with the noises is remarkably low, illustrating a clear contrast.

5.3.3 Alternative DVS Backbones

Although we employed the same 4-layer CNN structure [46] in the lower DVS branch as in other branches, our architecture exhibits minimal susceptibility to network structure. Specifically, within the DVS branch, we evaluated a Spiking Neural Network (SNN) structure which is known for its suitability for temporal information encoding [9], as presented in Tab. 3. It can be seen that DMR with SNN is better than the other single-modality DVS methods, suggesting that DMR has the potential of working for alternative backbone structures. The detailed SNN structure is provided in the supplementary materials.

Metrics	DVS-F+SNN	DVS+SNN	DMR+SNN
Distance	81±15	101±10	**143±30**
Reward	35±12	45±2	**117±34**

Table 3. Performance of DMR with alternative DVS backbones under the HB-Hardrain scenario.

Metrics	RGB	Depth	LiDAR	RGB	
				+Depth	+LiDAR
Distance	91±61	103±12	113±9	**109±25**	**145±19**
Reward	70±63	69±10	68±13	**80±20**	**112±17**

Table 4. Performance of DMR with different modality combinations under the HB-Hardrain scenario.

5.3.4 Different Modality Combinations

While we focus on the RGB frames and DVS events modalities, the DMR framework has broader applicability to various modalities. We conducted performance evaluations on additional modality combinations using DMR, as presented in Tab. 4. The results demonstrate that when RGB is fused with LiDAR or Depth, the performance of DMR exceeds that of the individual modalities. Besides, by combining the results in Tab. 1, we can implicitly observe the efficacy of the complementary modalities in our DMR. The pronounced complementarity of RGB and DVS modalities necessitates our joint learning approach. The detailed modality settings are provided in the supplementary materials.

6. Conclusion

This paper explores a new decomposition perspective to address the multi-modality visual RL problem. We propose a novel three-branch multi-modality fusion framework, called DMR, designed for highly-complementary frame- and event-based visual modalities. DMR can explicitly extract task-relevant features from both modalities while mitigating the impact of irrelevant information and noise from each modality. Experimental results demonstrate the efficacy and superiority of DMR in policy performance. Our future work includes improving generalization and stability in more diverse and realistic scenarios in a sim2real fashion [35].

Acknowledgement The study was funded by the National Natural Science Foundation of China under contracts No. 62372010, No. 62027804, No. 61825101, No. 62088102 and No. 62202010, and the major key project of the Peng Cheng Laboratory (PCL2021A13). Computing support was provided by Pengcheng Cloudbrain.

References

[1] Rishabh Agarwal, Marlos C. Machado, Pablo Samuel Castro, and Marc G. Bellemare. Contrastive behavioral similarity embeddings for generalization in reinforcement learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021. 2

[2] Kristoffer Fogh Andersen, Huy Xuan Pham, Halil Ibrahim Ugurlu, and Erdal Kayacan. Event-based navigation for autonomous drone racing with sparse gated recurrent network. In *2022 European Control Conference (ECC)*, pages 1342–1348. IEEE, 2022. 2

[3] Pablo Samuel Castro. Scalable methods for computing state similarity in deterministic markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10069–10076, 2020. 2

[4] Guang Chen, Hu Cao, Jorg Conradt, Huajin Tang, Florian Rohrbein, and Alois Knoll. Event-based neuromorphic vision for autonomous driving: A paradigm shift for bio-inspired visual sensing and perception. *IEEE Signal Processing Magazine*, 37(4):34–49, 2020. 2

[5] Kaiqi Chen, Yong Lee, and Harold Soh. Multi-modal mutual information (mummi) training for robust self-supervised deep reinforcement learning. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4274–4280. IEEE, 2021. 3

[6] Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):12878–12895, 2023. 3, 5, 6

[7] Yuanzheng Ci, Chen Lin, Lei Bai, and Wanli Ouyang. Fastmoco: Boost momentum-based contrastive learning with combinatorial patches. In *European Conference on Computer Vision*, pages 290–306. Springer, 2022. 5

[8] Benjamin Eysenbach, Tianjun Zhang, Sergey Levine, and Russ R Salakhutdinov. Contrastive learning as goal-conditioned reinforcement learning. *Advances in Neural Information Processing Systems*, 35:35603–35620, 2022. 2

[9] Wei Fang, Yanqi Chen, Jianhao Ding, Zhaofei Yu, Timothée Masquelier, Ding Chen, Liwei Huang, Huihui Zhou, Guoqi Li, and Yonghong Tian. Spikingjelly: An open-source machine learning infrastructure platform for spike-based intelligence. *Science Advances*, 9(40):eadi1480, 2023. 8

