
Vision Transformers for Cosmological Fields: Application to Weak Lensing κ Maps

Jash Kakadia

Department of Physics and Astronomy
University of Pennsylvania
Philadelphia, PA 19104
jkakadia@seas.upenn.edu

Shubh Agrawal

Department of Physics and Astronomy
University of Pennsylvania
Philadelphia, PA 19104
shubh@sas.upenn.edu

Kunhao Zhong

Department of Physics and Astronomy
University of Pennsylvania
Philadelphia, PA 19104
kunhaoz@sas.upenn.edu

Bhuvnesh Jain

Department of Physics and Astronomy
University of Pennsylvania
Philadelphia, PA 19104
bjain@physics.upenn.edu

Abstract

Weak gravitational lensing is a powerful probe of the universe’s growth history. While traditional two-point statistics capture only the Gaussian features of the convergence field, deep learning methods such as convolutional neural networks (CNNs) have shown promise in extracting non-Gaussian information from small-scale, nonlinear structures. In this work, we evaluate the effectiveness of attention-based architectures, including variants of vision transformers (ViTs) and shifted window (Swin) transformers, in constraining the cosmological parameters Ω_m and S_8 from weak lensing mass maps. Using a simulation-based inference (SBI) framework, we compare transformer-based methods to CNNs. We also examine performance scaling with the number of available N -body simulations, highlighting the importance of pre-training for transformer architectures. We find that the Swin transformer performs better than vanilla ViTs, especially with limited training data. Despite their higher representational capacity, the Figure of Merit for cosmology achieved by transformers is comparable to that of CNNs under realistic noise conditions.

1 Introduction

Spatial fluctuations in matter density encode information about the cosmological composition and evolution of the universe. Weak gravitational lensing is a key probe of structure formation, as it does not rely on modeling biased tracers (galaxies). Stage III experiments, such as the Dark Energy Survey [1, 2], Hyper Suprime-Cam Survey [3], and Kilo-Degree Survey [4, 5], have placed leading constraints on the matter density Ω_m and fluctuation amplitude σ_8 . However, the persistent tension between early- and late-time σ_8 measurements suggests possible new physics beyond Λ CDM model. This motivates the development of alternative statistics that go beyond the traditional two-point function, which will be even more important for upcoming surveys such as Euclid [6], LSST [7], and the Roman Space Telescope [8].

On large scales, weak lensing maps are nearly Gaussian and well described by two-point statistics [4, 9]. On small scales, however, nonlinear gravitational evolution induces non-Gaussian features, as overdensities collapse into halos and galaxies [10]. A variety of statistics have been developed to capture this information, including higher-order correlations [11, 12], peak counts [13, 14],

Minkowski functionals [15], scattering transforms [16], and deep learning approaches [17–20]. These have been applied to both simulations and real data [21–24].

Convolutional neural networks (CNNs) have been widely used in weak lensing to extract non-Gaussian information by mapping convergence fields to cosmological parameters. In this work, we explore attention-based models (Transformers). Transformers were originally developed for natural language processing [25] and later adapted for images [26, 27]. They process inputs in parallel using self-attention and have now found astronomical applications in galaxy morphology classification [28], strong gravitational lens parameter estimation [29], target detection in multi-band photometric surveys [30], and cosmology from 3D dark matter halo lightcones [31].

ViTs are especially attractive for weak lensing applications. They capture global context directly through attention, without requiring a hierarchical stack of local filters as in CNNs. This enables more efficient learning of long-range dependencies compared to CNNs, which learn such dependencies only in deeper layers. ViTs also have weaker inductive biases, which allows for more flexible representations but increases the need for training data and careful regularization. A key advantage is interpretability: attention weights can be analyzed using tools such as attention rollout, attention flow [32], and relevance propagation [33].

We present the first application of Vision Transformers—and more generally, self-attention-based deep learning—for predicting cosmological parameters from simulated weak lensing maps. Alongside three configurations of the original ViT [26], we study shifted window transformers (Swin Transformers) [34]. These use local self-attention within hierarchical windows, offering improved efficiency and better scaling to high-resolution inputs.

Our manuscript is organized as follows. In Section 2, we introduce simulation-based inference, how vision models can be used to constrain cosmology from the field level, the simulations we use, and how we train the models. In Section 3, we review the key differences of attention-based models. We present the results and conclude in Section 4.

