Visual Representation Learning with Stochastic Frame Prediction

Huiwon Jang[A], **Dongyoung Kim**[A], **Junsu Kim**[A], **Jinwoo Shin**[A], **Pieter Abbeel**[B], **Younggyo Seo**[A,C]
[A]KAIST, [B]UC Berkeley, [C]Dyson Robot Learning lab

KAIST · BAIR BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH · ICML International Conference On Machine Learning

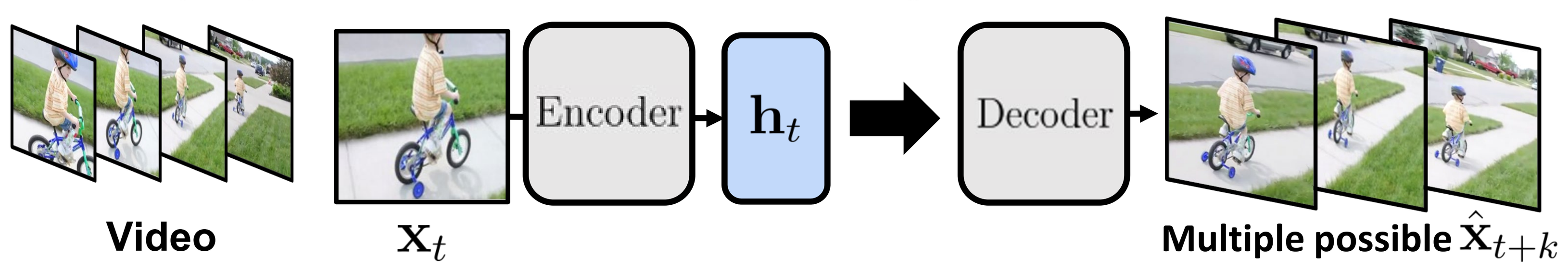
TL; DR. Learning stochastic frame prediction model from videos enhances the image representation to capture temporal information between frames.

Introduction

Learning image representation by **predicting the future** is promising direction. It enables models to understand temporal and causal relationships, improving their understanding of how the world operates.

However, predicting the future frame is *inherently under-determined*.

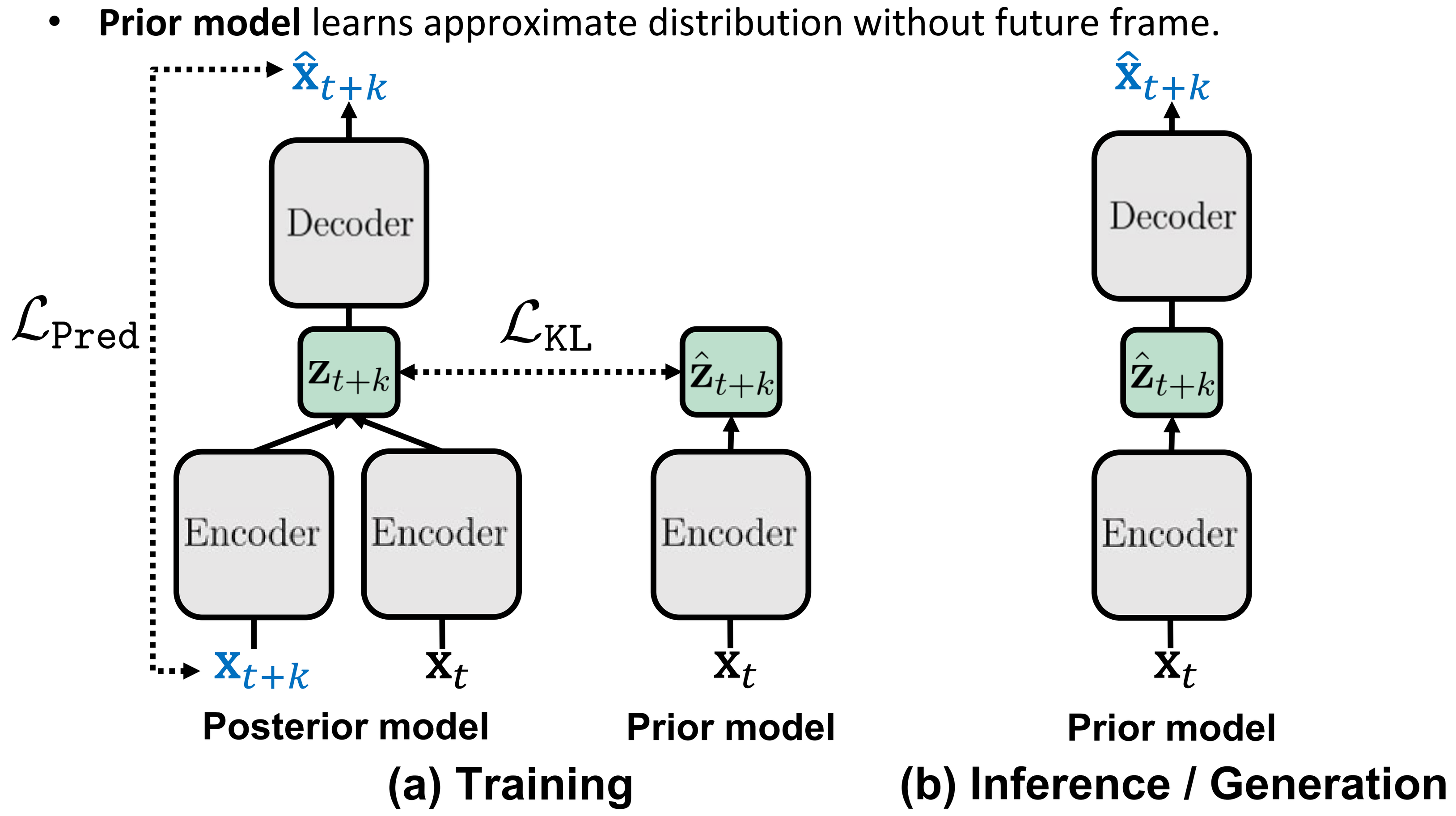
- **Multiple possible futures** can arise from a single current frame.



