Convolutional Vision Transformer for Cosmology Parameter Inference

Yash Gondhalekar
yashgondhalekar567@gmail.com

Kana Moriwaki
Research Center for the Early Universe, Graduate School of Science, The University of Tokyo

Abstract

Parameter inference is a crucial task in modern cosmology that requires accurate and fast computational methods to handle the high precision and volume of observational datasets. In this study, we explore a hybrid vision transformer, the Convolution vision Transformer (CvT), which combines the benefits of vision transformers (ViTs) and convolutional neural networks (CNNs). We use this approach to infer the Ω_m and σ_8 cosmological parameters from simulated dark matter and halo fields. Our experiments indicate that the constraints on Ω_m and σ_8 obtained using CvT are better than ViT and CNN, using either dark matter or halo fields. For CvT, pretraining on dark matter fields proves advantageous for improving constraints using halo fields compared to training a model from the beginning. However, ViT and CNN do not show these benefits. The CvT is more efficient than ViT since, despite having more parameters, it requires a training time similar to that of ViT and has similar inference times. The code is available at `https://github.com/Yash-10/cvt-cosmo-inference/`.

1 Introduction

An enthralling task in cosmology is accurately estimating the cosmological parameters describing the Universe from observational data, i.e., cosmological parameter inference. The widely accepted cosmological model, the ΛCDM (Λ Cold Dark Matter), describes the Universe using a few parameters: Ω_m (the matter density, including normal and dark matter), Ω_Λ (the dark energy density; Λ, the cosmological constant, represents dark energy), h (the Hubble parameter), n_s (the spectral index of density perturbations), σ_8 (the variance in the matter distribution smoothed over spheres of radius 8 h^{-1}Mpc). The overwhelming amount of cosmological information from current and upcoming observational surveys [e.g., 13, 11] will require sound statistical methodologies to achieve this goal.

Parameter inference aims to determine the posterior distributions of cosmological model parameters given a set of observations. Traditionally, this has been achieved by comparing the two-point correlation functions or power spectra of the tracers of large-scale structure (LSS) with theoretical predictions, or by using higher-order summary statistics that extract non-Gaussian information [e.g., 5, 27]. Such methods have analytically tractable likelihoods. However, predefined summary statistics inevitably fail to fully capture the rich non-Gaussian information at non-linear scales, which makes them suboptimal. Recently, 'field-level inference' [6] has gained a lot of attention as a potential alternative to these traditional techniques due to its ability to produce tighter constraints [see, e.g., 15, 1, 7, 16]. In this case, the likelihood is untractable since cosmological parameters are directly derived from the full, non-linear distribution of matter fields. Field-level inference allows access to higher-order information (e.g., from the phases of the fields), which is otherwise inaccessible through

Machine Learning and the Physical Sciences Workshop, NeurIPS 2024.

conventional summary statistics. Neural networks are a promising solution for field-level inference due to their demonstrated capabilities to extract features from complex data efficiently.

Since neural networks use the entire non-linear (and thus noisy) distribution of matter, they must effectively extract informative multi-scale features that help link those features to the underlying cosmological model parameters. Convolutional Neural Networks (CNNs) have consistently excelled in various tasks, primarily due to their localization through convolutional kernels, translation invariance property, and learning features hierarchically (i.e., local features in earlier layers and increasingly global features in later layers as its receptive field enlarges). Consequently, CNNs have become the dominant choice for cosmological parameter inference [e.g., 20, 21, 14, 25].

CNNs also have limitations because their receptive fields are constrained to grow larger as depth progresses, but that may not be necessary. The relatively newer Vision Transformers (ViTs) [8] do not take advantage of the strong inductive biases induced by convolutions in CNNs, allowing them to learn global spatial relationships through their self-attention layers even in the earlier layers of the model. ViTs break down an image into several patches, which are then flattened and considered tokens, similar to the terminology used in natural language processing. ViTs lack some biases, so they require large datasets for training, but given this constraint is satisfied, they have shown comparable or better performance than state-of-the-art CNNs such as ResNets. The applications of ViT for cosmological parameter inference are thus compelling, but only a few studies have explored their value [9, 10] and found them competitive with CNN. We hypothesize that combining the benefits of CNNs and ViTs may alleviate their individual deficiencies and improve parameter inference.

In this study, we perform likelihood-free inference to predict the marginal posterior mean and variance of the two cosmological parameters, Ω_m and σ_8, from simulated dark matter and halo distribution using the publicly available QUIJOTE simulation suite as our dataset. We use a convolutional vision transformer (CvT) that combines the advantages of both CNNs and ViTs. CvT has been previously applied in [4], who found it better than CNN and ViT for classifying galaxy morphologies.

2 Method

Data We use the N-body simulations of the publicly available QUIJOTE simulation suite [23] to obtain the DM density and halo catalogs; we use the friends-of-friends (FoF) halo catalogs. These simulations are performed with a box size of $L = 1\,h^{-1}$Gpc. We use standard Latin-hypercube simulations with massless neutrinos that contain 512^3 cold dark matter particles and use outputs at $z = 0$. This simulation set contains 2000 simulation data, and we divide it into training, validation, and testing sets using an 80-10-10% split. Since the splitting is performed at the simulation level, all data from a simulation are either in the training, validation, or testing set; such a non-random splitting is crucial to prevent obvious bias in the results [22]. All 1600 DM data are used for pretraining, but only 200 Halo data are used for finetuning. All simulations have different random seeds with Ω_m varied in [0.1, 0.5], σ_8 varied in [0.6, 1.0], and the other astrophysical parameters (Ω_b, h, n_s) varied within their appropriate ranges.