2 Methodology

Simulation-based inference (SBI), also known as likelihood-free inference (LFI), enables Bayesian parameter inference directly from simulated observables. It is especially useful when the likelihood is intractable, but forward simulations are feasible, like the case of higher-order statistics. Neural posterior estimation (NPE), or density-estimation LFI (DELFI), adopted this approach by training neural density estimators to learn the relationship between parameters and data. These methods can achieve posterior inference using orders of magnitude fewer simulations than traditional LFI approaches [35–38]. DELFI has been applied to supernova analysis [39], high-redshift Ly α forest modeling, tomographic weak lensing [40], and CMB parameter inference [41]. For an overview, see [42].

We frame our inference as a conditional density estimation task. Given simulation parameters θ and compressed data vectors d , we train an ensemble of neural density estimators (NDEs) to approximate $p(d|\theta)$, by auto-regressively reducing the Kullback–Leibler distance. Assuming a flat prior matching the simulation distribution, this yields posteriors from deep learning models. Our implementation uses the public `sbi` package. The ensemble consists of three Masked Autoregressive Flows (MAFs) and three Gaussian Mixture Density Networks (MDNs), all trained separately. The MAFs have 2, 3, and 4 MADE layers with 50, 50, and 25 transforms per layer, respectively. The MDNs have 2, 3, and 4 Gaussian components, using 50, 50, and 25 neurons per layer.

In the limit of abundant simulations, all NDEs should yield consistent posteriors. To reduce outlier effects, we impose a selection criterion: we discard any NDE whose 1σ constraint on S_8 deviates by more than 5% from the ensemble median. The remaining NDEs’ samples are concatenated to form the final posterior. We validate the inference using the empirical coverage test [43–45], implemented via the public TARP package. This test assesses whether the inferred posteriors are well-calibrated, i.e. whether they are too conservative or too narrow—by comparing against ground-truth parameter values across repeated trials.

We use weak lensing convergence maps generated from the DarkGridV1 N -body simulation suite [46, 47]. The convergence maps are constructed as weighted projections of 3D overdensity fields, following procedures similar to [48, 49]. We adopt the DES-Y3 redshift distribution for four

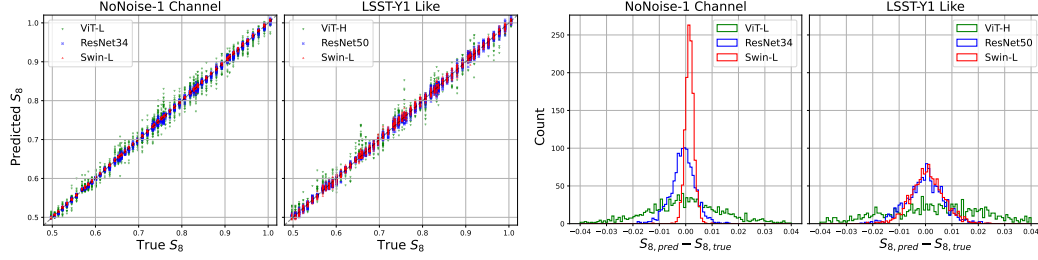


Figure 1: The vision model prediction of the parameter $S_8 = \sigma_8 \sqrt{\Omega_m}/0.3$. The first two plots show the prediction vs the true value, and the last two plots show the histogram of the residuals. The statistical measures are reported as root mean-square error and R^2 values, and we summarize them in Appendix B. Note that *NoNoise* is an exploratory case with only one channel, and *LSST-Y1 Like* contains the maps with 4 channels, as detailed in Section 2.

tomographic bins, but rescale the galaxy number density and survey area to match LSST Year 1. While real LSST-Y1 data will have different galaxy selections and lensing kernels, our conclusions are not sensitive to this, as the dominant factor remains the noise level. The simulations vary only Ω_m and σ_8 , while holding other cosmological parameters fixed and omitting systematics such as photometric redshift uncertainty. These simplifications are acceptable for our purpose of comparing different architectures. For testing purposes only, we also made a shape-noise free simulation set with only one channel to test the model difference with maximum signal-to-noise ratio.

