GSoC 2025 Project Proposal for

Foundation Model for Gravitational Lensing

(Duration: 350 hours)

SNEHANSHU MUKHERJEE
INDIAN INSTITUTE OF TECHNOLOGY (ISM), DHANBAD

1. About Me **3**
1.1 Student Information 3
1.2 Educational Background 3
1.3 Relevant Experience 3
1.4 Achievements 4
2. Evaluation Task **4**
3. Motivation for GSoC **6**
4. Why ML4SCI (DeepLense)? **6**
5. Project Information **7**
5.1 Background 7
5.2 Abstract 7
6. Approach/Implementation Details **8**
6.1 Exploration of Domain and Datasets 8
6.2 Development of family of Pre Trained models and Pre Training strategy 8
6.2.1 Exploration of Self-Supervised Methods and Code Preparation 8
6.2.2 Testing Masking Strategies & Loss Functions for Masked Autoencoders 9
6.2.3 Testing of I-JEPA 10
6.2.4 Scaling Experiments 10
6.2.5 Exploration of Context Autoencoder 11
6.2.6 Finalising a Pre-training Strategy 11
6.3 Fine Tuning 12
6.3.1 Classification 12
6.3.2 Super-Resolution 12
6.3.3 Regression (Dark matter property estimation) 13
6.3.4 Good-to-Have Extensions (Optional) 13
6.4 Evaluation & Benchmarking 14
6.5 Documentation 14
7. Time Commitments **14**
8. Proposed Timeline **15**
8.1 Pre-GSoC 15
8.2 Detailed Timeline 15
8.3 End Evaluation 17
9. Deliverables **17**
10. Further Scope **18**
10.1 Multi-Modal Training Pipeline 18
10.2. Sparsification and Efficient Attention 18
10.3. Synthetic Data Generation 18
11. Post GSoC **18**
12. References **19**

1. About Me

1.1 Student Information

Name: Snehanshu Mukherjee
Email: snehanshumukh@gmail.com, 21je0928@iitism.ac.in
Time Zone: Indian Standard Time (+5:30 GMT)
GitHub: pilot-j
LinkedIn: Snehanshu
Resume: Snehanshu_CV

1.2 Educational Background

University: Indian Institute of Technology (ISM), Dhanbad
Major: Electrical Engineering
Minor: Data Science
Current Year: 4th
Expected Graduation: May 2025
Degree: Bachelor of Technology

1.3 Relevant Experience

I am proficient in **Python**, **PyTorch**, and **C++**, with a strong foundation in deep learning and a growing interest in agentic frameworks and GPU computing.

- **Computer Vision Research Intern – CANDLE Research Lab, IIT Roorkee:** Worked on evaluating state-of-the-art object detection architectures and designing a lightweight object detection model. This work contributed to our research paper *ConstructNet.*
- **Data Science Intern – Axis Bank (Business Intelligence Unit):** Helped build the bank's first product recommendation engine. Optimized matrix operations on Spark DataFrames and developed a utility library to streamline matrix operation on internal data.
- **Research Intern (Part-Time) – NeurAI Lab, TU Eindhoven:** Exploring LLM-guided sparsity schemes, focusing on attention mechanisms and MoEfication techniques for LLMs. Work involves writing and testing new attention variants.
- **Open Source Contributions:** Contributed to the *MAYA project* under Aya Expedition 2024. Contributed to improving PyTorch documentation as part of the PyTorch Docathon 2023.
- **Free Time Coding:** Spending time to understand the internals of automatic differentiation libraries and CUDA programming.

1.4 Achievements

- **2nd runner up at Rakathon 2024**, annual AI innovation challenge organized by Rakuten India. Built multi-agent framework for hyper-personalized e-commerce shopping - *Rakumon*
- **10th team out of 3000+ teams at Fibe Hack The Vibe 2024**. Built *Newsense* - BERT-based text classifier for news articles.
- **Contributed to Maya.** Helped to integrate SigLip as vision encoder & wrote the evaluation script to calculate the BLEU score.
- **Presented our work *ConstructNet*** at *2nd International IEEE Applied Sensing Conference (APSCON), 2024*

2. Evaluation Task

This section outlines the evaluation tasks undertaken, the methods employed and the results achieved. It also briefly highlights my observations which act as a precursor to the way I have planned to implement this project.

Related code, approach and details can be found here.

Common Task: Image Classification of Lenses

- A modified ResNet18 architecture with Squeeze-and-Excitation (SE) units was used. The model was trained with the cross-entropy loss function.
- The classification task was relatively straightforward, and the model consistently achieved high AUC values for every class.

Specific Task 6A: Pre-training a Masked Autoencoder (MAE)

- A Masked Autoencoder (MAE) was trained on *no_sub samples* to learn feature representations of strong lensing images. The implementation followed the official MAE framework and was modularized into encoder, decoder, and helper functions. The reconstruction loss was modified from Mean Squared Error (MSE) to BCEWithLogitsLoss. The encoder and decoder were asymmetric to enforce better feature learning.
- The reconstructed images were identifiable but had a blurry white halo.