We project the particle and halo positions from the simulation snapshots onto a 256^3 grid using the cloud-in-cell (CIC) scheme. All the halos detected in QUIJOTE are contained in halo maps, i.e., we do not apply any mass cuts. Ten random two-dimensional maps (each of thickness $\sim$3.9 h^{-1}Mpc) along each projection direction (X, Y, and Z) are selected, producing 30 maps from each simulation (these maps are individually used as inputs to the neural network). The overdensities are first calculated ($\rho/\overline{\rho}$), followed by a logarithm-like transformation given by $\log_{10}(1 + \rho/\overline{\rho})$ to reduce the dynamic range of the pixel values, followed by standardization using the mean and standard deviation of the transformed fields from the training set. Each cosmological parameter is individually normalized to $[0, 1]$ using the corresponding minimum and maximum values calculated from the training set.

Fig. 1 shows an example comparison of the two-dimensional DM density and halo maps. It shows that the halos are biased tracers of the DM density as the former captures high local overdensities in the DM distribution, and thus leads to a visually sparser distribution.

Approach We use a Convolutional vision Transformer (CvT) [26], which uses a multi-stage hierarchical structure (see Appendix A for visualization of the architecture), where each stage contains a convolutional token embedding layer followed by convolutional transformer blocks. The

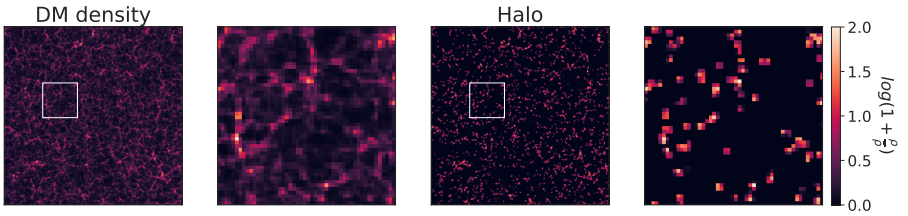

Figure 1: Sample visualization of the DM density and halo maps, extracted to show the same region from the simulation volume. The dimensions of the maps in the first and the third columns are 256×256 pixels, with thickness $\sim 3.9\,h^{-1}$Mpc. The second and the fourth columns show a zoomed inset of 50×50 pixels (and the same thickness) to better illustrate the comparison between DM density and halo distribution.

convolutional token embedding layers learn a convolution operation transforming input tokens into a new set of tokens. Its placement across different stages allows progressive spatial downsampling (i.e., reducing the number of tokens) together with increasing feature dimensions (i.e., increasing the width of tokens), and thus allows capturing local information as in CNNs. The convolutional transformer block uses a convolutional projection, implemented as a depth-wise separable convolution layer, instead of a linear projection used in traditional ViT. Ideologically, this transformer block generalizes the transformer in traditional ViT. Since local spatial relationships are modeled through the convolutional token embedding and projection, no positional encoding is required, which allows CvT to adapt to variable spatial resolution images. The use of efficient convolutions within the transformer in CvT also makes it computationally and memory-wise more efficient than traditional ViTs. The implementation is adapted from the `vit-pytorch` code[1]. Specifically, we use the lightweight CvT-13 model, i.e., with a total of 13 transformer blocks containing 17.6M parameters and the default model hyperparameters as used in the original paper.

Training details We modify the CvT architecture to perform a regression task to predict the two cosmological parameters, Ω_m and σ_8. Our model predicts the marginal posterior mean and variance for Ω_m and σ_8. The loss function is designed under the framework of moment networks [12] and used previously in works such as [24] and [25], and is given by:

$$\mathcal{L} = \sum_{i=1}^{2} \log \left(\sum_{j \in \text{batch}} \left(\theta_{i,j} - \mu_{i,j}\right)^2 \right) + \sum_{i=1}^{2} \log \left(\sum_{j \in \text{batch}} \left(\left(\theta_{i,j} - \mu_{i,j}\right)^2 - \sigma_{i,j}^2 \right)^2 \right). \quad (1)$$

where $\theta_{i,j}$ is the true value of parameter i from simulation j and $\mu_{i,j}$ and $\sigma_{i,j}$ are the network prediction of the mean and standard deviation of the marginal posterior of parameter i, respectively. During training, the 2D maps are rotated randomly by 90, 180, or 270 degrees. A batch size of 16, Adam with decoupled weight decay optimizer (AdamW; Loshchilov and Hutter 17) with weight decay of 10^{-5} and a learning rate of 5×10^{-6} is used. The learning rate is reduced by a factor of 0.3 if the validation loss does not improve for five epochs. Dropout is not used during training. Training is performed for 30 epochs in all experiments, and the model weights corresponding to the lowest validation loss are used for inference. Weights & Biases [3] was used to track training and evaluation. No specific hyperparameter tuning has been performed.