For each configuration, we use an 80%:10%:10% for our training, validation, and testing split. In total there are 13680 maps in the dataset, and each input is a 512×512 independent flattened patch of a convergence map, with either one or four channels. We train each network independently using the AdamW optimizer [50], with an early stopping patience of 30 epochs. A L2 error metric (root-mean-squared, RMSE) is used, computed over the validation dataset. Batch size is 32, except for the much larger networks ViT-L and ViT-H (each $> 300M$ parameters), where we set to 24 (16) and 12 (8) respectively for the one-channel (*four-channel*) dataset, due to GPU memory constraints (we use Nvidia A100s with 80 GBs of memory). We use the PyTorch ReduceLROnPlateau scheduler with a 10-epoch patience and reduction rate of 0.3. For all convolutional networks and vision transformers, the initial learning rate was set to 10^{-3} and 10^{-5} respectively. Other values were set to PyTorch defaults.

Vision Transformers are known to require large training sets [26]. To address this, we explore pre-training on synthetic data to improve performance. We generate pre-training data following [51], which uses an analytic model to rapidly produce maps with similar statistics to weak lensing fields. Our synthetic dataset includes 13680 maps, on which we pre-train the models in a fashion identical to the above. We then use the same method to train the pre-trained models on the real dataset.

3 Attention-Based Vision Models

Transformers, first introduced in [25], have achieved remarkable success in natural language processing (NLP). They process sequential inputs—such as words in a sentence—using self-attention, which models contextual relationships between tokens. Each token is embedded and transformed via three learned matrices: the query, key, and value. Together, these form an attention head, which enables the model to capture pairwise dependencies across the input sequence. Similarly to how CNNs can learn different types of features using multiple filters (i.e., learned weight kernels), multiple learnable attention heads can capture different relations between the input tokens. Positional encodings are added to preserve order information, which is otherwise absent in the architecture.

To adapt transformers to computer vision, images are divided into fixed-size patches. Each patch is flattened and linearly projected into an embedding space. These patch embeddings, along with positional encodings, are passed through a stack of transformer encoder blocks. Each block consists of multi-head self-attention and feedforward (MLP) layers, along with residual connections and normalization layers. For full architectural details, we refer the reader to [25, 26].

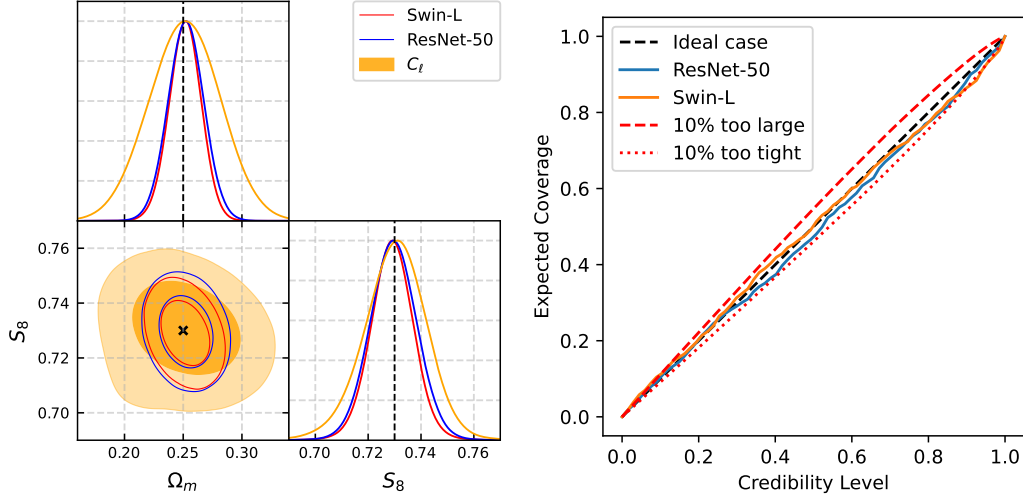


Figure 2: *Left*: SBI posteriors obtained from our NDE setup detailed in Section 2, for a CNN (blue) and a Swin-Transformer (red) for cosmological weak lensing κ fields. *Right*: Results from TARP [52], which compares empirical coverage against credibility level to gauge posterior performance. Posteriors obtained on the test set from our NDE setup yield curves closely aligned with the identity line, indicating that neither of the posteriors is under- or over-constrained.

The Swin Transformer [53] modifies the ViT structure to improve efficiency and scalability. Instead of applying global self-attention across all patches, Swin computes attention locally within non-overlapping windows. These windows shift between layers, allowing cross-window interactions while preserving locality. This design reduces the computational complexity of self-attention from quadratic to linear in image size, enabling Swin to scale to high-resolution inputs with fewer parameters than ViT.

