Lens-JEPA: Physics Informed Joint Embedding Predictive Architecture for Gravitational Lensing

J Rishi
Department of Computational and Data Sciences
Indian Institute of Science
Bangalore, 560012, India
rishij@iisc.ac.in

Pranath Reddy Kumbam
University of Florida,
Gainesville, FL 32611, USA
kumbam.pranath@gmail.com

Michael W. Toomey
Center for Theoretical Physics, Massachusetts Institute of Technology,
Cambridge, MA 02139, USA
mtoomey@mit.edu

Sergei Gleyzer
Department of Physics & Astronomy, University of Alabama,
Tuscaloosa, AL 35401, USA
sgleyzer@ua.edu

Abstract

We introduce Lens-JEPA, a novel approach towards building a foundation model for gravitational lensing images that extends I-JEPA to the astrophysical domain. Although recent advances in foundation models have transformed vision and language tasks by enabling generalization and transfer across other applications, astrophysical imaging still lacks such a framework. To address this gap, we develop a key component of Lens-JEPA, which is a Physics Encoder that introduces a transformer guided by the lensing equation, that combines the representational power of Vision Transformers with the rigor of Physics Informed Neural Networks (PINNs). This approach enhances the representation capabilities of gravitational lensing. This paper demonstrates the effectiveness of Lens-JEPA via a classification task, which shows that Lens-JEPA surpasses current top baseline architectures, highlighting the benefits of incorporating physics into transformer models. Although this study focuses on classification, it provides a foundation across other applications such as lens detection, mass modeling, and super-resolution, moving toward a foundation model for gravitational lensing

1 Introduction

Gravitational lensing, which occurs when light from a distant source is deflected by a massive object situated between the source and the observer, is a powerful probe in astrophysics and cosmology. Strong gravitational lensing, in particular, is very sensitive to the distribution of dark matter on

Machine Learning and the Physical Sciences Workshop, NeurIPS 2025.

subgalactic scales and provides an enhanced view of background sources, serving as a probe for the high redshift Universe. Strong gravitational lensing has established itself as a powerful probe of the nature of dark matter via substructure, from analyses of lensed quasars[1], results from ALMA [2], and a myriad of extended lensing images [3, 4, 5], among other studies. In the last several years, machine learning has been applied widely to address challenges related to gravitational lensing. Indeed, it is particularly well-suited for this field, as the analysis of even a single lens for the study of substructure can be computationally intensive. The applications of machine learning has ranged from classification [6, 7], regression [8, 9], segmentation analysis[10], domain adaptation[11], to anomaly detection[12].

Despite the progress of machine learning in gravitational lensing, most existing models remain highly task-specific and lack transferability. Training separate models for classification, regression, or super resolution not only requires large labeled datasets but also leads to fragmented pipelines with limited reusability. This motivates the development of it foundation models [13, 14, 15, 16], which aim to learn general-purpose representations from vast collections of unlabeled data and can be adapted to a variety of downstream tasks. Self-supervised[17, 18, 19, 20] learning provides a natural pathway toward such models, as it exploits the inherent structure of lensing images without requiring explicit labels. In the context of astrophysics, where simulations are expensive and labeled observations are scarce, a foundation model has the potential to unify disparate tasks such as lens detection, parameter estimation, and substructure analysis within a single framework, while significantly reducing the need for retraining and labeled supervision.

Within the landscape of self-supervised approaches for foundation models, the Image based Joint Embedding Predictive Architecture (I-JEPA)[21] has emerged as a compelling framework. Unlike generative methods that reconstruct data at the pixel level, I-JEPA learns to predict latent representations across masked contexts, yielding scalable and semantically rich embeddings[22, 21]. This makes it particularly suitable for scientific imaging domains such as gravitational lensing, where the goal is to capture high-level physical features rather than pixel-perfect reconstructions. However, conventional I-JEPA is designed to be domain-agnostic and does not explicitly incorporate astrophysical constraints. To overcome this limitation, we propose a `Lens-JEPA` for gravitational lensing images, where the lensing equation and relevant physical information are embedded into the learning process. By integrating physics into the representation space, our model leverages the strengths of self-supervised foundation models while ensuring physical consistency, enabling broad applicability across tasks such as lens detection, mass modeling, and probing dark matter substructure. Although in this paper we show results for only classification. future work includes other tasks, such as regression, lense finding, super resolution e.t.c.

