Learning Latent Linear Dynamics from Surgical Videos via Koopman-based Visual Representations

Luke Whatley[1], Andrei Cozma[1], Jie Fan[2], Hairong Qi[1], Fei Liu[1]

***Abstract*—Accurate prediction of future states from visual data is critical for surgical robotic control. We introduce a framework combining a convolutional autoencoder and a Koopman-inspired linear operator for modeling latent dynamics. The autoencoder extracts compact representations from surgical images, while the Koopman operator enforces linear temporal evolution in the latent space. Experiments on the AutoLaparo dataset confirm that our approach effectively captures key visual features and yields robust multi-step predictions, supporting applications such as motion planning and tool-tissue interaction analysis. This framework provides a promising foundation for real-time predictive modeling in surgical robotics.**

I. Introduction

Predicting future states from high-dimensional surgical video data remains a key challenge. In minimally invasive procedures, endoscopic videos offer rich visual cues, but their complexity limits the effectiveness of traditional prediction methods.

Convolutional autoencoders (CAEs) have demonstrated success in visual representation learning, but typically lack structured latent spaces suitable for dynamic prediction [1]. Koopman operator theory, however, provides linear approximations of nonlinear dynamics in a latent space [2], and has been effectively combined with autoencoders for temporal consistency in predictions [3], [4], though not extensively explored in surgical contexts.

We propose a preliminary framework combining a convolutional autoencoder with a linear Koopman operator for short-term surgical video prediction at 25 fps. Our approach compresses surgical images into latent embeddings, modeling temporal dynamics linearly. This provides an interpretable method to analyze tool-tissue interactions and supports robotic surgical planning. Our key contributions include:

- Integration of a convolutional autoencoder with a Koopman operator for surgical video prediction.
- A training strategy balancing reconstruction and linear latent-space prediction.
- Preliminary validation on the AutoLaparo dataset [5] demonstrating effective short-term predictions.

II. Methods

A. Backgrounds

Koopman Operator Theory. Originally introduced by Koopman [6] and extended by Mezić [2], Koopman operator theory models nonlinear dynamics via linear evolution in a lifted space. The operator $\mathcal{K}$ acts on observables $g(\mathbf{x})$ as:

$$\mathcal{K}g(\mathbf{x}) = g(F(\mathbf{x}))$$

Deep Learning for Koopman Analysis. Recent approaches integrate Koopman theory with deep autoencoders [3], [7]. An encoder $\varphi : \mathbb{R}^n \to \mathbb{R}^d$ maps input $\mathbf{x}_t$ to a latent state $\mathbf{y}_t$, and a decoder φ^{-1} reconstructs it:

$$\mathbf{y}_t = \varphi(\mathbf{x}_t), \quad \hat{\mathbf{x}}_t = \varphi^{-1}(\mathbf{y}_t)$$

A linear Koopman operator $\mathbf{K} \in \mathbb{R}^{d\times d}$ is learned to evolve the latent state:

$$\mathbf{y}_{t+1} = \mathbf{K}\mathbf{y}_t$$

Future states are predicted via:

$$\hat{\mathbf{x}}_{t+k} = \varphi^{-1}(\mathbf{K}^k\varphi(\mathbf{x}_t))$$

Visual Dynamics in Surgery. Surgical videos pose challenges due to the complexity of tool-tissue interactions. While deep learning has shown success in learning compact representations [1], most efforts emphasize reconstruction over temporal prediction, leaving Koopman-based forecasting relatively unexplored in this domain.

B. Methodology

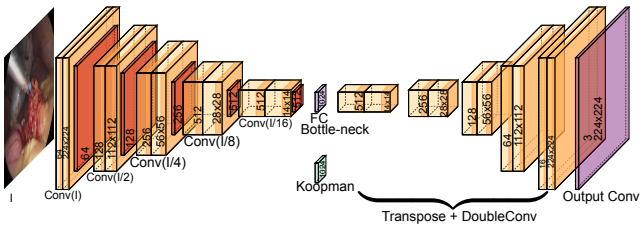


Fig. 1. Proposed convolutional autoencoder and Koopman operator framework for surgical video prediction.

Our framework combines a convolutional autoencoder and a linear Koopman operator to predict visual dynamics in surgical videos (Fig. 1). The encoder $\varphi : \mathbb{R}^n \to \mathbb{R}^d$ maps input frames $\mathbf{x}_t$ to compact latent vectors $\mathbf{y}_t$. A decoder $\varphi^{-1} : \mathbb{R}^d \to \mathbb{R}^n$ reconstructs images from these latent representations:

$$\mathbf{y}_t = \varphi(\mathbf{x}_t), \quad \hat{\mathbf{x}}_t = \varphi^{-1}(\mathbf{y}_t).$$

Latent dynamics evolve linearly via the learned Koopman operator $\mathbf{K} \in \mathbb{R}^{d\times d}$:

$$\mathbf{y}_{t+1} = \mathbf{K}\mathbf{y}_t, \quad \hat{\mathbf{x}}_{t+k} = \varphi^{-1}(\mathbf{K}^k\varphi(\mathbf{x}_t)).$$

We define the training loss component by:

[1] Department of Electrical Engineering and Computer Science, University of Tennessee Knoxville, Knoxville, TN 37996, USA {lwhatle3, acozma}@vols.utk.edu, {fliu33, hqi}@utk.edu
[2] jie.fan.0414@gmail.com