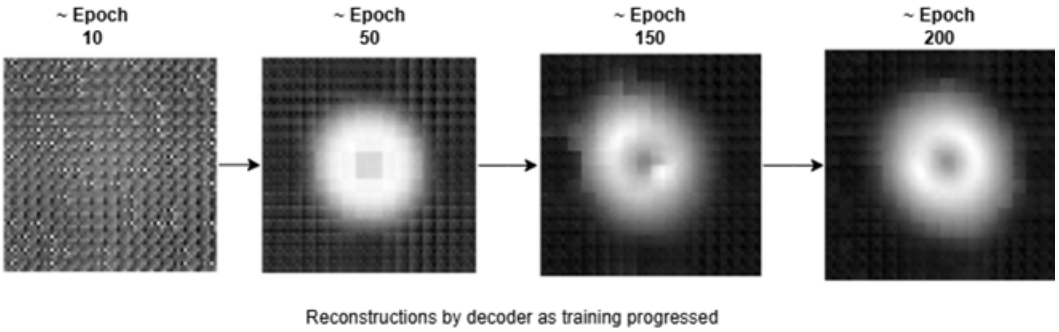

Fig. 1 Reconstruction output of 1st img of validation set

Specific Task 6A: Classification Using Pre-Trained MAE

- The pre-trained MAE was fine-tuned on the full dataset for multi-class classification. Fine Tuning was done in 2 stages - Linear Probing and Full Fine Tuning.
- Average micro AUC of 0.97 was achieved.

Specific Task 6B: Fine-Tuning for Super-Resolution

- The pre-trained model from Task 6A was fine-tuned for super-resolution. The SR head is a CNN, and L1 loss was used for the task. PSNR - 33.6579, SSIM - 0.9681, MSE-0.00042 was achieved.

Observations:

MSE performed poorly compared to BCEWithLogitsLoss.

- Most pixels in the images are black, the model easily learns to predict the background, causing MSE to drop quickly. However, this does not indicate good reconstruction, as the relevant lensing structures (lighter elliptical regions) remain challenging to construct.
- Images were later normalized to have pixel values between [0,1]. The reconstructed images from the validation set visibly produced related patterns when trained with BCE.

Training Instability (Effect of Scheduler and Optimizer)

- Pre-training required numerous epochs and was sensitive to learning rate variations. A manually adjusted StepLR scheduler improved training stability but had minimal effect on final performance.

Random Masking and Loss functions may be suboptimal for our use case

- Dataset images were sparse, with large black regions lacking geometric features. MAE masking and loss did not discriminate between information rich area vs black background.

Input Adaptation

- The ViT encoder was trained on 64x64 images (it expects input of the same dimensions), whereas the low-resolution (LR) inputs were 75x75. Simple resizing sufficed due to redundant background pixels, but resizing can cause information loss. A learnable pointwise convolutional layer was introduced before resizing to mitigate this.

Upsampling Method

- The model relied on convolutional operations for upsampling.Transposed convolutions require manual kernel size tuning for the desired output. The decoder could potentially upsample without explicit convolutional layers, but further exploration is needed.

3. Motivation for GSoC

Joining Google Summer of Code (GSoC) holds a special significance for me—it's much more than just a coding program. It represents the convergence of my passions for machine learning, research, and open-source. As a student, I have deeply benefited from open-source and open-science ecosystems—free science simulators, educational videos, open-access software—all of which helped spark my passion for technology and engineering. These resources made advanced knowledge and tools accessible, even to someone just starting out. Without the generosity of the global open-source community, I might not have discovered my interests so early on.

I view GSoC as a way to give back—a way to contribute to the very ecosystem that shaped me. I want to be part of the global effort to make 'cool' things accessible and meaningful for the next generation of learners, just as others did for me.

Beyond giving back, GSoC also offers an invaluable opportunity to enhance my skills, work on real-world problems, and grow under mentorship. The recognition and experience gained through GSoC would serve as a strong validation of my abilities and boost my confidence in navigating complex challenges. In a fast-evolving tech landscape, I believe that resilience and adaptability are just as crucial as technical prowess—and GSoC is the perfect platform to cultivate both.

4. Why ML4SCI (DeepLense)?

ML4SCI offers this unique opportunity to explore how machine learning can solve science problems, and I am excited to be a part of that mission. As someone who breathes science and engineering, the prospect of contributing to core scientific problems through ML is both refreshing and fulfilling.

I am applying exclusively to **DeepLense** for GSoC 2025 and do not intend to apply to any other organization. My motivation stems from a deep and long-standing fascination with astronomy, particularly since the discovery of gravitational waves. As both a ML practitioner and a physics enthusiast, I see Deeplense as the ideal project to bridge these two interests.

My long-term aspiration is to be involved in applied ML research and contribute to building scalable ML systems for scientific discovery with leading institutions like CERN. I also intend to pursue advanced studies in this direction. Collaborating with organizations like Deeplense places me among like-minded researchers and mentors whose experience and guidance can help me grow—both technically and personally.

Contributing to this project also helps me take meaningful steps toward becoming a better researcher and an advocate of open science. Submitting this proposal is my first concrete step toward that goal.

5. Project Information

5.1 Background

DeepLense is a research initiative at the intersection of astrophysics and machine learning, focusing on the use of advanced ML techniques to study strong gravitational lensing and the nature of dark matter substructures.