While more efficient than ViT, Swin models still have significantly more parameters than typical convolutional architectures. They can also exhibit training instabilities when scaled up. SwinV2 [54] addresses this issue, introducing improved normalization and scaling techniques for stable training at larger model sizes.

Both ViT and Swin models have much higher parameter counts than standard CNNs. For instance, the smallest ViT model used here has approximately 86 million parameters, while ResNet-50 has about 24 million. This makes architecture tuning impractical in our setting, where simulation and computational resources are limited. Instead, we adopt standard configurations: three ViTs (ViT-B, ViT-L, ViT-H) from [26], and four Swin models (Swin-T, Swin-S, Swin-B, Swin-L) from [53].

4 Results and Conclusion

As shown in Figure 1 and Figure 2, the best attention-based model and the best convolution-based model achieve comparable constraining power. The full results are shown in Appendix B. Both of the models passed the coverage test that they are within 10% of over-constraining. In the exploratory noiseless case, Swin Transformers initially appeared to outperform CNNs, likely due to their greater model flexibility. However, this advantage vanished once realistic shape noise was introduced. Vanilla Vision Transformers consistently underperformed relative to both CNNs and Swin models, possibly due to inefficient training on small datasets and difficulty capturing fine-scale features.

We also find that pre-training significantly improves the performance of large models such as Swin Transformers, especially when training data are limited (Figure 3). In contrast, pre-training has minimal effect on CNNs, consistent with previous findings [55].

In conclusion, our experiments show that attention-based transformers do not outperform CNNs in constraining cosmological parameters from weak lensing convergence maps. With sufficient training data or pre-training, their performance can match that of CNNs. Although no gain in constraining

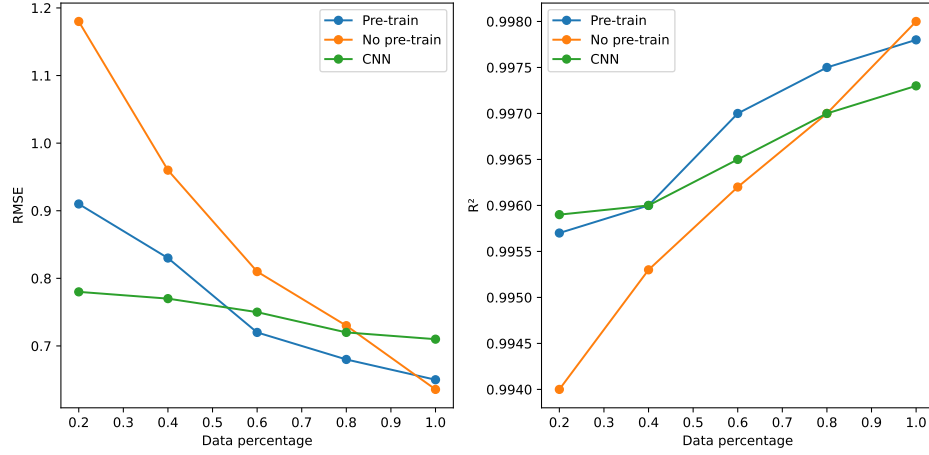


Figure 3: RMSE and R^2 for S_8 as a function of the training data fraction. Pre-training significantly improves the performance of attention-based models, especially in low-data regimes. In contrast, CNN baselines—shown without pre-training—exhibit stronger regularization and more stable performance.

power is observed, the interpretability features of transformer architectures remain a promising direction for future exploration.