3 Results

Experimental details and evaluation We pretrain CvT on DM fields and report the test results. We also show test results when the pretrained model is finetuned on halo fields and compare its performance against the model trained from scratch on halo fields. Before finetuning, we only reinitialized the fully connected MLP head. We chose not to freeze any weights, as doing so only provided marginally better constraints on Ω_m: the RMSE was almost the same, with error bars increasing by about 0.014, reflecting the RMSE better, but constraints on σ_8 were significantly worse, with the RMSE increasing by about 0.014 and error bars underpredicted by about 0.03, than not freezing. The optimal pretrained model on DM data was found after 28k iterations, the optimal

[1] `https://github.com/lucidrains/vit-pytorch`

finetuned model on halo data after 2.4k iterations, and the optimal model trained from scratch on halo data after 1.4k iterations, although it had a higher validation loss than using transfer learning.

We used two metrics to evaluate the prediction for each cosmological parameter: the root mean squared error (RMSE), defined as $\mathrm{RMSE}_i = \sqrt{\frac{\sum_{j=1}^{N}(\theta_{i,j}-\mu_{i,j})^2}{N}}$, where N is the number of test examples, and $\bar{\sigma_i} = \frac{1}{N}\sum_{j=1}^{N}\sigma_{i,j}$ denotes the averaged errors. RMSE_i determines the accuracy of the predictions, and σ_i denotes the 1σ error in the prediction of the parameter value.

Prediction performance Fig. 2 shows the predictions versus the true parameter values for DM pretraining (a), halo transfer learning (b), and halo training from scratch (c), from left to right. σ_8 predictions for (a) show excellent agreement with a near 1:1 relationship (RMSE = 0.005 and appropriately small error bars), while Ω_m predictions are moderately good (RMSE = 0.059) and appropriate error bars. To put these values in context, we report the RMSE in the case of a constant prediction equal to the mean of the true value and obtain RMSE = 0.118 for both parameters.

Using the pretrained model from dark matter and transfer learning using halo fields (case b) gives slightly worse constraints for Ω_m than (a) (RMSE = 0.064 and underestimated error bars), but the σ_8 constraints are more prominently deteriorated (RMSE = 0.079 and slightly underestimated error bars). This deterioration is not unexpected, as halos are biased tracers of the underlying dark matter field, so they contain less pertinent information. It is currently elusive why the error bars for Ω_m are severely underestimated, but the fact that this also happens for (c) suggests that this is not necessarily due to transfer learning but a characteristic feature when using halo fields. The error bars for σ_8 are also underestimated, but this is less severe than Ω_m.

For (b), it can be seen that large Ω_m values tend to be underestimated, small σ_8 values are overestimated, whereas large σ_8 values are underestimated. So, predictions near the edges are affected and there is a tendency to regress towards the mean of the cosmological parameter set[2]. The RMSE for constant prediction is 0.121 and 0.109 for Ω_m and σ_8, respectively. Thus, this accuracy is still better than simply predicting the mean value. We do not have a clear explanation for these biases, but they may be due to insufficient expressiveness of the MLP head (since we only use a single-layered MLP) or due to overfitting (see [19] and [18], respectively).

The constraints in (c) are worse than in (b) as shown by the lower values of RMSE and $\bar{\sigma}$, and therefore the transfer learning approach (DM pretraining followed by halo finetuning) seems more beneficial than training on halo data from scratch. This can be expected because the large-scale features in the DM and halo fields are similar, so the pretrained weights of the model that is trained on DM data serve as a better starting point to learn features from the halo fields.

Comparison with traditional ViT We compare the CvT (used in this work) with the simpler version of the traditional ViT architecture discussed in Beyer et al. [2], which we dub the 'ViT'[3]. We used a patch size of 8 for the ViT, but other common hyperparameters are the same as CvT. The results for (a), (b), and (c) for Ω_m are as follows: RMSE = 0.066 and $\bar{\sigma}$ = 0.254, RMSE = 0.068 and $\bar{\sigma}$ = 0.299, RMSE = 0.074 and $\bar{\sigma}$ = 0.281. Thus, for (a) and (b), ViT is less accurate than CvT. For (c), the RMSEs are similar, but the error bar for CvT is more representative of the accuracy. For σ_8, the results for (a), (b), and (c) are: RMSE = 0.1 and $\bar{\sigma}$ = 0.24, RMSE = 0.106 and $\bar{\sigma}$ = 0.314, RMSE = 0.112 and $\bar{\sigma}$ = 0.247. However, the predictions are 'near-flat'[4] in all cases. Thus, CvT can constrain the cosmological parameters more tightly than ViT, especially σ_8.

Comparison with CNN The CNN architecture consists of five convolutional layers and batch normalization, followed by a fully connected layer that predicts the mean and standard deviation, just like the ViT, and other common hyperparameters are the same as CvT. The results for (a), (b), and (c) for Ω_m are as follows: RMSE = 0.073 and $\bar{\sigma}$ = 0.086, RMSE = 0.21 and $\bar{\sigma}$ = 0, RMSE = 0.106 and $\bar{\sigma}$ = 0.139. CNN is less accurate than ViT and CvT, and also yields an overconfident prediction for (b) ($\bar{\sigma}$ = 0). For σ_8, the results for (a), (b), and (c) are: RMSE = 0.035 and $\bar{\sigma}$ = 0.075, RMSE = 0.151 and

[2]Although we intend to talk about the mean of the 'test' parameter set here, we have checked the mean of the training parameter set is also similar which is because we randomly split the simulations.