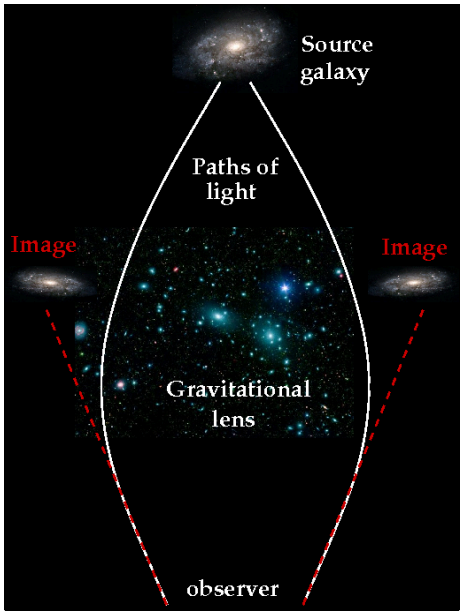

Fig. 2 - Gravitational Lensing

Gravitational lensing occurs when massive celestial objects bend light from background sources, creating distorted or multiple images. This phenomenon offers a powerful observational tool to infer the composition and structure of dark matter.

Work at Deeplense involves analyzing both real and simulated lensing images to identify and characterize different dark matter models. There are three main types: *Axion Dark Matter* with vortex substructures, *Cold Dark Matter* modeled as point-mass subhalos, and a *No-Substructure model* representing smooth dark matter distributions. Previous GSoC projects have contributed in developing ML models for various tasks (Classification, Regression, Super Resolution) helping researchers gain better insights about the role of dark matter.

5.2 Abstract

Organization: ML4SCI
Project Name: Foundation Model for Gravitational Lensing
Mentors: Michael Toomey, Sergei Gleyzer, Pranath Reddy, Anna Parul

Over the years, DeepLense has developed several models to address different tasks in gravitational lensing. This project aims to develop a **foundation model** capable of addressing a wide range of tasks with minimal modifications. The ultimate goal is to build a flexible, **end-to-end pipeline** that can seamlessly adapt to various downstream applications, reducing the need for task-specific redesigns.

The project focuses on two core objectives:

1. **Representation Learning -** Developing a pre-training strategy for learning robust and generalizable representations of gravitational lensing data.

2. **Downstream Fine Tuning -** Fine Tuning the base (backbone) model on a variety of downstream tasks, including classification, regression and super-resolution.

In addition to model development, the project will conduct an exhaustive comparative analysis between the proposed approach and previous work. The final deliverable will also include a

detailed report covering architectural comparisons, task-specific performance, and broader insights into the application of deep learning to strong lensing data.

Project Plan:

- Exploration of domain and datasets
- Development of a family of pretrained models and a pre training strategy
- Fine Tuning on downstream tasks
- Evaluation and benchmarking
- Documentation
- Further scope and future directions

6. Approach/Implementation Details

6.1 Exploration of Domain and Datasets

This project requires a holistic approach as it spans multiple domains and datasets. To support this, I will begin by understanding the fundamental physics of gravitational lensing, focusing particularly on the concepts outlined in DeepLense-related work/papers. I also plan to explore loss function design with an emphasis on incorporating physics-informed variants, which necessitates a solid grasp of the relevant physical equations.

Given the potential use of multiple datasets, an initial statistical analysis is essential. This will include evaluating image pixel distribution, signal-to-noise ratio (SNR), overall data distribution, and the distribution of relevant physical quantities.

Additionally, I intend to study the geometric (among others) properties of the gravitational images. I will review and familiarize myself with the methodologies adopted in PyAutoLens [1] and Lenstronomy [2] to guide this analysis.

6.2 Development of family of Pre Trained models and Pre Training strategy

6.2.1 Exploration of Self-Supervised Methods and Code Preparation

For our purpose we need models that can learn good representations from our data, are generalisable and do not depend on handcrafted augmentations.

My goal is to investigate and compare different self supervised approaches, including:

- **Masked Autoencoders [3] (MAE):** Focused solely on reconstructing masked patches.
- **I-JEPA [4] (Image- Based Joint Embedding Predictive Architecture):** Learns representations by predicting features in masked regions. Non-generative and the predictions are made in representation space.
- **Context Autoencoders [5] (CAE) :** Combines reconstruction with embedding alignment between visible and masked patches.

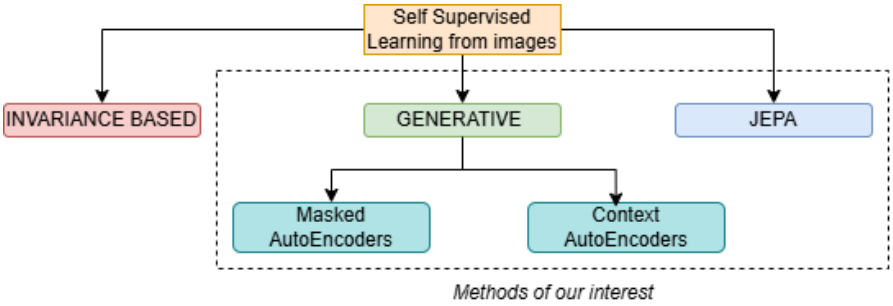

Fig. 3 - SSL Methods

I will begin by thoroughly analyzing the theoretical foundations of these SSL methods, emphasizing their core motivations and architectural distinctions. More development time will be allocated to implementing I-JEPA (and CAE), while MAE will be used for initial experiments to establish performance baselines. The code for MAE is nearly complete as part of the evaluation task.

6.2.2 Testing Masking Strategies & Loss Functions for Masked Autoencoders

To establish a strong baseline, I will focus on Masked Autoencoders while experimenting with various masking strategies. During the evaluation task, I observed that the model tends to quickly learn to produce a black background, likely due to the dominance of black pixels. Random masking fails to encourage learning of the lensing structures. To address this, I plan to implement region-aware masking (AutoMAE **[6]**), which prioritizes masking areas with higher information density. This strategy is straightforward to implement and avoids the need for architectural changes.