2 Datasets

We use two simulated galaxy-galaxy strong lensing datasets generated with `lenstronomy` [23], each mimicking distinct observational instruments. Images are 150 × 150 pixels, single-channel, with Gaussian and with a signal-to-noise ratio chosen such that SNR$\sim$ 25. We have modeled the lensed background galaxies with a simple Sérsic light profile. The lensing galaxy's dark matter halo is modeled as a Singular Isothermal Sphere (SIS). Altogether, we model three substructure classes: (1) standard CDM substructure with truncated NFW subhalos; (2) axion dark matter with a mass of $m \sim 10^{-23}$ eV with vortex-like defects [24] (3) no-substructure baseline.

From this we have two base data sets tuned for different mock telescopes which we denote Models A, and B. Model A represents a generic mock data set for an arbitrary instrument with a Gaussian PSF of 0.05 arcsecond resolution. Models B represent mock Euclid instruments respectively. In this work we use Model A for training due to its generic, instrument-agnostic setup, enabling models to learn core lensing features. Models B are used for finetuning to adapt to realistic Euclid conditions.

We initially utilized 10,000 simulations from Model A for pretraining. For downstream tasks, we employed 3,000 simulations per category (axion, CDM, and without substructure) derived from Models B, implementing an 80:20 train-test division across the three dark matter models.

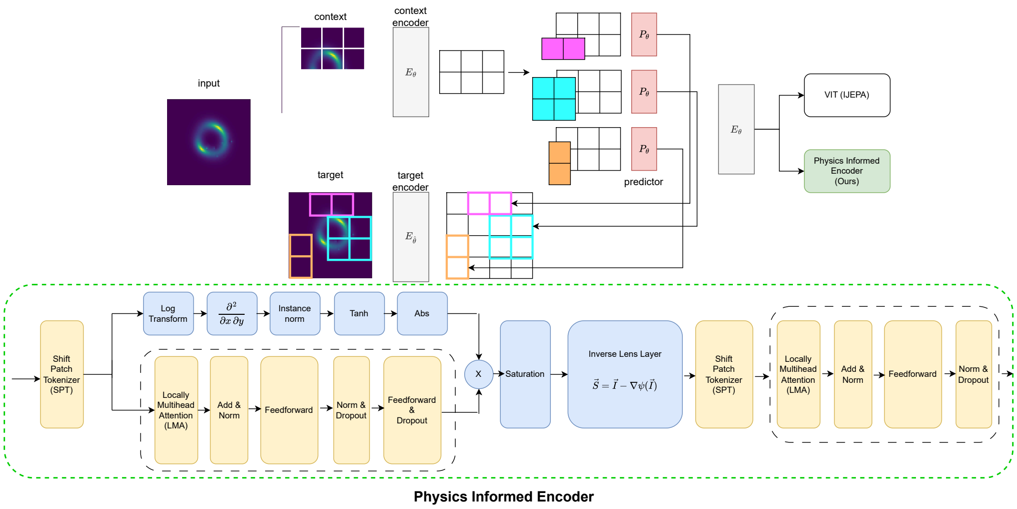

Figure 1: `Lens-JEPA Architecture:` Modified I- JEPA framework where the standard ViT encoder is replaced with a physics-informed encoder incorporating gravitational lensing equations. The input image is processed through context and target encoders, with masked patches predicted by a predictor network. The physics-informed encoder (bottom, green box) applies gravitational lensing transformations including the inverse lens equation before attention-based feature extraction, enabling the model to learn representations constrained by the physical principles of gravitational lensing.