References

- [1] Amon A., et al., 2022, *Physical Review D*, 105, 023514
- [2] Secco L. F., et al., 2022, *Physical Review D*, 105, 023515
- [3] Miyatake H., et al., 2023 ([arXiv:2304.00704](#))
- [4] Hildebrandt H., et al., 2017, *Monthly Notices of the Royal Astronomical Society*, 465, 1454
- [5] Heymans C., et al., 2021, *Astronomy & Astrophysics*, 646, A140
- [6] Laureijs R., et al., 2011 ([arXiv:1110.3193](#))
- [7] Ivezić Ž., et al., 2019, *ApJ*, 873, 111
- [8] Dore O., et al., 2019, *BAAS*, 51, 341
- [9] Troxel M. A., et al., 2018, *Phys. Rev. D*, 98, 043528
- [10] Kilbinger M., 2015, *Reports on Progress in Physics*, 78, 086901
- [11] Zaldarriaga M., Scoccimarro R., 2003, *The Astrophysical Journal*, 584, 559
- [12] Takada M., Jain B., 2003, *Monthly Notices of the Royal Astronomical Society*, 344, 857
- [13] Kratochvil J. M., Haiman Z., May M., 2010, *PRD*, 81, 043519
- [14] Dietrich J. P., Hartlap J., 2010, *Monthly Notices of the Royal Astronomical Society*, 402, 1049
- [15] Parroni C., Cardone V. F., Maoli R., Scaramella R., 2020, *Astronomy & Astrophysics*, 633, A71
- [16] Cheng S., Ting Y.-S., Ménard B., Bruna J., 2020, *Monthly Notices of the Royal Astronomical Society*, 499, 5902
- [17] Gupta A., Matilla J. M. Z., Hsu D., Haiman Z., 2018, *Physical Review D*, 97, 103515
- [18] Fluri J., Kacprzak T., Refregier A., Amara A., Lucchi A., Hofmann T., 2018, *Physical Review D*, 98, 123518

- [19] Lu T., Haiman Z., Matilla J. M. Z., 2022, *Monthly Notices of the Royal Astronomical Society*, 511, 1518
- [20] Zhong K., Gatti M., Jain B., 2024, *Phys. Rev. D*, 110, 043535
- [21] Petri A., Liu J., Haiman Z., May M., Hui L., Kratochvil J. M., 2015, *Physical Review D*, 91, 103511
- [22] Gatti M., et al., 2020, *Monthly Notices of the Royal Astronomical Society*, 498, 4060
- [23] Valogiannis G., Dvorkin C., 2022, *Physical Review D*, 106, 103509
- [24] Lu T., Haiman Z., Li X., 2023, *Monthly Notices of the Royal Astronomical Society*, 521, 2050
- [25] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I., 2017 ([arXiv:1706.03762](https://arxiv.org/abs/1706.03762))
- [26] Dosovitskiy A., et al., 2021, [doi:10.48550/arXiv.2010.11929](https://doi.org/10.48550/arXiv.2010.11929), <http://arxiv.org/abs/2010.11929>
- [27] Carion N., Massa F., Synnaeve G., Usunier N., Kirillov A., Zagoruyko S., 2020, <http://arxiv.org/abs/2005.12872>
- [28] Lin J. Y.-Y., Liao S.-M., Huang H.-J., Kuo W.-T., Ou O. H.-M., 2022, [doi:10.48550/arXiv.2110.01024](https://doi.org/10.48550/arXiv.2110.01024), <http://arxiv.org/abs/2110.01024>
- [29] Huang K.-W., Chen G. C.-F., Chang P.-W., Lin S.-C., Hsu C.-J., Thengane V., Lin J. Y.-Y., 2022, <http://arxiv.org/abs/2210.04143>
- [30] Jia P., Zheng Y., Wang M., Yang Z., 2023, *Astronomy and Computing*, 42, 100687
- [31] Hwang S. Y., Sabiu C. G., Park I., Hong S. E., 2023, [doi:10.48550/arXiv.2304.08192](https://doi.org/10.48550/arXiv.2304.08192), <http://arxiv.org/abs/2304.08192>
- [32] Abnar S., Zuidema W., 2020, <http://arxiv.org/abs/2005.00928>
- [33] Chefer H., Gur S., Wolf L., 2021, <http://arxiv.org/abs/2012.09838>
- [34] Liu Z., Lin Y., Cao Y., Hu H., Wei Y., Zhang Z., Lin S., Guo B., 2021, *arXiv e-prints*, p. [arXiv:2103.14030](https://arxiv.org/abs/2103.14030)
- [35] Bonassi F. V., You L., West M., 2011, *Statistical Applications in Genetics and Molecular Biology*, 10
- [36] Fan Y., Nott D. J., Sisson S. A., 2012 ([arXiv:1212.1479](https://arxiv.org/abs/1212.1479))
- [37] Papamakarios G., Murray I., 2018 ([arXiv:1605.06376](https://arxiv.org/abs/1605.06376))
- [38] Lueckmann J.-M., Bassetto G., Karaletsos T., Macke J. H., 2019 ([arXiv:1805.09294](https://arxiv.org/abs/1805.09294))
- [39] Alsing J., Wandelt B., Feeney S., 2018, *Monthly Notices of the Royal Astronomical Society*, 477, 2874
- [40] Alsing J., Charnock T., Feeney S., Wandelt B., 2019, *Monthly Notices of the Royal Astronomical Society*, p. stz1960
- [41] Lemos P., Cranmer M., Abidi M., Hahn C., Eickenberg M., Massara E., Yallup D., Ho S., 2023, *Machine Learning: Science and Technology*, 4, 01LT01
- [42] Lueckmann J.-M., Boelts J., Greenberg D. S., Gonçalves P. J., Macke J. H., 2021, *arXiv e-prints*, p. [arXiv:2101.04653](https://arxiv.org/abs/2101.04653)
- [43] Guo C., Pleiss G., Sun Y., Weinberger K. Q., 2017, *arXiv e-prints*, p. [arXiv:1706.04599](https://arxiv.org/abs/1706.04599)
- [44] Hermans J., Delaunoy A., Rozet F., Wehenkel A., Begy V., Louppe G., 2021, *arXiv e-prints*, p. [arXiv:2110.06581](https://arxiv.org/abs/2110.06581)

