USC Viterbi
Information Sciences Institute

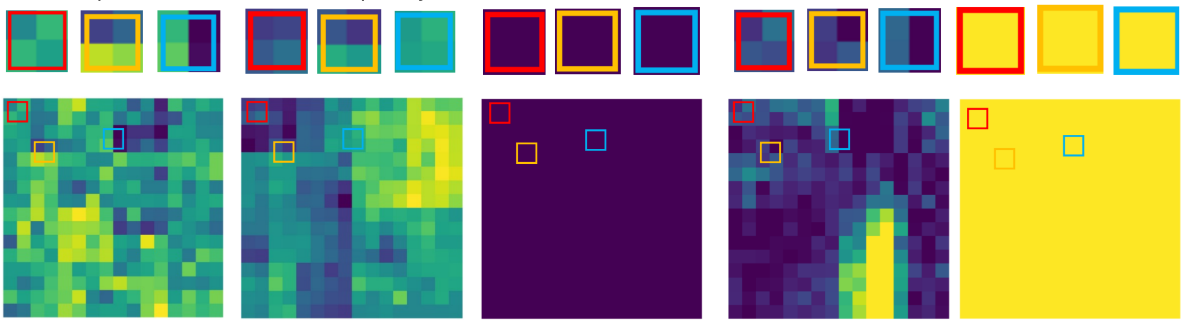
An Investigation on The Position Encoding in Vision-Based Dynamics Prediction

Jiageng Zhu*, Hanchen Xie*, Jiazhi Li, Mahyar Kkhayatkhoei, Wael AdbAlmageed

1. Motivation

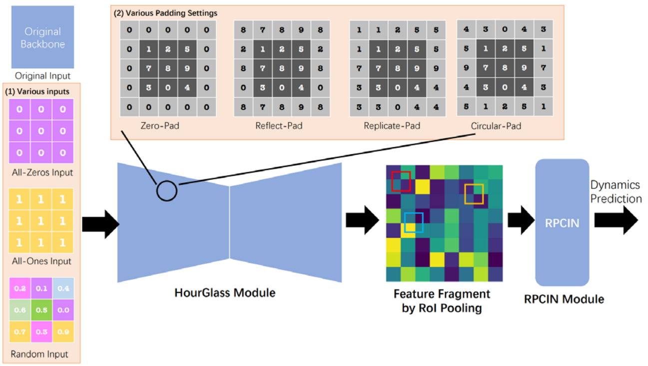
The paper aims to provide a comprehensive investigation into how position information is encoded and utilized in vision-based dynamics prediction models. This investigation is motivated by

- Previous success in vision-based dynamics prediction models was challenged by environment misalignments, suggesting a need for better understanding of how these models work.
- While prior work showed that object abstracts (like bounding boxes) could mitigate visual domain misalignment, the insight into using bounding boxes as object abstracts was under-explored.
- Empirical results in literature showed that object bounding boxes alone could provide sufficient position information for dynamics prediction through Region of Interest (RoI) Pooling. However, how this position information is implicitly encoded was overlooked.



(a) Visual Inputs Zero Pad w/ bias | (b) All-Zeros Inputs Zero Pad w/ bias | (c) All-Zeros Inputs Zero Pad w/o bias | (d) Fixed-Random Inputs Reflect Pad w/ bias | (e) All-Ones Inputs Reflect Pad w/ bias

2. Investigation Details



RPCIN forward propagation [P1]. b_i^t is the ball state feature extracted from feature map by RoI Pooling. f are various networks to analyze object dynamics.

$$e_i^t = f_A(f_O(b_i^t) + \sum_{j \neq i} f_R(b_i^t, b_j^t)),$$

$$z_i^t = f_Z(b_i^t, e_i^t),$$

$$b_i^{t+1} = f_P(z_i^t, z_i^{t-1}, ..., z_i^{t-T_{ref}+1})$$

3. Experiment Results

3.1 The differences on the feat map, introduced by proper padding setting, can be utilized by models to infer position information.

Table 1. Quantitative comparison of different padding modes with bias weights within CNN kernels trained on different types of input.

Padding Mode (w/ bias)	Zero		Reflect		Replicate		Circular	
Eval Period	P1 ↓	P2 ↓	P1 ↓	P2 ↓	P1 ↓	P2 ↓	P1 ↓	P2 ↓
Visual Inputs	$2.72_{\pm 0.31}$	$27.94_{\pm 1.08}$	$2.74_{\pm 0.30}$	$28.43_{\pm 1.21}$	$2.82_{\pm 0.42}$	$28.94_{\pm 1.11}$	$2.73_{\pm 0.42}$	$28.03_{\pm 1.09}$
All-Zeros Inputs	$2.97_{\pm 0.53}$	$29.83_{\pm 1.13}$	$\mathbf{144.34}_{\pm 0.21}$	$\mathbf{145.14}_{\pm 0.31}$	$\mathbf{144.35}_{\pm 0.31}$	$\mathbf{145.51}_{\pm 0.30}$	$\mathbf{144.42}_{\pm 0.30}$	$\mathbf{145.08}_{\pm 0.31}$
All-Ones Inputs	$3.11_{\pm 0.49}$	$30.48_{\pm 1.48}$	$\mathbf{144.43}_{\pm 0.30}$	$\mathbf{145.17}_{\pm 0.32}$	$\mathbf{144.43}_{\pm 0.31}$	$\mathbf{145.17}_{\pm 0.29}$	$\mathbf{144.43}_{\pm 0.32}$	$\mathbf{145.09}_{\pm 0.21}$
Fixed-Random Inputs	$2.91_{\pm 0.47}$	$30.03_{\pm 1.15}$	$3.01_{\pm 0.42}$	$31.48_{\pm 1.67}$	$3.04_{\pm 0.52}$	$29.56_{\pm 1.08}$	$3.17_{\pm 0.46}$	$30.46_{\pm 1.81}$
Random Inputs	$3.00_{\pm 0.55}$	$29.40_{\pm 0.96}$	$2.90_{\pm 0.41}$	$30.68_{\pm 1.09}$	$3.17_{\pm 0.48}$	$31.09_{\pm 1.09}$	$2.98_{\pm 0.39}$	$29.07_{\pm 1.73}$

Table 2. Quantitative comparison of different padding modes without bias weights within CNN kernels trained on different types of input.