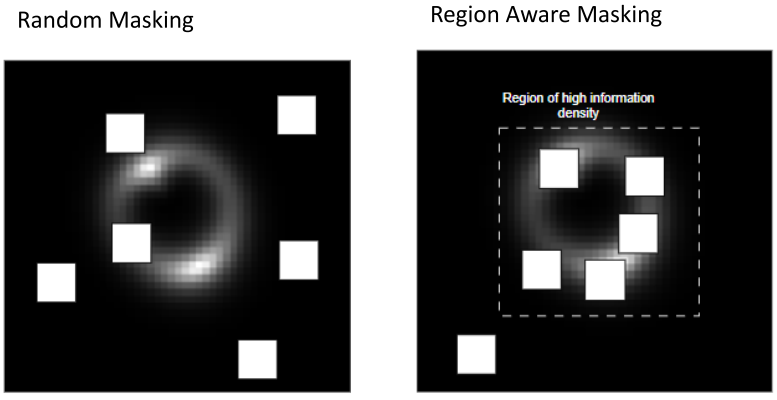

Fig. 4 Random vs Region Aware masking

Given the sparsity of the dataset, pixel-based losses tend to converge to zero without offering meaningful interpretability. Moreover, a lower pixel-based loss does not necessarily translate to better representations—this was evident during the evaluation task, where lower MSE did not yield qualitatively better reconstructions.

Since our use case focuses on ring-like structures, I aim to explore the integration of geometry-based (perceptual loss **[7]**) or physics-informed loss terms, which could enforce relevant priors during pre-training. This may also enhance performance in downstream tasks like super-resolution and regression.

We will employ a custom CNN-based feature extractor to assess the perceptual similarity between reconstructed and input images, complementing metrics such as PSNR and SSIM to provide a more comprehensive evaluation of reconstruction quality. Model performance will be benchmarked using both reconstruction metrics and linear probing scores.

6.2.3 Testing of I-JEPA

I-JEPA has demonstrated notable advantages over MAEs in two key areas: improved representation learning and faster convergence. A consistent issue I observed with MAEs is training instability—these models often require significantly more epochs to converge. I-JEPA addresses this by enforcing alignment in the representation space, leading to more stable and efficient training.

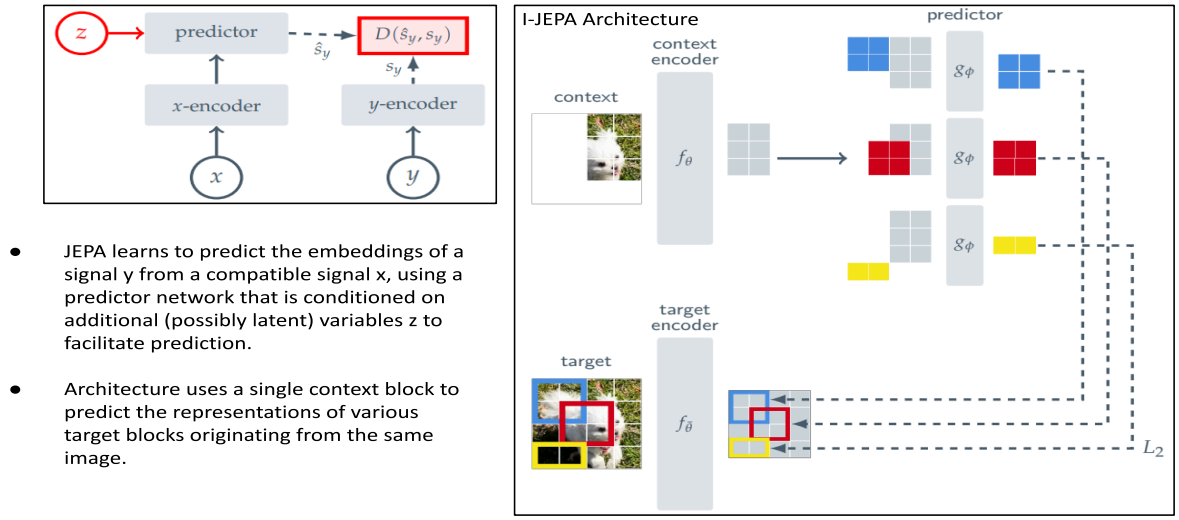

Fig. 5 I-JEPA Overview

Since the I-JEPA architecture has not yet been thoroughly explored in previous DeepLense projects, this presents an opportunity to establish concrete baselines and assess its potential in our domain. Alongside code preparation for a similar ViT base, I'll train I-JEPA and evaluate its performance.

6.2.4 Scaling Experiments

Once a common baseline is established, I plan to explore how the architecture performs across different scales and hyperparameters. This includes varying the masking ratio, patch size, embedding dimensions, and other hyperparameters. The goal is to gather concrete insights into model scalability within the first six weeks of work.

Reconstruction quality and linear probing scores will serve as the primary evaluation metrics. Additionally, I will visualize the learned representations using t-SNE, UMAP and attention maps plots for each model variant.

6.2.5 Exploration of Context Autoencoder

Context Autoencoders (CAEs) improve upon MAEs by not only reconstructing masked patches but also aligning the representations of visible and masked regions—similar in spirit to I-JEPA. Comparing CAEs with MAEs and I-JEPA will help us better understand the latent space associated with gravitational lensing data.

I aim to implement the architecture and tabulate a base results for CAEs. However, extensive testing and evaluation will depend on time availability. My primary interest lies in studying the loss function, as CAEs introduce a clever representation alignment term.