3 Methodology

Developing a foundation model for gravitational lensing is a crucial extension to circumvent the challenges of training from scratch for various tasks. Thus, creating a versatile model for all tasks is essential because these datasets are rich in information yet computationally demanding to examine. Although several self-supervised methods have been suggested for constructing such models, our attention is on I-JEPA [21], which has demonstrated strong capabilities in learning general-purpose representations. Nonetheless, applying I-JEPA directly to gravitational lensing images presents difficulties, as capturing the complex physical nuances of lensing goes beyond basic representation learning. To tackle this challenge, we propose `Lens-JEPA`, a physics-informed extension that integrates two main components: (i) a transformer backbone designed to capture intricate dependencies within lensing images, and (ii) the inclusion of the lens equation via a specialized module in the encoder, embedding the essential physics of strong lensing into the architecture itself. Further details are explored in subsequent sections.

3.1 Preliminaries

Joint-Embedding Predictive Architectures (JEPA) [22] represent a class of self-supervised learning methods that aim to learn representations by matching the embeddings of context and target, rather than learn representations of the raw data directly. Learning to match embeddings of context and target, as in JEPA, is better suited than direct input reconstruction because it emphasizes capturing high-level semantic relationships rather than low-level pixel statistics. Unlike other selfsupervised techniques [17, 18, 19, 20], which may overfit to superficial details of the input, JEPA style objective encourage representations that are more invariate, robust and transferable to downstream tasks. I-JEPA [21] is a self-supervised framework designed for images that predicts the representations of masked image blocks from visible context blocks, using a Vision Transformer (ViT) backbone. I-JEPA avoids pixel-level reconstruction, focusing on semantic abstraction and enabling efficient, scalable pre-training for visual tasks.

3.2 Lens-JEPA

In Lens-JEPA we have introduced new encoder where we explicitly incorporate the governing physics of strong lensing by leveraging the *lens equation* (see [25] for more details). To capture the gravitational potential of the lensing galaxy, we include an analytic form of the potential for a *Singular Isothermal Sphere (SIS)* profile as an ansatz [26]. This choice provides a first-order approximation to the potential of the dark matter halo. The lens equation in its dimensionless form is given by:

$$\vec{S} = \vec{I} - \nabla\Psi(\vec{I}), \tag{1}$$

where $\mathbf{S} = (x_s, y_s)$ is the dimensionless source-plane position of the galaxy, $\mathbf{I} = (x_i, y_i)$ is the observed image-plane position, and $\nabla\Psi(\mathbf{I}) = (\Psi_x(x_i, y_i), \Psi_y(x_i, y_i))$ represents the gradient of the dimensionless gravitational potential. Since only $\mathbf{I}$ is observed, assumptions about the lens potential must be made. We adopt the following ansatz:

$$\Psi(x_i, y_i) = k(x_i, y_i) \cdot \Psi_{\text{SIS}}(x_i, y_i), \tag{2}$$

where

$$\Psi_{\text{SIS}}(x_i, y_i) = \sqrt{x_i^2 + y_i^2}, \tag{3}$$

as given in [27], and $k(x_i, y_i)$ is a learnable function predicted from the data. To estimate k_{ij} at each pixel (i, j), the encoder employs a ViTSD [28], trained on lensed images and their transformations to capture subtle variations in gradients induced by dark matter substructures. This process provides the potential $\Psi(x_i, y_i)$, enabling reconstruction of the source galaxy image.

This physics encoder integrates both physics and transformer-based learning by combining *Shifted Patch Tokenization (SPT)* and *Locality Self-Attention (LSA)* [28]. The reconstructed source galaxy, derived from solving the gravitational lens equation using the estimated potential, is tokenized with SPT and used as the input to the transformer blocks. This hybrid design enhances model adaptability across diverse lensing configurations and improves performance by embedding physical priors directly into the learning framework. This architecture is inspired from [29]. Overview of architecture is shown in Figure 1.