Padding Mode (w/o bias)	Zero		Reflect		Replicate		Circular	
Eval Period	P1 ↓	P2 ↓	P1 ↓	P2 ↓	P1 ↓	P2 ↓	P1 ↓	P2 ↓
Visual Inputs	$2.82_{\pm 0.36}$	$28.31_{\pm 1.31}$	$2.84_{\pm 0.33}$	$29.02_{\pm 0.91}$	$2.88_{\pm 0.34}$	$29.02_{\pm 1.01}$	$2.93_{\pm 0.63}$	$29.95_{\pm 1.46}$
All-Zero Inputs	$\mathbf{144.41}_{\pm 0.21}$	$\mathbf{145.13}_{\pm 0.11}$	$\mathbf{144.43}_{\pm 0.29}$	$\mathbf{145.27}_{\pm 0.34}$	$\mathbf{144.31}_{\pm 0.20}$	$\mathbf{145.14}_{\pm 0.27}$	$\mathbf{144.42}_{\pm 0.31}$	$\mathbf{145.07}_{\pm 0.29}$
All-Ones Inputs	$3.36_{\pm 0.45}$	$31.45_{\pm 1.37}$	$\mathbf{144.37}_{\pm 0.21}$	$\mathbf{145.17}_{\pm 0.37}$	$\mathbf{144.43}_{\pm 0.29}$	$\mathbf{145.08}_{\pm 0.21}$	$\mathbf{144.43}_{\pm 0.36}$	$\mathbf{145.14}_{\pm 0.30}$
Fixed-Random Inputs	$3.15_{\pm 0.39}$	$31.83_{\pm 0.96}$	$3.26_{\pm 0.40}$	$31.98_{\pm 1.14}$	$3.22_{\pm 0.52}$	$30.56_{\pm 1.51}$	$3.21_{\pm 0.47}$	$31.46_{\pm 1.31}$
Random inputs	$3.19_{\pm 0.42}$	$31.36_{\pm 1.10}$	$3.09_{\pm 0.32}$	$31.03_{\pm 1.25}$	$3.08_{\pm 0.44}$	$31.12_{\pm 1.46}$	$3.02_{\pm 0.35}$	$30.07_{\pm 0.58}$

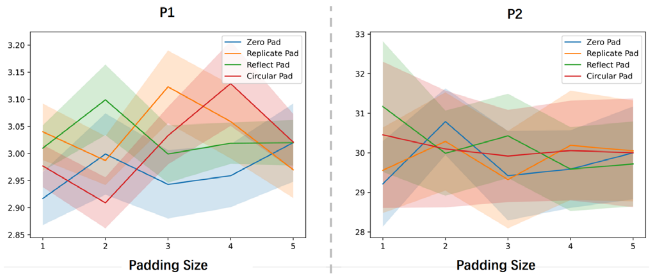


Figure 1. Quantitative comparison between different padding modes and padding size with bias weight trained on Fixed-Random Inputs.

3.2 When utilizing the environment information is necessary, the naïve differences on the feature map are insufficient for reaching the optimal solution .

Table 3. Quantitative comparison of different padding modes with Fix-Random Inputs on SimB-Border and SimB-Split datasets.

Dataset	SimB-Border		SimB-Split	
Eval Period	P1 ↓	P2 ↓	P1 ↓	P2 ↓
Baseline [19]	**1.13±0.01**	**9.57±0.12**	**0.91±0.02**	**7.73±0.21**
Zero	2.05±0.02	12.58±0.20	3.65±0.03	16.86±0.06
Reflect	2.04±0.02	12.23±0.25	3.61±0.05	16.71±0.38
Replicate	2.04±0.01	11.93±0.26	3.61±0.05	16.96±0.68
Circular	2.04±0.01	12.34±0.08	3.63±0.04	16.55±0.47

Table 4. Quantitative comparison of various inputs on SimB-Border and SimB-Split datasets. Results reported in [19]([P2]) was used as the Visual Inputs Performance..

Dataset	SimB-Border		SimB-Split	
Eval Period	P1 ↓	P2 ↓	P1 ↓	P2 ↓
Visual Inputs [19]	**1.13±0.01**	**9.57±0.12**	**0.91±0.02**	**7.73±0.21**
All-Zero Inputs	2.04±0.02	11.89±0.22	3.68±0.05	16.85±0.13
Fixed-Random Inputs	2.05±0.02	12.58±0.20	3.65±0.03	16.86±0.06

[P1] H. Qi, et al., "Learning Long-term Visual Dynamics with Region Proposal Interaction Networks," ICLR 2021
[P2] H. Xie, et al., "A Critical View of Vision-Based Long-Term Dynamics Prediction Under Environment Misalignment," ICML 2023

*: Equal Contributions; Corresponding Author: hanchenx@isi.edu