- [45] Lemos P., Coogan A., Hezaveh Y., Perreault-Levasseur L., 2023, *40th International Conference on Machine Learning*, 202, 19256
- [46] Zürcher D., Fluri J., Sgier R., Kacprzak T., Refregier A., 2021, *JCAP*, 2021, 028
- [47] Zürcher D., et al., 2022, *MNRAS*, 511, 2075
- [48] Gatti M., et al., 2024, *MNRAS*, 527, L115
- [49] Gatti M., et al., 2023, *arXiv e-prints*, p. arXiv:2310.17557
- [50] Loshchilov I., Hutter F., 2019 (arXiv:1711.05101), <https://arxiv.org/abs/1711.05101>
- [51] Zhong K., Bernstein G., Boruah S. S., Jain B., Kobla S., 2025, *Phys. Rev. D*, 111, 103507
- [52] Lemos P., Coogan A., Hezaveh Y., Perreault-Levasseur L., 2023 (arXiv:2302.03026), <https://arxiv.org/abs/2302.03026>
- [53] Liu Z., Lin Y., Cao Y., Hu H., Wei Y., Zhang Z., Lin S., Guo B., 2021 (arXiv:2103.14030), <https://arxiv.org/abs/2103.14030>
- [54] Liu Z., et al., 2022 (arXiv:2111.09883), <https://arxiv.org/abs/2111.09883>
- [55] Sharma D., Dai B., Seljak U., 2024, *JCAP*, 08, 010
- [56] Zhong K., Gatti M., Jain B., 2024 (arXiv:2403.01368), <https://arxiv.org/abs/2403.01368>
- [57] He K., Zhang X., Ren S., Sun J., 2015 (arXiv:1512.03385)
- [58] Barchi P., et al., 2020, *Astronomy and Computing*, 30, 100334
- [59] Fielding E., Nyirenda C. N., Vaccari M., 2021, in 2021 International Conference on Electrical, Computer and Energy Technologies (ICECET). pp 1–5, doi:10.1109/ICECET52533.2021.9698414
- [60] Ćiprijanović A., et al., 2022, *Machine Learning: Science and Technology*, 3, 035007
- [61] Andersen T. E., Owner-Petersen M., Enmark A., 2020, *Journal of Astronomical Telescopes, Instruments, and Systems*, 6, 034002
- [62] Li R., Shu Y., Su J., Feng H., Zhang G., Wang J., Liu H., 2019, *Monthly Notices of the Royal Astronomical Society*, 482, 313
- [63] Caldeira J., Wu W., Nord B., Avestruz C., Trivedi S., Story K., 2019, *Astronomy and Computing*, 28, 100307
- [64] He S., Ravanbakhsh S., Ho S., 2018
- [65] Myles J., et al., 2021, *Monthly Notices of the Royal Astronomical Society*, 505, 4249–4277

A Model Architectures

To adequately benchmark the performance of vision transformer models for weak lensing, we also explore a number of convolutional models. Convolutional Neural Networks (CNNs) are a standard deep learning tool for regressing parameters from cosmological and astrophysical multi-dimensional data [56], with built-in translation invariance and assumption of hierarchical local spatial information, due to the underlying convolution operation. As performance benchmarks, we use a simple baseline CNN with six convolutional layers feeding into a four-layer feedforward perceptron, with pooling and LeakyReLU activation functions. While we have not optimized the hyperparameters and architecture of this baseline CNN, this architecture has been shown to achieve competitive results for these types of tasks [56].