Loss in CAE

$$\ell_y(\mathbf{Y}_m, \bar{\mathbf{Y}}_m) + \lambda\, \ell_z(\mathbf{Z}_m, \mathrm{sg}[\bar{\mathbf{Z}}_m]).$$

$\bar{\mathbf{Y}}_m$: Target for patch reconstruction

$\bar{\mathbf{Z}}_m$: Target for representation of masked patches

(a) feeds visible patches into the encoder and extracts their representations
(b) completes the pretext tasks: predict the representations of the masked patches from the visible patches in the encoded representation space

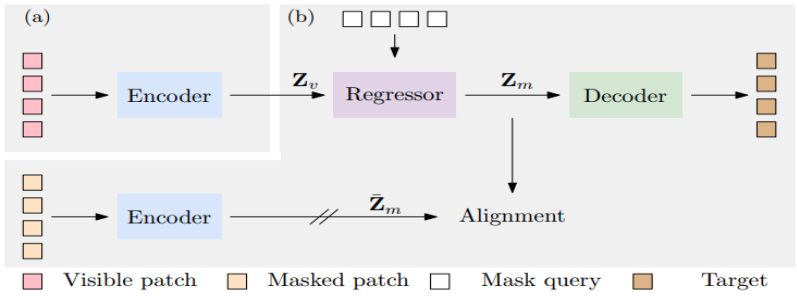

Fig. 6 Context AutoEncoder Overview

This addresses the limitation observed during MAE training, where improved reconstruction did not necessarily lead to better latent representations. Introducing alignment in MAEs could potentially mitigate this issue. This exploration is primarily for research purposes and may not be included in the final pipeline.

6.2.6 Finalising a Pre-training Strategy

By the end of the pre training phase, we will have tested multiple architectures and established a set of common baselines. At this stage, key hyperparameters and a core strategy for each model family will be finalized based on empirical results and performance benchmarks. All relevant details necessary for reproducing our results will be compiled and consolidated into a cohesive technical document.

Deliverable: Final code versions, pre-training scripts, configuration files, and a summary of insights gained during experiments. In addition, I will prepare pre-training guidelines to help other researchers adopt and adapt our methods in their own work.

This deliverable will serve as a technical reference for the mid-term evaluation, alongside the source code and experimental reports.

6.3 Fine Tuning

The next step is to evaluate effectiveness of pretrained backbone models on downstream tasks — **classification**, **super-resolution**, and **regression**. We will start by assessing the datasets available for each of these tasks to determine suitability and any required preprocessing.

My approach will involve reproducing results from existing DeepLense work (where available) and systematically comparing the performance of our fine tuned models. For each task, we aim to define a robust finetuning strategy, tailored to its specific requirements. We will also investigate the impact of input noise on model performance to assess robustness under realistic conditions. Throughout the process, we will also identify and document the limitations of our approach, offering insights for future improvements.

Deliverable: Task-specific data loaders, utility functions, and a complete end-to-end pipeline for each downstream task. Additionally, we will provide fine tuning scripts for task-specific heads built on top of the pretrained backbones.

6.3.1 Classification

Since we will be evaluating models using linear probing during pre training, we will already have preliminary insights into their classification capabilities. The goal here is to compare linear probing performance with full finetuning on selected models. We will use ROC-AUC as the primary evaluation metric, and analyze the effect of data augmentation and noise on classification performance. If time allows, we will also explore using a transformer-based classification head as an alternative to the standard MLP.

I also plan to assess how a backbone fine tuned on classification affects performance on subsequent tasks like regression and super-resolution. This is motivated by observations that fine tuned classification models often show a well-separated CLS token distribution in t-SNE visualizations—implying the model has learned the concept of discrete classes. Whether this prior bias is helpful or detrimental for other tasks remains an open question, which will be explored.

6.3.2 Super-Resolution

The goal of the super-resolution task is to enhance the spatial resolution of gravitational lensing images while preserving physically meaningful features. We will begin by exploring CNN-based super-resolution heads, which can be efficiently integrated with our pretrained backbones. This approach was previously used in the evaluation task, where I manually calculated the appropriate kernel size for transpose convolutions.

Next, I will implement GAN-based training [8] [9] [10] (adapted for our pre-trained ViT based components), which may improve output quality by better recovering fine lensing structures. To further understand how different methods scale, models will be tested across multiple upsampling factors, evaluating their effectiveness in different resolution scenarios. The final deliverable will be a generalized pipeline that can be adapted for varying scale factors (2x, 4x).
If a suitable dataset becomes available, we may explore a physics-informed formulation of the super-resolution problem—drawing inspiration from related research (LensPINN [11] [12]).
Another optional direction will be to leverage the native decoder of the pretrained architecture for upsampling, eliminating the need for a dedicated super-resolution head.

6.3.3 Regression (Dark matter property estimation)

For the regression task, the first step will be to reproduce existing results to establish a strong and consistent baseline. I will also review the regression tasks and physical quantities of interest in gravitational lensing, such as mass density of vortex substructure.