4 Experiments

To assess the performance of our model, we initially pre-trained the Lens-JEPA on the model A dataset, as detailed in Section 2. We then fine-tuned the model using the model B dataset. This was followed by a series of comparisons with state-of-the-art supervised models trained on the model B dataset, including Vision Transformer (ViT) [30], Vision Transformers for Small Datasets (ViTSD) [28], Convolutional Transformer (CvT) [31], Class-Attention in Image Transformers (CaiT) [32], and Residual Networks (ResNet) [33]. All these models share a similar number of parameters and employ the AdamW [34] optimizer with a learning rate of 0.00001, using the CrossEntropy loss function over 50 training epochs. Performance was assessed across several classification metrics.

5 Results and Discussion

Table 1 presents the performance comparison of different models on Model B dataset. Conventional supervised architectures such as ResNet, ViT, CaiT, and ViTSD achieve competitive results but fall short in capturing the underlying physical priors of gravitational lensing. While Lensformer and I-JEPA improve upon these baslines, the integration of physics-based inductive biases in Lens-JEPA leads to consistent gains across all metrics. Notably, Lens-JEPA achieves the highest accuracy (0.9120) and outperforms both supervised and original I-JEPA, with ROC AUC scores of 0.97 (Axion), 0.96 (CDM), and 1.00 (No subs). These results highlight the effectiveness of embedding lensing physics into the JEPA framework, positioning Lens-JEPA as a strong foundation model for astrophysical tasks. Importantly, the improvements are consistent across different dark matter scenarios, underscoring the model's robustness. this establishes a promising path toward scalable foundation models in astrophysics that unify physics-informed reasoning with modern transformer architectures.

Table 1: Performance Metrics of Different Models

Model	Model B			
	Accuracy	Roc AUC Score		
		Axion	CDM	No Subs
Resnet18	0.8207	0.94	0.88	0.97
ViT	0.8931	0.96	0.94	0.99
CaiT	0.8065	0.94	0.85	0.95
ViTSD	0.8838	0.96	0.93	0.99
Lensformer	0.8969	0.94	0.92	0.98
`I-JEPA`	0.9017	0.96	0.95	0.98
`Lens-JEPA`	**0.9120**	**0.97**	**0.96**	**1.00**

Despite our focus on classification, the results with `Lens-JEPA` demonstrate a robust foundation for a diversity of future gravitiational lensing analyses. Indeed, follow-up analyses for lens detection, super-resolution, and regression are natural to the work done here. Given the difficulty of probing the microphysical nature of dark matter via lensing, leveraging the added power of physics-informed architecture like `Lens-JEPA` will better position future analyses to maximize extraction of information from lensing data. However, limitations remain, including the complexity of training foundation models and our current reliance on simulated datasets. Future work will necessitate optimizing efficiency and validating performance on real world data sets. `Lens-JEPA` represents a significant step towards establishing a physics-informed foundation model for strong gravitational lensing studies.

Acknowledgements

J Rishi was a participant in the Google Summer of Code 2025 program. S.G. was supported in part by U.S. National Science Foundation award No. 2108645. Portions of this work were conducted in MIT's Center for Theoretical Physics and partially supported by the U.S. Department of Energy under grant Contract Number DE-SC0012567. M.W.T. acknowledges financial support from the Simons Foundation (Grant Number 929255).

References

[1] Shude Mao and Peter Schneider. Evidence for substructure in lens galaxies? *Monthly Notices of the Royal Astronomical Society*, 295(3):587–594, 1998.

[2] Yashar D Hezaveh, Neal Dalal, Daniel P Marrone, Yao-Yuan Mao, Warren Morningstar, Di Wen, Roger D Blandford, John E Carlstrom, Christopher D Fassnacht, Gilbert P Holder, et al. Detection of lensing substructure using alma observations of the dusty galaxy sdp. 81. *The Astrophysical Journal*, 823(1):37, 2016.