Name	Parameter Count
Baseline CNN	500K
ResNet-18	11M
ResNet-34	21M
ResNet-50	24M
ResNet-101	44M
ViT-B	86M
ViT-L	307M
ViT-H	632M
Swin-T	29M
Swin-S	50M
Swin-B	88M
Swin-L	197M

Table 1: Model names and number of parameters

Setting	B-CNN	RN-18	RN-34	RN-50	RN-101	ViT-B	ViT-L	ViT-H	Swin-T	Swin-S	Swin-B	Swin-L
NoN (RMSE)	0.6298	0.5239	0.4724	0.5084	0.5294	1.860	1.549	1.662	0.3724	0.3088	0.3209	0.2232
NoN (R^2)	0.9981	0.9987	0.9989	0.9987	0.9986	0.9831	0.9883	0.9865	0.9993	0.9995	0.9994	0.9998
T1 (RMSE)	0.7436	0.7852	0.6170	0.6227	0.7613	1.725	1.621	1.859	0.7766	0.6759	0.6480	0.5492
T1 (R^2)	0.9973	0.9970	0.9981	0.9981	0.9972	0.9855	0.9872	0.9832	0.9971	0.9978	0.9980	0.9985
T4 (RMSE)	-	1.058	0.8986	0.7118	0.8734	2.005	2.105	1.988	1.061	1.021	0.8830	0.6357
T4 (R^2)	-	0.995	0.996	0.998	0.996	0.980	0.978	0.981	0.995	0.995	0.996	0.998

Table 2: Performance comparison of different ML models across various settings.

In addition, we use residual convolutional neural networks (ResNets) introduced in He et al. [57]. Contrary to naive expectations, increasing the depth of standard convolutional networks did not improve performance due to the degradation problem where an optimizer is unable to approach identity mapping; He et al. [57] circumvented this by adding non-trainable residual connections that allow the model to *skip* layers through connections of identity or projection mappings. Thus, ResNets are able to be deeper than conventional CNNs; ResNet50 specifically has 50 layers structured in several *bottleneck* blocks that are shortcut by identity mappings. Beyond conventional computer vision tasks, residual convolutional networks have found contemporary applications in galaxy morphology classification [58–60], wavefront sensing for image sharpening [61], Ly- α emitter identification [62], and Cosmic Microwave Background analysis [63, 64], making it an apt choice for our comparison standard. The full set of models, along with their respective parameter counts, are detailed in Table 1.

B Detailed Model Performance

We test our models with three simulation cases. For the first case (NoNoise-1 Channel), we choose not to include observational shape noise or tomography. For the second (LSST-Y1 1 Channel), we include shape noise at the expected level of LSST-Y1, with effective source number density $n_{\text{eff}} \sim 12$. For both, we use a redshift kernel alike the DES third tomography bin [65]. The final case (LSST-Y1 Like), we implement a full tomographic map-making procedure, generating 4-channel maps, each corresponding to the four DES tomographic bins.

We test each model outlined in Table 1 against each simulation case. For each configuration, we use an 80%:10%:10% for our training, validation, and testing split. Each input is a 512×512 independent flattened patch of a convergence map, with either one or four channels. We train each network using the AdamW optimizer [50], with an early stopping patience of 30 epochs. A L2 error metric (root-mean-squared, RMSE) is used, computed over the validation dataset. Batch size is 32, except for the much larger networks ViT-L and ViT-H (each $> 300\text{M}$ parameters), where we set to 24 and 12 respectively, due to GPU memory constraints (we use Nvidia A100s with 80 GBs of memory). We use the PyTorch ReduceLROnPlateau scheduler with a 10-epoch patience and reduction rate of 0.3. For all convolutional networks and vision transformers, the initial learning rate was set to 10^{-3} and 10^{-5} respectively. Other values were set to PyTorch defaults. After running the full sweep, the Swin-L model was observed to achieve the lowest RMSE value. See Table 2 for the full results.