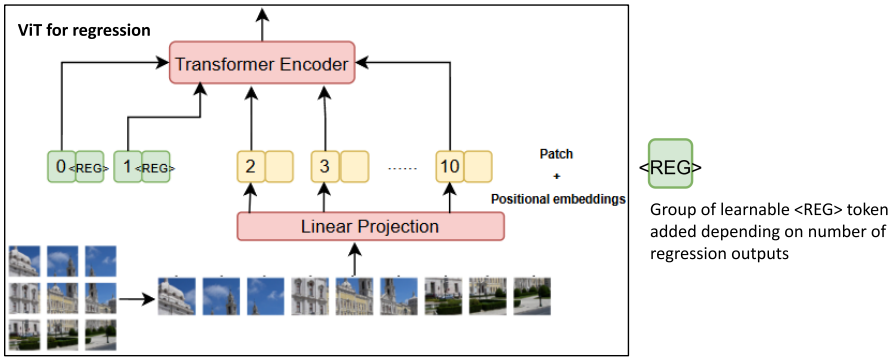

Fig. 7 Proposed change in ViT

Given the emerging use of transformers in regression, I will study their applicability, with a particular focus on incorporating a dedicated REG token—analogous to the CLS token in classification tasks—to represent the global regression output. This could offer a cleaner way for the model to aggregate information relevant to continuous-valued predictions.

Notably, previous work has shown that Vision Transformers (ViTs) [13] underperform compared to CNNs [14] in this context. The goal is to understand whether carefully adapted and pretrained ViT-based models can match or outperform previous approaches in predicting physically meaningful lensing parameters.

6.3.4 Good-to-Have Extensions (Optional)

In alignment with mentor guidance and organizational priorities, I would like to explore additional downstream tasks beyond the core objectives. Two areas of particular interest are lens finding and anomaly detection.

Lens finding resembles a long-tail object detection problem, where the number of positive (lensed) samples is significantly outnumbered by negatives. This task is crucial in wide-field surveys like HSC-SSP. The idea is to integrate pre-trained backbone with an object detection-like pipeline that models elliptical bounding boxes instead of standard rectangular ones. We will evaluate how well pretrained models generalize to lens finding in real observational data, and to assess their limitations.

Another direction I would like to explore is anomaly detection, particularly in identifying rare or unexpected features in astrophysical datasets. SSL models are well-suited for learning general representations, and with appropriate fine-tuning, may be leveraged to spot out-of-distribution signals or deviations from known physical patterns. This could provide valuable insights for future astronomical surveys where manual inspection is infeasible due to scale.

However, my current exposure to this domain is limited, and meaningful progress will require thorough reading and familiarization with the existing literature, datasets, and previous/ongoing work done within DeepLense.

6.4 Evaluation & Benchmarking

Evaluation and benchmarking will be an ongoing process, conducted at key stages—i.e after each sub-phase—to systematically track performance improvements. The process will begin with a comprehensive listing of datasets of interest, determined in collaboration with mentors. The goal is to compile and analyze results across all performance metrics using both provided datasets and other relevant astrophysical datasets.

Deliverable: Compiled Results. Generalized evaluation scripts that can be reused across the organization, facilitating consistent and reproducible benchmarking for future projects.

6.5 Documentation

Since this project aims to develop a **family of models** capable of performing multiple tasks and adapting to custom datasets, comprehensive documentation is crucial. Inspired by the *YOLO* series maintained by Ultralytics, I envision extending this work into a dedicated library, making it easily accessible for researchers across various use cases.

To ensure clarity and usability, documentation will be integrated at multiple levels:

- **Code-Level Documentation:** All functions will include detailed docstrings outlining their functionality, input parameters, and return values. Key implementation details will also be annotated where necessary.
- **Workflow & Approach:** I will document my methodology through blogs, covering key design choices, challenges, and solutions.
- **Starter & Contributor Guides:** Step-by-step guides will help users fine-tune the models for their specific tasks, while contributor guides will provide clear instructions for extending or improving the library.
- **Final Documentation Phase:** Towards the end of the project, I will structure the documentation into library-style guides, complete with starter notebooks and reference materials, ensuring seamless onboarding for new users - provided sufficient time is available for this transition.

7. Time Commitments

I will be able to dedicate approximately 30 to 35 hours per week to the project — around 20 hours on weekends, with the remaining time distributed across weekdays Starting June, I will be working full-time at *Meesho*, but I have already discussed the expected workload with my team there and can confidently manage both commitments without any overlap. To avoid any hiccups during the project timeline, I plan to start preparing relevant code and boilerplate scripts in May itself. This will ensure that the majority of the GSoC period is focused on critical components and meaningful experimentation

Weekly updates and meeting schedules can be coordinated based on the availability of mentors to ensure smooth progress and regular communication.

8. Proposed Timeline

8.1 Pre-GSoC

Before the official coding period begins, I plan to:

- **Familiarize myself with the Deeplense codebase** and review previous works relevant to this project to understand existing approaches and architectural decisions. Engage actively with mentors to understand Deeplense's coding practices and collaborative norms.
- **Study PyAutoLens (and Lenstronomy)** especially its data generation pipeline and the underlying physics of gravitational lensing, to build a stronger foundation for informed model design.
- **Deep dive into self-supervised learning (SSL)** in vision
- Spend some time on **transformer-based regression techniques**, exploring both theoretical concepts and practical implementations to prepare for downstream task adaptation.

8.2 Detailed Timeline

Phase	Period	Task
Community Bonding Period		
Exploration of domain and datasets.	Week 1 [June 2 - June 8]	• Compile a list of relevant datasets and surveys. • Perform statistical analysis on datasets to generate insights. • Read research papers on self supervised learning models of interest and document their motivations and methods. • Add REG token in transformer architecture. Perform a preliminary test to make sure adding REG token does not negatively affect pre-training performance - use evaluation task as a reference.
Code preparation and Experiments	Week 2 Week 3 [June 9 - June 22]	• Test MAE with different masking and loss functions. • Perform scaling experiments. • Prepare code for I-JEPA and CAE.
	Week 4 Week 5 [June 23 - July 6]	• Test I - JEPA. Conduct similar scaling experiments. • Test CAE. • Tabulate results for all coded architecture. • Start writing a blog on work done up till now.