[10] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020. 2

[11] Daniel Gehrig, Michelle Rüegg, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction. *IEEE Robotics and Automation Letters*, 6(2):2822–2829, 2021. 2

[12] Mathias Gehrig and Davide Scaramuzza. Recurrent vision transformers for object detection with event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13884–13893, 2023. 1, 2

[13] Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G. Bellemare. Deepmdp: Learning continuous latent space models for representation learning. In *International Conference on Machine Learning*, pages 2170–2179. PMLR, 2019. 2, 3, 5

[14] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018. 3, 5

[15] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018. 3, 5

[16] Philippe Hansen-Estruch, Amy Zhang, Ashvin Nair, Patrick Yin, and Sergey Levine. Bisimulation makes analogies in goal-conditioned reinforcement learning. In *International Conference on Machine Learning*, pages 8407–8426. PMLR, 2022. 2

[17] Javier Hidalgo-Carrió, Guillermo Gallego, and Davide Scaramuzza. Event-aided direct sparse odometry. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5771–5780. IEEE, 2022. 2

[18] Hanjiang Hu, Zuxin Liu, Sharad Chitlangia, Akhil Agnihotri, and Ding Zhao. Investigating the impact of multi-lidar placement on object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2550–2559, 2022. 5

[19] Yangru Huang, Peixi Peng, Yifan Zhao, Guangyao Chen, and Yonghong Tian. Spectrum random masking for generalization in image-based reinforcement learning. *Advances in Neural Information Processing Systems*, 35:20393–20406, 2022. 1, 3, 5

[20] Yangru Huang, Peixi Peng, Yifan Zhao, Yunpeng Zhai, Haoran Xu, and Yonghong Tian. Simoun: Synergizing interactive motion-appearance understanding for vision-based reinforcement learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 176–185, 2023. 1

[21] Xiaosong Jia, Yulu Gao, Li Chen, Junchi Yan, Patrick Langechuan Liu, and Hongyang Li. Driveadapter: Breaking the coupling barrier of perception and planning in end-to-end autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7953–7963, 2023. 5

[22] Yasser H. Khalil and Hussein T. Mouftah. Exploiting multimodal fusion for urban autonomous driving using latent deep reinforcement learning. *IEEE Trans. Veh. Technol.*, 72(3): 2921–2935, 2023. 3

[23] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. CURL: contrastive unsupervised representations for reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, pages 5639–5650. PMLR, 2020. 2

[24] Michael Laskin, Hao Liu, Xue Bin Peng, Denis Yarats, Aravind Rajeswaran, and Pieter Abbeel. Unsupervised reinforcement learning with contrastive intrinsic control. *Advances in Neural Information Processing Systems*, 35: 34478–34491, 2022. 2

[25] Minhyeok Lee, Chaewon Park, Suhwan Cho, and Sangyoun Lee. Spsn: Superpixel prototype sampling network for rgb-d salient object detection. In *European Conference on Computer Vision*, pages 630–647. Springer, 2022. 3

[26] Dianze Li, Jianing Li, and Yonghong Tian. Sodformer: Streaming object detection with transformer using events and frames. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2, 4, 6

[27] Jianing Li, Jia Li, Lin Zhu, Xijie Xiang, Tiejun Huang, and Yonghong Tian. Asynchronous spatio-temporal memory network for continuous event-based object detection. *IEEE Trans. Image Process.*, 31:2975–2987, 2022. 1, 2

[28] Jianing Li, Xiao Wang, Lin Zhu, Jia Li, Tiejun Huang, and Yonghong Tian. Retinomorphic object detection in asynchronous visual streams. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1332–1340, 2022. 2

[29] Siao Liu, Zhaoyu Chen, Yang Liu, Yuzheng Wang, Dingkang Yang, Zhile Zhao, Ziqing Zhou, Xie Yi, Wei Li, Wenqiang Zhang, et al. Improving generalization in visual reinforcement learning via conflict-aware gradient agreement augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23436–23446, 2023. 5

[30] Jinming Ma, Feng Wu, Yingfeng Chen, Xianpeng Ji, and Yu Ding. Effective multimodal reinforcement learning with modality alignment and importance enhancement. *arXiv preprint arXiv:2302.09318*, 2023. 3

[31] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013. 1, 3

[32] Masashi Okada and Tadahiro Taniguchi. Dreaming: Model-based reinforcement learning by latent imagination without reconstruction. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4209–4215. IEEE, 2021. 2

[33] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):1964–1980, 2019. 2, 4

[34] Tobias Renzler, Michael Stolz, Markus Schratter, and Daniel Watzenig. Increased accuracy for fast moving lidars: Correction of distorted point clouds. In *2020 IEEE international instrumentation and measurement technology conference (I2MTC)*, pages 1–6. IEEE, 2020. 2