[3]Note that this is a slightly modified version of the original ViT architecture proposed in [8]

[4]The predicted parameters are visually similar irrespective of the true parameter value when visualized like Fig. 2.

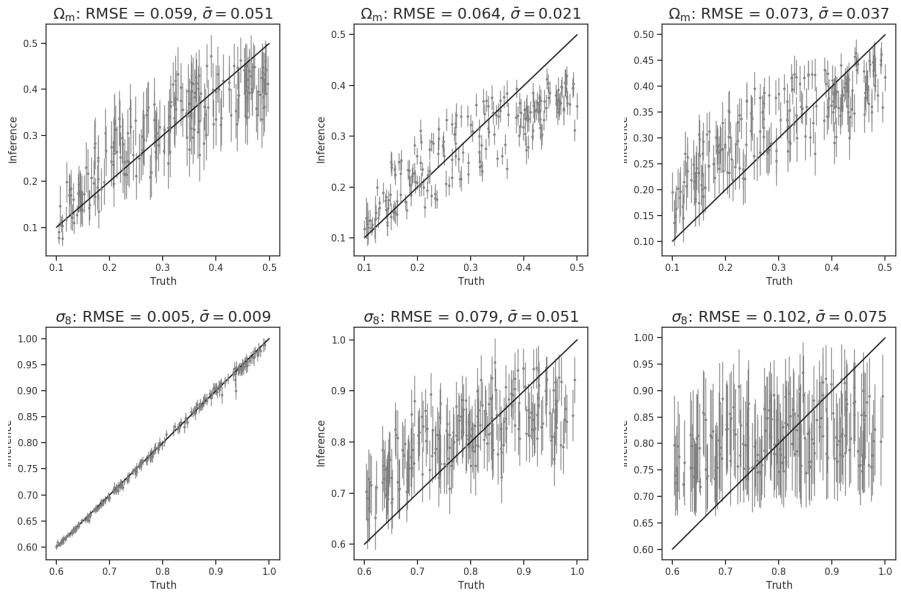

Figure 2: Comparison of predicted (y-axis) and ground-truth (x-axis) Ω_m and σ_8 cosmological parameters on the test set. The first, second, and third columns show the test results of pretraining on DM data, transfer learning on halo data, and training a model from scratch on halo data, respectively. Each data point shows the averaged values and errors across all 30 2D maps from a single 3D simulation volume. The title of each panel shows the RMSE and $\bar{\sigma}$ (see text for description).

$\bar{\sigma}$ = 0, RMSE = 0.116 and $\bar{\sigma}$ = 0.137. CNN is better than ViT to predict σ_8 for (a), but is worse than both ViT and CvT for all other cases.

Execution time The ViT used in this work contains far fewer parameters (1.6M vs. 17.6M) but requires only marginally shorter training time than the CvT ($\sim$2.6 vs. 2.7 hours) for the DM pretraining. Thus, the operations in CvT are more efficient than those in ViT, probably due to the introduction of convolutions and the convolutional projection operation [26]. At test time, CvT requires $\sim$24 seconds, whereas ViT requires $\sim$19 seconds for inference on 6000 maps (this experiment was performed when the model was trained using halo data from scratch). Although CvT is cumulatively slightly slower, the per-map inference times are almost the same.

4 Conclusion

We have applied the convolution-based vision transformer (CvT) to infer the Ω_m and σ_8 cosmological parameters using data obtained from the QUIJOTE simulation. We find that CvT constrains both parameters better than CNN and ViT when using dark matter and halo fields. CNN is found to be more beneficial than ViT only for inferring σ_8 from dark matter fields, whereas ViT outperforms in all other cases. Pretraining CvT on dark matter fields has proven beneficial in better constraining the parameters when finetuned on halo fields rather than training a model from scratch on halo data, but these benefits are not apparent for CNN and ViT. One possible interpretation is that CvT is able to effectively leverage the large-scale structure similarities between dark matter and halo fields; however, more detailed tests are necessary to validate this finding. The demonstrated constraining power of CvT is noteworthy given that it was finetuned using $8\times$ lesser data than pretraining. As a result, it may be advantageous to apply CvT on data such as galaxy distribution, which require hydrodynamic simulations that are often computationally prohibitive. We also briefly demonstrate that CvT is more efficient than ViT due to the use of convolutions and has a similar inference time to it.

Some future aims of this work are to interpret CvT, apply it to real data and develop guidelines for observational cosmologists instructing the regions to look at in the data, and integrate it with data simulation approaches based on deep learning (i.e., emulators).

Acknowledgments and Disclosure of Funding

This work was supported by JSPS KAKENHI Grant Number 23K03446.

References

[1] Adam Andrews, Jens Jasche, Guilhem Lavaux, and Fabian Schmidt. Bayesian field-level inference of primordial non-Gaussianity using next-generation galaxy surveys. *Monthly Notices of the RAS*, 520(4): 5746–5763, April 2023. doi: 10.1093/mnras/stad432.

[2] Lucas Beyer, Xiaohua Zhai, and Alexander Kolesnikov. Better plain ViT baselines for ImageNet-1k. *arXiv e-prints*, art. arXiv:2205.01580, May 2022. doi: 10.48550/arXiv.2205.01580.