TABLE I
SUMMARY OF TRAINING HYPERPARAMETERS

Parameter	Value	Parameter	Value
in_dim	256	H	4
mask_dilation_radius	20	batch_size	16
epochs	15	n_layers	4
dropout	0.0	base_channels	16
latent_dim	512	learning_rate	0.0001
Kinit	ORTHO	optimizer	Adam
loss_function	MSELoss		

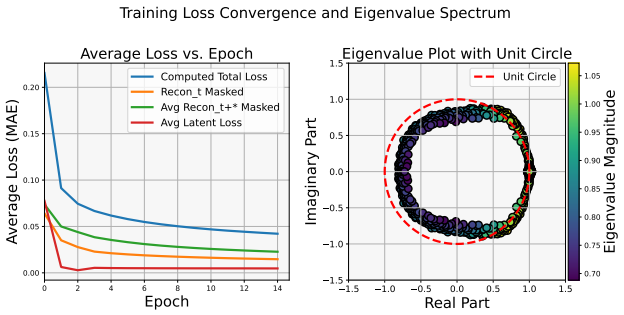


Fig. 2. Average training loss plot for end-to-end training (left) and Eigenvalue components and magnitudes of trained Koopman operator (right).

- **Reconstruction loss** $\mathcal{L}_{\text{recon}}$: preserves key visual features by ensuring accurate input reconstruction.
- **Prediction loss** $\mathcal{L}_{\text{pred}}$: enforces temporal consistency between predicted and actual future states via the Koopman.
- **Linearity loss** $\mathcal{L}_{\text{lin}}$: encourages latent dynamics to follow $\mathbf{y}_{t+1} = \mathbf{K}\mathbf{y}_t$, guiding the encoder to a Koopman-consistent representation.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \alpha_1 \mathcal{L}_{\text{pred}} + \alpha_2 \mathcal{L}_{\text{lin}}, \tag{1}$$

where α_1, α_2 are hyperparameters. To emphasize relevant surgical interactions, we constrain reconstruction loss using dilated masks around tool-tissue regions.

III. RESULTS

A. Training Loss Convergence

We first evaluate convergence by plotting loss components across epochs. The total loss—including reconstruction, future prediction, and linearity constraints—stabilizes around epoch 15, achieving a masked-region reconstruction MSE of approximately 0.0132. This indicates effective reconstruction of tool-tissue interactions, though fine details may be slightly reduced.

Training was performed using the entire AutoLaparo dataset (10-second videos at 25 fps) [5], sampling $N+1$ sequential frames with sliding-window sampling and random augmentation. Hyperparameters are summarized in Table I, and the loss progression is shown in Fig. 2-Left. The learned Koopman operator converges to a linear mapping for reliable future state prediction, with linearity regularization constraining latent embeddings to maintain prediction accuracy.

B. Eigenvalue Spectrum

We analyze the eigenvalues of the learned Koopman operator K on the complex plane (Fig. 2-Right). Most eigenvalues cluster near or within the unit circle, indicating stable or mildly decaying latent dynamics. A few eigenvalues slightly exceed magnitude one (up to 1.07), suggesting potential minor drift during extended predictions, though short-term forecasts remain robust.

C. Multi-Step Prediction Error

We evaluate predictive accuracy using mean squared error (MSE) between predicted and ground-truth frames over multiple future timesteps. Table II reports average MSE for 1–4 step horizons, showing gradual error increase with longer predictions while remaining acceptable within 4 steps.

TABLE II
RECONSTRUCTION MSE AT VARYING PREDICTION HORIZONS ON THE TEST SET.

Horizon (k)	t	t+1	t+2	t+3	t+4
MSE	0.0132	0.0178	0.0196	0.0208	0.0227

D. Feature Representation

The learned Koopman operator demonstrates reliable multi-step predictive capability within short horizons. However, the presence of eigenvalues outside the unit circle (modes slightly larger than 1) suggests these might introduce reconstruction drift over longer time horizons. Nevertheless, for practical applications in minimally invasive surgery, these short-term predictions remain highly beneficial and sufficiently accurate.

IV. DISCUSSION

Our results demonstrate that combining convolutional autoencoders with a linear Koopman operator effectively models and predicts short-term dynamics in surgical videos. The convergence of training losses (Fig. 2) highlights the model's capability to balance accurate reconstruction with linear predictive constraints in latent space.

Future work could enhance long-term stability through eigenvalue regularization or spectral radius constraints.

REFERENCES

[1] I. Higgins, L. Matthey, X. Glorot, A. Pal, C. Burgess, S. Botvinick, R. Gregor, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," *Proc. ICLR*, 2017.

[2] I. Mezić, "Spectrum of the Koopman operator, spectral expansions in functional spaces, and state-space geometry," *Journal of Nonlinear Science*, vol. 27, pp. 897–915, 2017.

[3] B. Lusch, J. N. Kutz, and S. L. Brunton, "Deep learning for universal linear embeddings of nonlinear dynamics," *Nature Communications*, vol. 9, no. 1, p. 4950, 2018.

[4] N. Takeishi, Y. Kawahara, and T. Yairi, "Learning Koopman invariant subspaces for dynamic mode decomposition," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[5] Z. Wang, B. Lu, Y. Long, F. Zhong, T.-H. Cheung, Q. Dou, and Y. Liu, "Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 486–496.

[6] B. O. Koopman, "Hamiltonian systems and transformation in Hilbert space," *Proceedings of the National Academy of Sciences*, vol. 17, no. 5, pp. 315–318, 1931.

[7] E. Yeung, S. Kundu, and N. Hodas, "Learning deep neural network representations for Koopman operators of nonlinear dynamical systems," *American Control Conference (ACC)*, pp. 4832–4839, 2019.