[35] Sruthi Sudhakar, Jon Hanzelka, Josh Bobillot, Tanmay Randhavane, Neel Joshi, and Vibhav Vineet. Exploring the sim2real gap using digital twins. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20418–20427, 2023. 8

[36] Lei Sun, Christos Sakaridis, Jingyun Liang, Qi Jiang, Kailun Yang, Peng Sun, Yaozu Ye, Kaiwei Wang, and Luc Van Gool. Event-based fusion for motion deblurring with cross-modal attention. In *European Conference on Computer Vision*, pages 412–428. Springer, 2022. 2, 6

[37] Zhaoning Sun, Nico Messikommer, Daniel Gehrig, and Davide Scaramuzza. Ess: Learning event-based semantic segmentation from still images. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*, pages 341–357. Springer, 2022. 1, 2

[38] Abhishek Tomy, Anshul Paigwar, Khushdeep Singh Mann, Alessandro Renzaglia, and Christian Laugier. Fusing event-based and RGB camera for robust object detection in adverse conditions. In *2022 International Conference on Robotics and Automation, ICRA 2022, Philadelphia, PA, USA, May 23-27, 2022*, pages 933–939. IEEE, 2022. 2, 3, 6

[39] Sai Vemprala, Sami Mian, and Ashish Kapoor. Representation learning for event-based visuomotor policies. *Advances in Neural Information Processing Systems*, 34:4712–4724, 2021. 1, 2

[40] Yannick Verdié, Jifei Song, Barnabé Mas, Benjamin Busam, Aleš Leonardis, and Steven McDonagh. Cromo: Cross-modal learning for monocular depth estimation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3927–3937. IEEE, 2022. 2

[41] Celyn Walters and Simon Hadfield. Ceril: Continuous event-based reinforcement learning. *arXiv preprint arXiv:2302.07667*, 2023. 2

[42] Xiangjun Wang, Junxiao Song, Penghui Qi, Peng Peng, Zhenkun Tang, Wei Zhang, Weimin Li, Xiongjun Pi, Jujie He, Chao Gao, et al. Scc: An efficient deep reinforcement learning agent mastering the game of starcraft ii. In *International conference on machine learning*, pages 10905–10915. PMLR, 2021. 1

[43] Penghao Wu, Xiaosong Jia, Li Chen, Junchi Yan, Hongyang Li, and Yu Qiao. Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline. *Advances in Neural Information Processing Systems*, 35:6119–6132, 2022. 5

[44] Denis Yarats, Amy Zhang, Ilya Kostrikov, Brandon Amos, Joelle Pineau, and Rob Fergus. Improving sample efficiency in model-free reinforcement learning from images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10674–10681, 2021. 2

[45] Tao Yu, Zhizheng Zhang, Cuiling Lan, Yan Lu, and Zhibo Chen. Mask-based latent reconstruction for reinforcement learning. *Advances in Neural Information Processing Systems*, 35:25117–25131, 2022. 1, 2

[46] Yunpeng Zhai, Peixi Peng, Yifan Zhao, Yangru Huang, and Yonghong Tian. Stabilizing visual reinforcement learning via asymmetric interactive cooperation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 207–216, 2023. 1, 2, 5, 6, 8

[47] Amy Zhang, Rowan Thomas McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2021. 1, 2, 6

[48] Biao Zhang, Xiaoyuan Zhang, Baochen Wei, and Chenkun Qi. A point cloud distortion removing and mapping algorithm based on lidar and imu ukf fusion. In *2019 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, pages 966–971. IEEE, 2019. 2

[49] Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Delivering arbitrary-modal semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1136–1147, 2023. 2

[50] Yinuo Zhao, Kun Wu, Zhiyuan Xu, Zhengping Che, Qi Lu, Jian Tang, and Chi Harold Liu. Cadre: A cascade deep reinforcement learning framework for vision-based autonomous urban driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3481–3489, 2022. 1, 5

[51] Zhuyun Zhou, Zongwei Wu, Rémi Boutteau, Fan Yang, Cédric Demonceaux, and Dominique Ginhac. Rgb-event fusion for moving object detection in autonomous driving. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7808–7815. IEEE, 2023. 1, 2, 6

[52] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 3357–3364. IEEE, 2017. 1

[53] Yi-Fan Zuo, Jiaqi Yang, Jiaben Chen, Xia Wang, Yifu Wang, and Laurent Kneip. DEVO: depth-event camera visual odometry in challenging conditions. In *2022 International Conference on Robotics and Automation, ICRA 2022, Philadelphia, PA, USA, May 23-27, 2022*, pages 2179–2185. IEEE, 2022. 2