[3] Lukas Biewald. Experiment tracking with weights and biases, 2020. URL `https://www.wandb.com/`. Software available from wandb.com.

[4] Jie Cao, Tingting Xu, Yuhe Deng, Linhua Deng, Mingcun Yang, Zhijing Liu, and Weihong Zhou. Galaxy morphology classification based on Convolutional vision Transformer (CvT). *Astronomy and Astrophysics*, 683:A42, March 2024. doi: 10.1051/0004-6361/202348544.

[5] Sihao Cheng, Yuan-Sen Ting, Brice Ménard, and Joan Bruna. A new approach to observational cosmology using the scattering transform. *Monthly Notices of the RAS*, 499(4):5902–5914, December 2020. doi: 10.1093/mnras/staa3165.

[6] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Science*, 117(48):30055–30062, December 2020. doi: 10.1073/pnas.1912789117.

[7] Natalí S. M. de Santi, Helen Shao, Francisco Villaescusa-Navarro, L. Raul Abramo, Romain Teyssier, Pablo Villanueva-Domingo, Yueying Ni, Daniel Anglés-Alcázar, Shy Genel, Elena Hernández-Martínez, Ulrich P. Steinwandel, Christopher C. Lovell, Klaus Dolag, Tiago Castro, and Mark Vogelsberger. Robust Field-level Likelihood-free Inference with Galaxies. *Astrophysical Journal*, 952(1):69, July 2023. doi: 10.3847/1538-4357/acd1e2.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv e-prints*, art. arXiv:2010.11929, October 2020. doi: 10.48550/arXiv.2010.11929.

[9] Kuan-Wei Huang, Geoff Chih-Fan Chen, Po-Wen Chang, Sheng-Chieh Lin, Chia-Jung Hsu, Vishal Thengane, and Joshua Yao-Yu Lin. Strong Gravitational Lensing Parameter Estimation with Vision Transformer. *arXiv e-prints*, art. arXiv:2210.04143, October 2022. doi: 10.48550/arXiv.2210.04143.

[10] Se Yeon Hwang, Cristiano G. Sabiu, Inkyu Park, and Sungwook E. Hong. The universe is worth 64^3 pixels: convolution neural network and vision transformers for cosmology. *Journal of Cosmology and Astroparticle Physics*, 2023(11):075, November 2023. doi: 10.1088/1475-7516/2023/11/075.

[11] Željko Ivezić, Steven M. Kahn, J. Anthony Tyson, Bob Abel, Emily Acosta, Robyn Allsman, David Alonso, Yusra AlSayyad, Scott F. Anderson, John Andrew, James Roger P. Angel, George Z. Angeli, Reza Ansari, Pierre Antilogus, Constanza Araujo, Robert Armstrong, Kirk T. Arndt, Pierre Astier, Éric Aubourg, Nicole Auza, Tim S. Axelrod, Deborah J. Bard, Jeff D. Barr, Aurelian Barrau, James G. Bartlett, Amanda E. Bauer, Brian J. Bauman, Sylvain Baumont, Ellen Bechtol, Keith Bechtol, Andrew C. Becker, Jacek Becla, Cristina Beldica, Steve Bellavia, Federica B. Bianco, Rahul Biswas, Guillaume Blanc, Jonathan Blazek, Roger D. Blandford, Josh S. Bloom, Joanne Bogart, Tim W. Bond, Michael T. Booth, Anders W. Borgland, Kirk Borne, James F. Bosch, Dominique Boutigny, Craig A. Brackett, Andrew Bradshaw, William Nielsen Brandt, Michael E. Brown, James S. Bullock, Patricia Burchat, David L. Burke, Gianpietro Cagnoli, Daniel Calabrese, Shawn Callahan, Alice L. Callen, Jeffrey L. Carlin, Erin L. Carlson, Srinivasan Chandrasekharan, Glenaver Charles-Emerson, Steve Chesley, Elliott C. Cheu, Hsin-Fang Chiang, James Chiang, Carol Chirino, Derek Chow, David R. Ciardi, Charles F. Claver, Johann Cohen-Tanugi, Joseph J. Cockrum, Rebecca Coles, Andrew J. Connolly, Kem H. Cook, Asantha Cooray, Kevin R. Covey, Chris Cribbs, Wei Cui, Roc Cutri, Philip N. Daly, Scott F. Daniel, Felipe Daruich, Guillaume Daubard, Greg Daues, William Dawson, Francisco Delgado, Alfred Dellapenna, Robert de Peyster, Miguel de Val-Borro, Seth W. Digel, Peter Doherty, Richard Dubois, Gregory P. Dubois-Felsmann, Josef Durech, Frossie Economou, Tim Eifler, Michael Eracleous, Benjamin L. Emmons, Angelo Fausti Neto, Henry Ferguson, Enrique Figueroa, Merlin Fisher-Levine, Warren Focke, Michael D. Foss, James Frank, Michael D. Freemon, Emmanuel Gangler, Eric Gawiser, John C. Geary, Perry Gee, Marla Geha, Charles