Q) How can we address the **ambiguity of the future** to learn representations from videos?

Key idea: Learning a stochastic frame prediction model with videos to learn *image* representations that capture temporal information between frames.

- **Posterior model** predict the future frame from posterior distribution.
- **Prior model** learns approximate distribution without future frame.

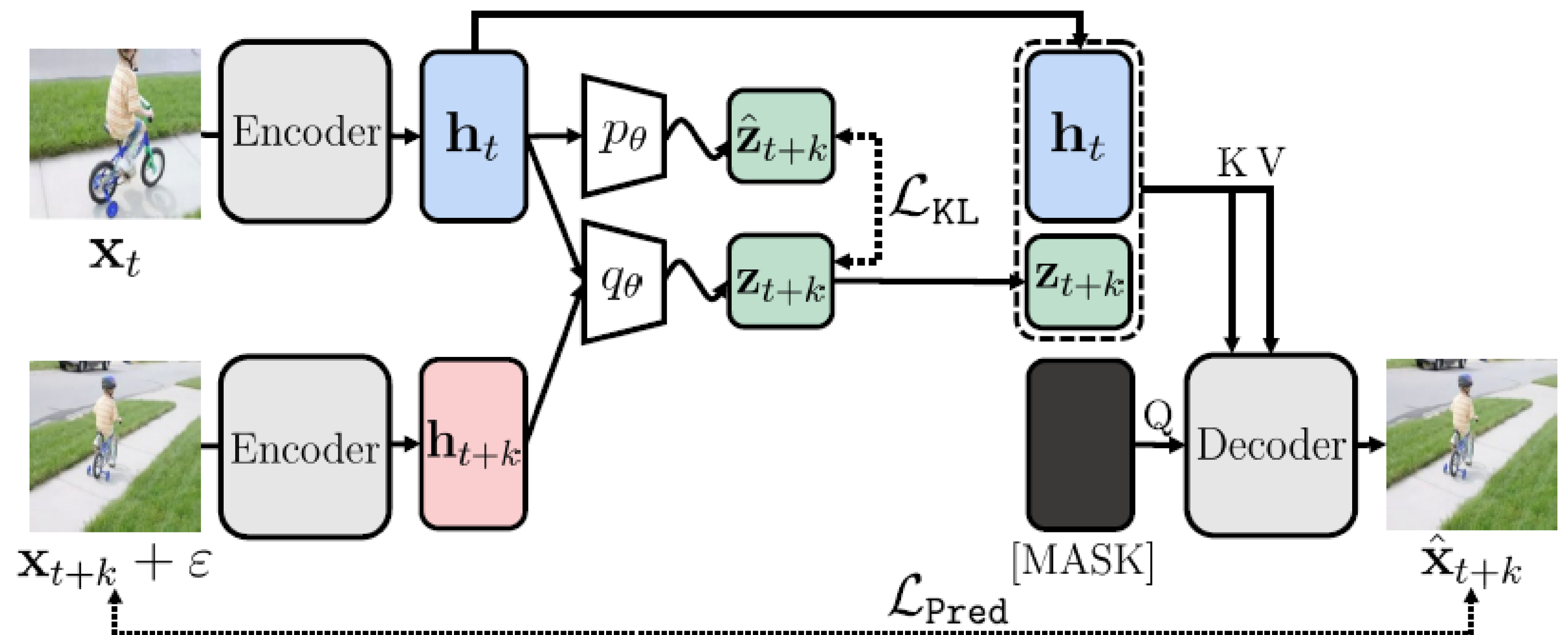


Summary of Contribution

We propose **RSP**, a framework for visual representation learning from videos via stochastic future frame prediction.

- We learn a stochastic frame prediction model to capture uncertainty in future frame prediction: *Posterior and learned prior.*
- Extensive experiments demonstrate that **RSP** consistently achieves competitive or superior performance to various SSL baselines on variety of tasks.

Method: RSP



Inputs. Given a video $\mathbf{x}$, we randomly sample two frames $\{\mathbf{x}_t, \mathbf{x}_{t+k}\} \in \mathbf{x}$.

Patch representations. We obtain patch representations.