Transition Phase	Week 6 [July 7 - July 12]	• Document pre training strategy. • Complete pre training scripts and related notebooks. • Start classification finetuning. Perform full finetuning.
Mid Term Evaluations		
Classification	Week 7 [July 14 - July 20]	• Visualise attention heads, t-SNE and UMAPs • Test transformer based heads vs MLPs. • Tabulate results and organise related code into scripts and notebooks
Super resolution	Week 8 Week 9 [July 21 - Aug 3]	• Test CNN based SR head with pre trained backbone • Develop GAN based training method and compare it with CNN approach • Explore scope of integrating special loss functions to above tested methods (if specific dataset is available) • Test classification backbone and compare performance • Tabulate results
Deep Regression	Week 10 Week 11 [Aug 4 - Aug 17]	• Read about deep regression in gravitational lensing. • Reproduce previous work based on CNNs and Hybrid architecture. • Test pre trained backbone with plain regression head. • Test fine tuned classification backbone.
Documentation	Week 12 [Aug 18 - Aug 24]	• Compile evaluation and benchmarking results for fine-tuning • Prepare an end to end pipeline for pre-training, finetuning and inference. • Finalise code and repository. • Hardware optimization in model training & inference code. *(Optional)* • Add contributor guides/ starter notebooks *(Optional)* • Write unit tests *(Optional)* • Write blog
Wrapping Up	Week 13 [Aug 25 - Sept 1]	• Start model migration on hugging face (if allowed) • Finalise technical report. • Document scope of further improvement and alternate directions of research. • Buffer time for any blockers.
End Term Evaluations		
Post GSoC		

8.3 End Evaluation

I have carefully structured the project timeline, taking into account other commitments, and I am confident that all milestones will be achieved on time. By the final evaluation phase, I will ensure that all implementations are well-documented, tested, and integrated into the overall pipeline. Additionally, I will actively explore opportunities to publish a research paper based on the outcomes of this project.

9. Deliverables

- **Github Repo** containing implementation of pre-trained and fine-tuned models for classification, super-resolution, and regression.
- **Technical report** detailing the methodology, experiments and results.
- **Published models on Hugging Face** (subject to licensing and project policies).
- **Well-structured documentation** including contributor guides and starter notebooks.
- **End-to-End pipeline** for easy reproduction, evaluation, and adaptation of the models to new datasets or tasks.
- **GPU optimization insights or improvements** for accelerating training/inference, if large-scale finetuning is involved *(Optional)*

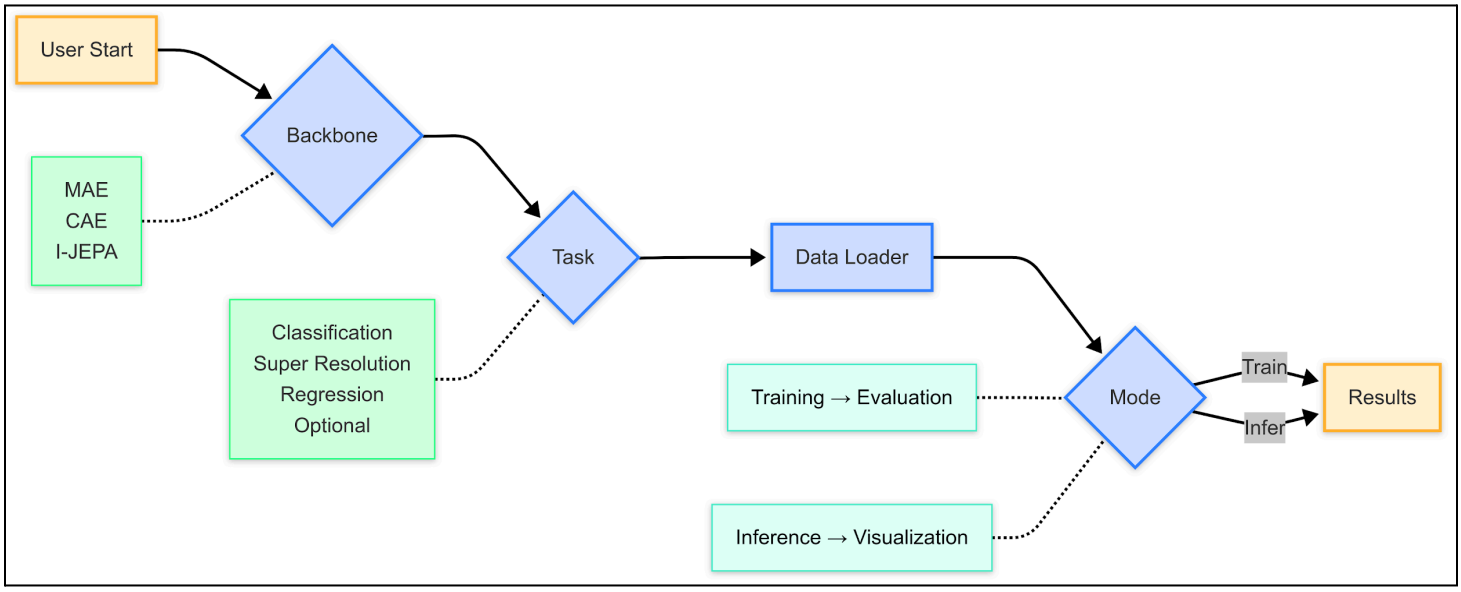

Fig. 8 Foundation Model E2E Pipeline

10. Further Scope

The following list outlines potential research directions and future extensions of this project.