J. B. Gessner, Robert R. Gibson, D. Kirk Gilmore, Thomas Glanzman, William Glick, Tatiana Goldina, Daniel A. Goldstein, Iain Goodenow, Melissa L. Graham, William J. Gressler, Philippe Gris, Leanne P. Guy, Augustin Guyonnet, Gunther Haller, Ron Harris, Patrick A. Hascall, Justine Haupt, Fabio Hernandez, Sven Herrmann, Edward Hileman, Joshua Hoblitt, John A. Hodgson, Craig Hogan, James D. Howard, Dajun Huang, Michael E. Huffer, Patrick Ingraham, Walter R. Innes, Suzanne H. Jacoby, Bhuvnesh Jain, Fabrice Jammes, M. James Jee, Tim Jenness, Garrett Jernigan, Darko Jevremović, Kenneth Johns, Anthony S. Johnson, Margaret W. G. Johnson, R. Lynne Jones, Claire Juramy-Gilles, Mario Jurić, Jason S. Kalirai, Nitya J. Kallivayalil, Bryce Kalmbach, Jeffrey P. Kantor, Pierre Karst, Mansi M. Kasliwal, Heather Kelly, Richard Kessler, Veronica Kinnison, David Kirkby, Lloyd Knox, Ivan V. Kotov, Victor L. Krabbendam, K. Simon Krughoff, Petr Kubánek, John Kuczewski, Shri Kulkarni, John Ku, Nadine R. Kurita, Craig S. Lage, Ron Lambert, Travis Lange, J. Brian Langton, Laurent Le Guillou, Deborah Levine, Ming Liang, Kian-Tat Lim, Chris J. Lintott, Kevin E. Long, Margaux Lopez, Paul J. Lotz, Robert H. Lupton, Nate B. Lust, Lauren A. MacArthur, Ashish Mahabal, Rachel Mandelbaum, Thomas W. Markiewicz, Darren S. Marsh, Philip J. Marshall, Stuart Marshall, Morgan May, Robert McKercher, Michelle McQueen, Joshua Meyers, Myriam Migliore, Michelle Miller, David J. Mills, Connor Miraval, Joachim Moeyens, Fred E. Moolekamp, David G. Monet, Marc Moniez, Serge Monkewitz, Christopher Montgomery, Christopher B. Morrison, Fritz Mueller, Gary P. Muller, Freddy Muñoz Arancibia, Douglas R. Neill, Scott P. Newbry, Jean-Yves Nief, Andrei Nomerotski, Martin Nordby, Paul O'Connor, John Oliver, Scot S. Olivier, Knut Olsen, William O'Mullane, Sandra Ortiz, Shawn Osier, Russell E. Owen, Reynald Pain, Paul E. Palecek, John K. Parejko, James B. Parsons, Nathan M. Pease, J. Matt Peterson, John R. Peterson, Donald L. Petravick, M. E. Libby Petrick, Cathy E. Petry, Francesco Pierfederici, Stephen Pietrowicz, Rob Pike, Philip A. Pinto, Raymond Plante, Stephen Plate, Joel P. Plutchak, Paul A. Price, Michael Prouza, Veljko Radeka, Jayadev Rajagopal, Andrew P. Rasmussen, Nicolas Regnault, Kevin A. Reil, David J. Reiss, Michael A. Reuter, Stephen T. Ridgway, Vincent J. Riot, Steve Ritz, Sean Robinson, William Roby, Aaron Roodman, Wayne Rosing, Cecille Roucelle, Matthew R. Rumore, Stefano Russo, Abhijit Saha, Benoit Sassolas, Terry L. Schalk, Pim Schellart, Rafe H. Schindler, Samuel Schmidt, Donald P. Schneider, Michael D. Schneider, William Schoening, German Schumacher, Megan E. Schwamb, Jacques Sebag, Brian Selvy, Glenn H. Sembroski, Lynn G. Seppala, Andrew Serio, Eduardo Serrano, Richard A. Shaw, Ian Shipsey, Jonathan Sick, Nicole Silvestri, Colin T. Slater, J. Allyn Smith, R. Chris Smith, Shahram Sobhani, Christine Soldahl, Lisa Storrie-Lombardi, Edward Stover, Michael A. Strauss, Rachel A. Street, Christopher W. Stubbs, Ian S. Sullivan, Donald Sweeney, John D. Swinbank, Alexander Szalay, Peter Takacs, Stephen A. Tether, Jon J. Thaler, John Gregg Thayer, Sandrine Thomas, Adam J. Thornton, Vaikunth Thukral, Jeffrey Tice, David E. Trilling, Max Turri, Richard Van Berg, Daniel Vanden Berk, Kurt Vetter, Francoise Virieux, Tomislav Vucina, William Wahl, Lucianne Walkowicz, Brian Walsh, Christopher W. Walter, Daniel L. Wang, Shin-Yawn Wang, Michael Warner, Oliver Wiecha, Beth Willman, Scott E. Winters, David Wittman, Sidney C. Wolff, W. Michael Wood-Vasey, Xiuqin Wu, Bo Xin, Peter Yoachim, and Hu Zhan. LSST: From Science Drivers to Reference Design and Anticipated Data Products. *Astrophysical Journal*, 873(2):111, March 2019. doi: 10.3847/1538-4357/ab042c.