- Encoder: $\begin{cases} \mathbf{h}_{t+k} = f_\theta^{\text{enc}}(\mathbf{x}_{t+k}+\varepsilon) \\ \mathbf{h}_t \quad = f_\theta^{\text{enc}}(\mathbf{x}_t) \end{cases}$
- Gaussian noise $\varepsilon \sim \mathcal{N}(0, \sigma)$ prevents copying pixels from $\mathbf{x}_{t+k}$ to predict $\hat{\mathbf{x}}_{t+k}$.

Posterior and learned prior. We predict the future frame from *posterior distribution*, which captures the uncertainty over future. A *prior* learns approximate distribution without access to the future frame.

- Posterior: $\mathbf{z}_{t+k} \sim q_\theta(\mathbf{z}_{t+k}|\mathbf{h}_t, \mathbf{h}_{t+k})$
- Learned prior: $\hat{\mathbf{z}}_{t+k} \sim p_\theta(\hat{\mathbf{z}}_{t+k}|\mathbf{h}_t)$

Decoder. We decode [MASK] tokens through cross-attention:

- Decoder: $\hat{\mathbf{x}}_{t+k} \sim p_\theta(\hat{\mathbf{x}}_{t+k}|\mathbf{h}_t, \mathbf{z}_{t+k})$

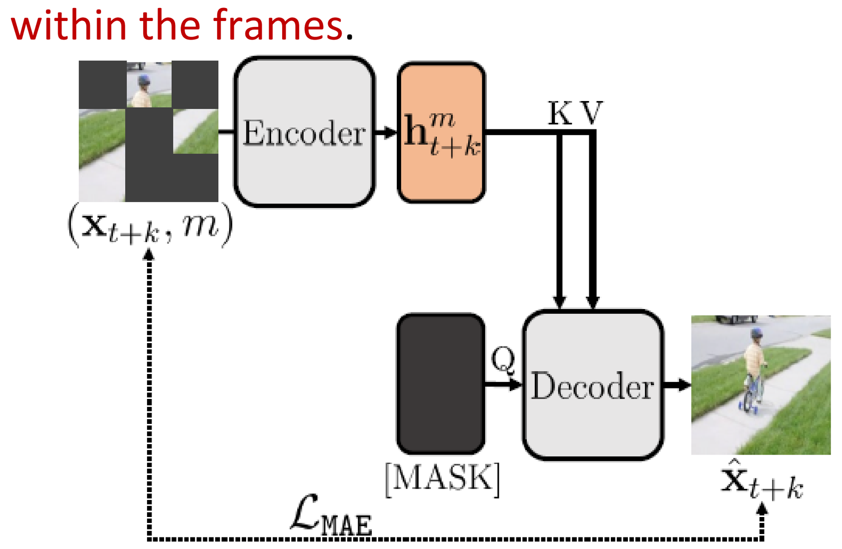
Objective. We train future frame prediction model to provide accurate prediction $\mathbf{x}_{t+k}$ while minimizing KL loss to learn the prior network for future predictions.