Over the past month, I've explored a wide range of literature and identified the following gaps and opportunities. If time permits, I intend to initiate some of these explorations during GSoC and remain engaged with them beyond the program.

10.1 Multi-Modal Training Pipeline

Recent work **[15]** in self-supervised learning suggests the potential for multi-modal training, integrating complementary sensory outputs to enrich representations. Given the sparse nature of image data and the limited availability of labeled samples, it's worth exploring whether heatmaps, contour maps, or other physically meaningful data modalities can be integrated alongside traditional imaging datasets. This could lead to more robust pretraining and improved generalization across tasks.

10.2. Sparsification and Efficient Attention

Vision Transformers (ViTs) are often parameter-heavy, and sparsification has been actively explored in both NLP and computer vision to improve efficiency. In language modeling, newer attention variants **[16] [17]** have reduced computational overhead without sacrificing performance. CNNs have also benefited from structured sparsity schemes. Given the scale of models we work with, experimenting with lightweight attention mechanisms or sparsity-inducing techniques may result in significant gains. This direction is inspired by my prior research work at TU Eindhoven, and I'm particularly interested in assessing whether vision models can achieve similar efficiency gains as LLMs.

10.3. Synthetic Data Generation

Another promising direction is to evaluate whether our pretrained backbone can aid in synthetic data generation. A natural starting point is to use the super-resolution (SR) network to upscale low-resolution images, which could help augment existing datasets. We can explore whether the backbone can be adapted for diffusion-based generation, potentially allowing us to simulate lensing images under different physical conditions or data regimes.

11. Post GSoC

After the conclusion of GSoC, I plan to continue exploring additional research directions (as mentioned above) and improvements to the work done during the summer. I am genuinely interested in remaining involved with the DeepLense project and contributing to its open-source community. I also hope to take on a more active role in the community—potentially contributing as a mentor in future GSoC seasons.

12. References

1. Nightingale, James, Richard G. Hayes, Ashley Kelly, Aristeidis Amvrosiadis, Amy Etherington, Qiuhan He, Nan Li et al. "PyAutoLens: open-source strong gravitational lensing." *arXiv preprint arXiv:2106.01384* (2021).
2. Birrer, Simon, and Adam Amara. "lenstronomy: Multi-purpose gravitational lens modelling software package." *Physics of the Dark Universe* 22 (2018): 189-201.
3. He, Kaiming, 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*, pp. 16000-16009. 2022.
4. Assran, Mahmoud, 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*, pp. 15619-15629. 2023.
5. Chen, Xiaokang, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. "Context autoencoder for self-supervised representation learning." *International Journal of Computer Vision* 132, no. 1 (2024): 208-223.
6. Chen, Haijian, Wendong Zhang, Yunbo Wang, and Xiaokang Yang. "Improving masked autoencoders by learning where to mask." In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 377-390. Singapore: Springer Nature Singapore, 2023.
7. Johnson, Justin, Alexandre Alahi, and Li Fei-Fei. "Perceptual losses for real-time style transfer and super-resolution." In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 694-711. Springer International Publishing, 2016.
8. Ledig, Christian, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken et al. "Photo-realistic single image super-resolution using a generative adversarial network." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681-4690. 2017.
9. Lee, Kwonjoon, Huiwen Chang, Lu Jiang, Han Zhang, Zhuowen Tu, and Ce Liu. "Vitgan: Training gans with vision transformers." *arXiv preprint arXiv:2107.04589* (2021).
10. Baghel, Neeraj, Shiv Ram Dubey, and Satish Kumar Singh. "SRTransGAN: Image Super-Resolution using Transformer based Generative Adversarial Network." *arXiv preprint arXiv:2312.01999* (2023).
11. Ojha, Ashutosh, Sergei Gleyzer, Michael W. Toomey, and Pranath Reddy Kumbam. "LensPINN: Physics Informed Neural Network for Learning Dark Matter Morphology in Lensing."
12. Shankar, Anirudh, Michael W. Toomey, and Sergei Gleyzer. "Unsupervised Physics-Informed Super-Resolution of Strong Lensing Images for Sparse Datasets."
13. Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).
14. 'Deep Regression Exploration' https://medium.com/@gg884691896/gsoc-2021-with-ml4sci-deep-regression-exploration-34d5d8fb4643
15. Bachmann, Roman, David Mizrahi, Andrei Atanov, and Amir Zamir. "Multimae: Multi-modal multi-task masked autoencoders." In *European Conference on Computer Vision*, pp. 348-367. Cham: Springer Nature Switzerland, 2022.
16. Ainslie, Joshua, James Lee-Thorp, Michiel De Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. "Gqa: Training generalized multi-query transformer models from multi-head checkpoints." *arXiv preprint arXiv:2305.13245* (2023).
17. Anagnostidis, Sotiris, Dario Pavllo, Luca Biggio, Lorenzo Noci, Aurelien Lucchi, and Thomas Hofmann. "Dynamic context pruning for efficient and interpretable autoregressive transformers." *Advances in Neural Information Processing Systems* 36 (2023): 65202-65223.