[12] Niall Jeffrey and Benjamin D. Wandelt. Solving high-dimensional parameter inference: marginal posterior densities & Moment Networks. *arXiv e-prints*, art. arXiv:2011.05991, November 2020. doi: 10.48550/arXiv.2011.05991.

[13] R. Laureijs, J. Amiaux, S. Arduini, J. L. Augueres, J. Brinchmann, R. Cole, M. Cropper, C. Dabin, L. Duvet, A. Ealet, B. Garilli, P. Gondoin, L. Guzzo, J. Hoar, H. Hoekstra, R. Holmes, T. Kitching, T. Maciaszek, Y. Mellier, F. Pasian, W. Percival, J. Rhodes, G. Saavedra Criado, M. Sauvage, R. Scaramella, L. Valenziano, S. Warren, R. Bender, F. Castander, A. Cimatti, O. Le Fèvre, H. Kurki-Suonio, M. Levi, P. Lilje, G. Meylan, R. Nichol, K. Pedersen, V. Popa, R. Rebolo Lopez, H. W. Rix, H. Rottgering, W. Zeilinger, F. Grupp, P. Hudelot, R. Massey, M. Meneghetti, L. Miller, S. Paltani, S. Paulin-Henriksson, S. Pires, C. Saxton, T. Schrabback, G. Seidel, J. Walsh, N. Aghanim, L. Amendola, J. Bartlett, C. Baccigalupi, J. P. Beaulieu, K. Benabed, J. G. Cuby, D. Elbaz, P. Fosalba, G. Gavazzi, A. Helmi, I. Hook, M. Irwin, J. P. Kneib, M. Kunz, F. Mannucci, L. Moscardini, C. Tao, R. Teyssier, J. Weller, G. Zamorani, M. R. Zapatero Osorio, O. Boulade, J. J. Foumond, A. Di Giorgio, P. Guttridge, A. James, M. Kemp, J. Martignac, A. Spencer, D. Walton, T. Blümchen, C. Bonoli, F. Bortoletto, C. Cerna, L. Corcione, C. Fabron, K. Jahnke, S. Ligori, F. Madrid, L. Martin, G. Morgante, T. Pamplona, E. Prieto, M. Riva, R. Toledo, M. Trifoglio, F. Zerbi, F. Abdalla, M. Douspis, C. Grenet, S. Borgani, R. Bouwens, F. Courbin, J. M. Delouis, P. Dubath, A. Fontana, M. Frailis, A. Grazian, J. Koppenhöfer, O. Mansutti, M. Melchior, M. Mignoli, J. Mohr, C. Neissner, K. Noddle, M. Poncet, M. Scodeggio, S. Serrano, N. Shane, J. L. Starck, C. Surace, A. Taylor, G. Verdoes-Kleijn, C. Vuerli, O. R. Williams, A. Zacchei, B. Altieri, I. Escudero Sanz, R. Kohley, T. Oosterbroek, P. Astier, D. Bacon, S. Bardelli, C. Baugh, F. Bellagamba, C. Benoist, D. Bianchi, A. Biviano, E. Branchini, C. Carbone, V. Cardone, D. Clements, S. Colombi, C. Conselice, G. Cresci, N. Deacon, J. Dunlop, C. Fedeli, F. Fontanot, P. Franzetti, C. Giocoli, J. Garcia-Bellido, J. Gow, A. Heavens, P. Hewett, C. Heymans, A. Holland, Z. Huang, O. Ilbert, B. Joachimi, E. Jennins, E. Kerins, A. Kiessling, D. Kirk, R. Kotak, O. Krause, O. Lahav, F. van Leeuwen, J. Lesgourgues, M. Lombardi, M. Magliocchetti, K. Maguire, E. Majerotto, R. Maoli, F. Marulli, S. Maurogordato, H. McCracken, R. McLure, A. Melchiorri, A. Merson,

M. Moresco, M. Nonino, P. Norberg, J. Peacock, R. Pello, M. Penny, V. Pettorino, C. Di Porto, L. Pozzetti, C. Quercellini, M. Radovich, A. Rassat, N. Roche, S. Ronayette, E. Rossetti, B. Sartoris, P. Schneider, E. Semboloni, S. Serjeant, F. Simpson, C. Skordis, G. Smadja, S. Smartt, P. Spano, S. Spiro, M. Sullivan, A. Tilquin, R. Trotta, L. Verde, Y. Wang, G. Williger, G. Zhao, J. Zoubian, and E. Zucca. Euclid Definition Study Report. *arXiv e-prints*, art. arXiv:1110.3193, October 2011. doi: 10.48550/arXiv.1110.3193.

[14] Andrei Lazanu. Extracting cosmological parameters from N-body simulations using machine learning techniques. *Journal of Cosmology and Astroparticle Physics*, 2021(9):039, September 2021. doi: 10.1088/1475-7516/2021/09/039.

[15] Florent Leclercq and Alan Heavens. On the accuracy and precision of correlation functions and field-level inference in cosmology. *Monthly Notices of the RAS*, 506(1):L85–L90, September 2021. doi: 10.1093/mnrasl/slab081.

[16] Pablo Lemos, Liam H. Parker, ChangHoon Hahn, Shirley Ho, Michael Eickenberg, Jiamin Hou, Elena Massara, Chirag Modi, Azadeh Moradinezhad Dizgah, Bruno Régaldo-Saint Blancard, and David Spergel. SimBIG: Field-level Simulation-based Inference of Large-scale Structure. In *Machine Learning for Astrophysics*, page 18, July 2023. doi: 10.48550/arXiv.2310.15256.