$$\mathcal{L}(\theta) = \mathbb{E}_{q_\theta(\mathbf{z}_{t+k}|\mathbf{x}_t,\mathbf{x}_{t+k})}\big[\underbrace{-\ln p_\theta(\mathbf{x}_{t+k}|\mathbf{x}_t,\mathbf{z}_{t+k})}_{\mathcal{L}_{\text{Pred}}} + \underbrace{\beta \text{KL}[q_\theta(\mathbf{z}_{t+k}|\mathbf{x}_t,\mathbf{x}_{t+k}) || p_\theta(\hat{\mathbf{z}}_{t+k}|\mathbf{x}_t)]}_{\mathcal{L}_{\text{KL}}}\big]$$

Masked autoencoding with shared decoder. We introduce *auxiliary MAE objective* to learn dense representation within the frames.

We *share the decoder* for both the frame prediction and MAE objective.

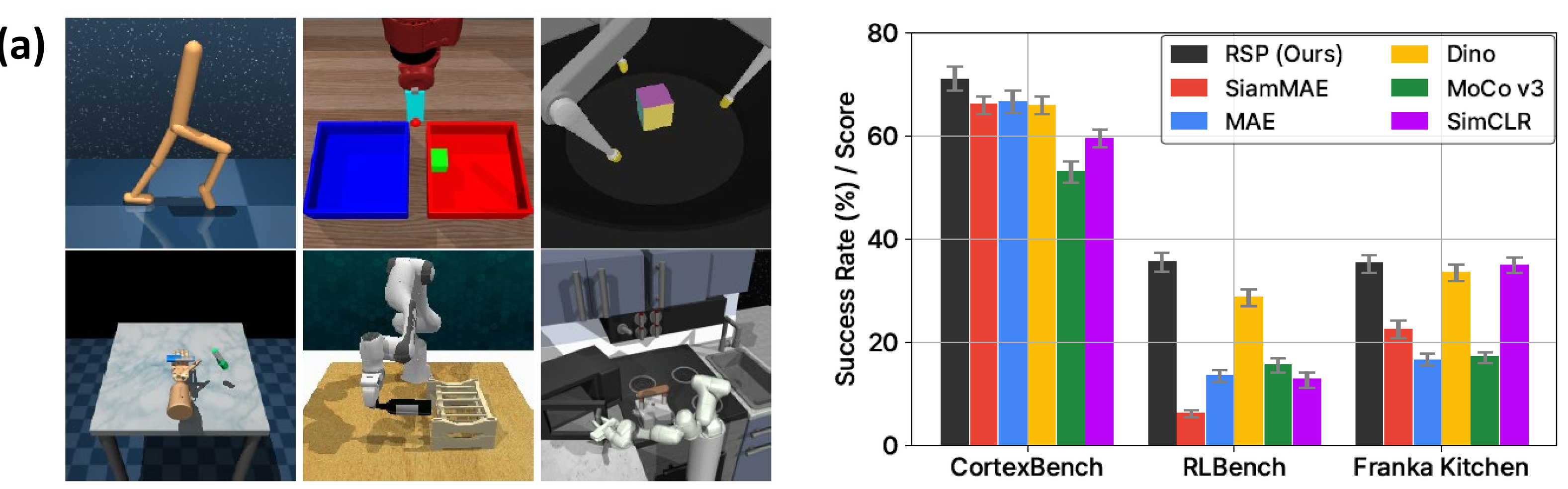
- Masking: $\mathbf{x}_{t+k}^m \sim p^{\text{mask}}(\mathbf{x}_{t+k}, m)$
- Encoder: $\mathbf{h}_{t+k}^m = f_\theta^{\text{enc}}(\mathbf{x}_{t+k}^m)$
- Decoder: $\hat{\mathbf{x}}_{t+k} \sim p_\theta(\hat{\mathbf{x}}_{t+k}|\mathbf{h}_{t+k}^m)$