[3] Simona Vegetti and Léon VE Koopmans. Bayesian strong gravitational-lens modelling on adaptive grids: objective detection of mass substructure in galaxies. *Monthly Notices of the Royal Astronomical Society*, 392(3):945–963, 2009.

[4] LVE Koopmans. Gravitational imaging of cold dark matter substructures. *Monthly Notices of the Royal Astronomical Society*, 363(4):1136–1144, 2005.

[5] Simona Vegetti and LVE Koopmans. Statistics of mass substructure from strong gravitational lensing: quantifying the mass fraction and mass function. *Monthly Notices of the Royal Astronomical Society*, 400(3):1583–1592, 2009.

[6] Stephon Alexander, Sergei Gleyzer, Evan McDonough, Michael W Toomey, and Emanuele Usai. Deep learning the morphology of dark matter substructure. *The Astrophysical Journal*, 893(1):15, 2020.

[7] Ana Diaz Rivero and Cora Dvorkin. Direct detection of dark matter substructure in strong lens images with convolutional neural networks. *Physical Review D*, 101(2):023515, 2020.

[8] Laurence Perreault Levasseur, Yashar D Hezaveh, and Risa H Wechsler. Uncertainties in parameters estimated with neural networks: Application to strong gravitational lensing. *The Astrophysical Journal Letters*, 850(1):L7, 2017.

[9] Johann Brehmer, Siddharth Mishra-Sharma, Joeri Hermans, Gilles Louppe, and Kyle Cranmer. Mining for dark matter substructure: Inferring subhalo population properties from strong lenses with machine learning. *The Astrophysical Journal*, 886(1):49, 2019.

[10] Bryan Ostdiek, Ana Diaz Rivero, and Cora Dvorkin. Image segmentation for analyzing galaxy-galaxy strong lensing systems. *Astronomy & Astrophysics*, 657:L14, 2022.

[11] Stephon Alexander, Sergei Gleyzer, Pranath Reddy, Marcos Tidball, and Michael W Toomey. Domain adaptation for simulation-based dark matter searches using strong gravitational lensing. *arXiv preprint arXiv:2112.12121*, 2021.

[12] Stephon Alexander, Sergei Gleyzer, Hanna Parul, Pranath Reddy, Michael W Toomey, Emanuele Usai, and Ryker Von Klar. Decoding dark matter substructure without supervision. *arXiv preprint arXiv:2008.12731*, 2020.

[13] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[15] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[17] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020.

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[19] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

[20] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[21] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023.

[22] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 2022.

[23] Simon Birrer, Adam Amara, and Alexandre Refregier. Gravitational lens modeling with basis sets. *The Astrophysical Journal*, 813(2):102, 2015.

[24] Stephon Alexander, Sergei Gleyzer, Evan McDonough, Michael W. Toomey, and Emanuele Usai. Deep learning the morphology of dark matter substructure. *The Astrophysical Journal*, 893(1):15, apr 2020.

[25] Ramesh Narayan and Matthias Bartelmann. Lectures on gravitational lensing. *arXiv preprint astro-ph/9606001*, 1996.

[26] Charles R Keeton. A catalog of mass models for gravitational lensing. *arXiv preprint astro-ph/0102341*, 2001.

[27] Robert Kormann, Peter Schneider, and Matthias Bartelmann. Isothermal elliptical gravitational lens models. *Astronomy and Astrophysics (ISSN 0004-6361), vol. 284, no. 1, p. 285-299*, 284:285–299, 1994.

[28] Seung Hoon Lee, Seunghyun Lee, and Byung Cheol Song. Vision transformer for small-size datasets. *arXiv preprint arXiv:2112.13492*, 2021.

[29] Lucas Velôso, Michael Toomey, and Sergei Gleyzer. Lensformer: A physics-informed vision transformer for gravitational lensing. *Journal of neural information processing*, 2023.

[30] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[31] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22–31, 2021.

[32] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 32–42, 2021.

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.