Experiment

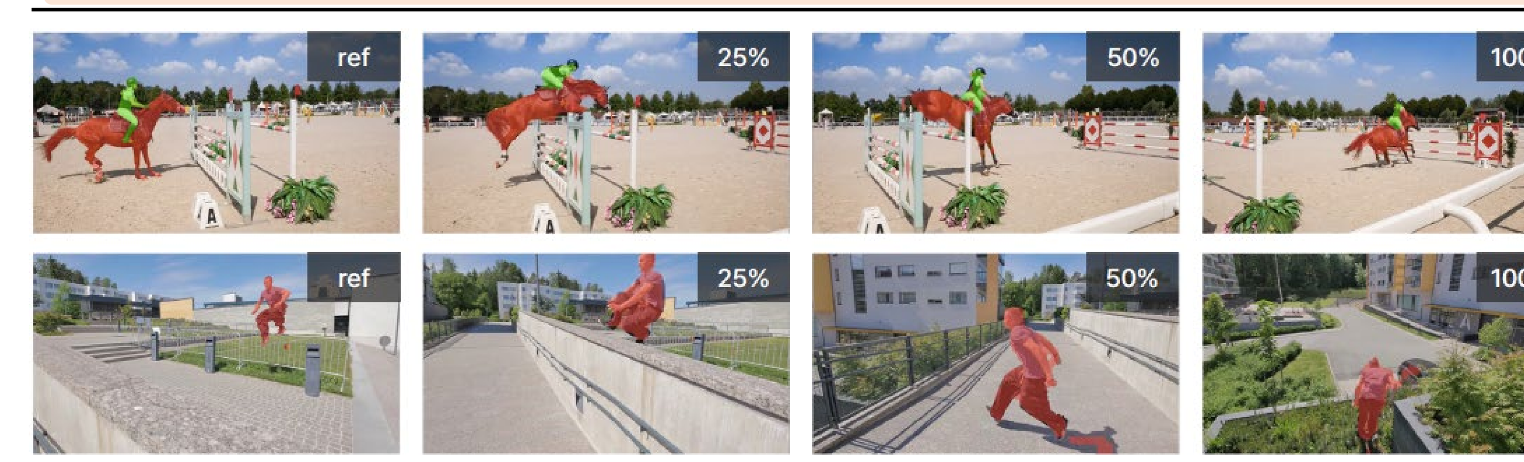
RSP consistently outperforms visual self-supervised learning methods in (a) vision-based robot learning tasks, and (b) video label propagation tasks.

(a)



(b)

Method	Architecture	DAVIS $\mathcal{J}\&\mathcal{F}_m$	DAVIS $\mathcal{J}_m$	DAVIS $\mathcal{F}_m$	VIP mIoU	JHMDB PCK@0.1	JHMDB PCK@0.2
SimCLR (Chen et al., 2020b)	ViT-S/16	53.9	51.7	56.2	31.9	37.9	66.1
MoCo v3 (Chen et al., 2021b)	ViT-S/16	57.7	54.6	60.8	32.4	38.4	67.6
Dino (Caron et al., 2021)	ViT-S/16	59.5	56.5	62.5	33.4	41.1	70.3
MAE (He et al., 2022)	ViT-S/16	53.5	50.4	56.7	32.5	43.0	71.3
SiamMAE (Gupta et al., 2023)	ViT-S/16	58.1	56.6	59.6	33.3	**44.7**	73.0
RSP (Ours)	ViT-S/16	**60.1**	**57.4**	**62.8**	**33.8**	44.6	**73.4**
RSP (Ours)	ViT-B/16	60.5	57.8	63.2	34.0	46.0	74.6



We conduct extensive ablation studies and analysis: Design choices for RSP

Stochastic	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
✗	54.4	50.7	58.1
✓	**60.1**	**57.4**	**62.8**

Deterministic prediction

Latent	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
Gaussian	54.1	52.9	55.9
Categorical	**60.1**	**57.4**	**62.8**

Stochastic latent variable

Same aug	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
✗	53.7	52.2	55.2
✓	**60.1**	**57.4**	**62.8**

Applying the same augmentation

w/ MAE	Decoder	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
✗	-	57.7	54.9	60.5
✓	Separate	58.1	55.4	60.7
✓	Shared	**60.1**	**57.4**	**62.8**

Auxiliary MAE objective

KL scale	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
0.1	56.1	52.9	59.3
0.01	**60.1**	**57.4**	**62.8**
0.001	59.1	56.6	61.5

KL objective scale

Future frame aug	Scale	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
None	-	58.3	56.1	60.6
Masking	0.75	57.7	54.8	60.6
Masking	0.95	55.8	52.7	58.9
Noise	0.1	58.4	56.0	60.7
Noise	0.5	**60.1**	**57.4**	**62.8**
Noise	1.0	58.9	56.3	61.4

Future frame augmentation