[17] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. *arXiv e-prints*, art. arXiv:1711.05101, November 2017. doi: 10.48550/arXiv.1711.05101.

[18] Michelle Ntampaka, Daniel J. Eisenstein, Sihan Yuan, and Lehman H. Garrison. A Hybrid Deep Learning Approach to Cosmological Constraints from Galaxy Redshift Surveys. *Astrophysical Journal*, 889(2):151, February 2020. doi: 10.3847/1538-4357/ab5f5e.

[19] ShuYang Pan, MiaoXin Liu, Jaime Forero-Romero, Cristiano G. Sabiu, ZhiGang Li, HaiTao Miao, and Xiao-Dong Li. Cosmological parameter estimation from large-scale structure deep learning. *Science China Physics, Mechanics, and Astronomy*, 63(11):110412, November 2020. doi: 10.1007/s11433-020-1586-3.

[20] Siamak Ravanbakhsh, Junier Oliva, Sebastien Fromenteau, Layne C. Price, Shirley Ho, Jeff Schneider, and Barnabas Poczos. Estimating Cosmological Parameters from the Dark Matter Distribution. *arXiv e-prints*, art. arXiv:1711.02033, November 2017. doi: 10.48550/arXiv.1711.02033.

[21] Dezső Ribli, Bálint Ármin Pataki, and István Csabai. An improved cosmological parameter inference scheme motivated by deep learning. *Nature Astronomy*, 3:93–98, January 2019. doi: 10.1038/s41550-018-0596-8.

[22] Dezső Ribli, Bálint Ármin Pataki, José Manuel Zorrilla Matilla, Daniel Hsu, Zoltán Haiman, and István Csabai. Weak lensing cosmology with convolutional neural networks on noisy data. *Monthly Notices of the RAS*, 490(2):1843–1860, December 2019. doi: 10.1093/mnras/stz2610.

[23] Francisco Villaescusa-Navarro, ChangHoon Hahn, Elena Massara, Arka Banerjee, Ana Maria Delgado, Doogesh Kodi Ramanah, Tom Charnock, Elena Giusarma, Yin Li, Erwan Allys, Antoine Brochard, Cora Uhlemann, Chi-Ting Chiang, Siyu He, Alice Pisani, Andrej Obuljen, Yu Feng, Emanuele Castorina, Gabriella Contardo, Christina D. Kreisch, Andrina Nicola, Justin Alsing, Roman Scoccimarro, Licia Verde, Matteo Viel, Shirley Ho, Stephane Mallat, Benjamin Wandelt, and David N. Spergel. The Quijote Simulations. *Astrophysical Journal, Supplement*, 250(1):2, September 2020. doi: 10.3847/1538-4365/ab9d82.

[24] Francisco Villaescusa-Navarro, Daniel Anglés-Alcázar, Shy Genel, David N. Spergel, Yin Li, Benjamin Wandelt, Andrina Nicola, Leander Thiele, Sultan Hassan, Jose Manuel Zorrilla Matilla, Desika Narayanan, Romeel Dave, and Mark Vogelsberger. Multifield Cosmology with Artificial Intelligence. *arXiv e-prints*, art. arXiv:2109.09747, September 2021. doi: 10.48550/arXiv.2109.09747.

[25] Francisco Villaescusa-Navarro, Shy Genel, Daniel Anglés-Alcázar, Leander Thiele, Romeel Dave, Desika Narayanan, Andrina Nicola, Yin Li, Pablo Villanueva-Domingo, Benjamin Wandelt, David N. Spergel, Rachel S. Somerville, Jose Manuel Zorrilla Matilla, Faizan G. Mohammad, Sultan Hassan, Helen Shao, Digvijay Wadekar, Michael Eickenberg, Kaze W. K. Wong, Gabriella Contardo, Yongseok Jo, Emily Moser, Erwin T. Lau, Luis Fernando Machado Poletti Valle, Lucia A. Perez, Daisuke Nagai, Nicholas Battaglia, and Mark Vogelsberger. The CAMELS Multifield Data Set: Learning the Universe's Fundamental Parameters with Artificial Intelligence. *Astrophysical Journal, Supplement*, 259(2):61, April 2022. doi: 10.3847/1538-4365/ac5ab0.

[26] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. CvT: Introducing Convolutions to Vision Transformers. *arXiv e-prints*, art. arXiv:2103.15808, March 2021. doi: 10.48550/arXiv.2103.15808.

[27] Dominik Zürcher, Janis Fluri, Raphael Sgier, Tomasz Kacprzak, and Alexandre Refregier. Cosmological forecast for non-Gaussian statistics in large-scale weak lensing surveys. *Journal of Cosmology and Astroparticle Physics*, 2021(1):028, January 2021. doi: 10.1088/1475-7516/2021/01/028.

A Architecture of the convolutional vision transformer

Fig. 3 shows the architecture of the CvT network. The primary modifications in CvT compared to the traditional ViT are the presence of a convolutional token embedding layer, whose presence across multiple stages resembles the design of CNNs, and the presence of a convolutional projection instead of a linear projection.

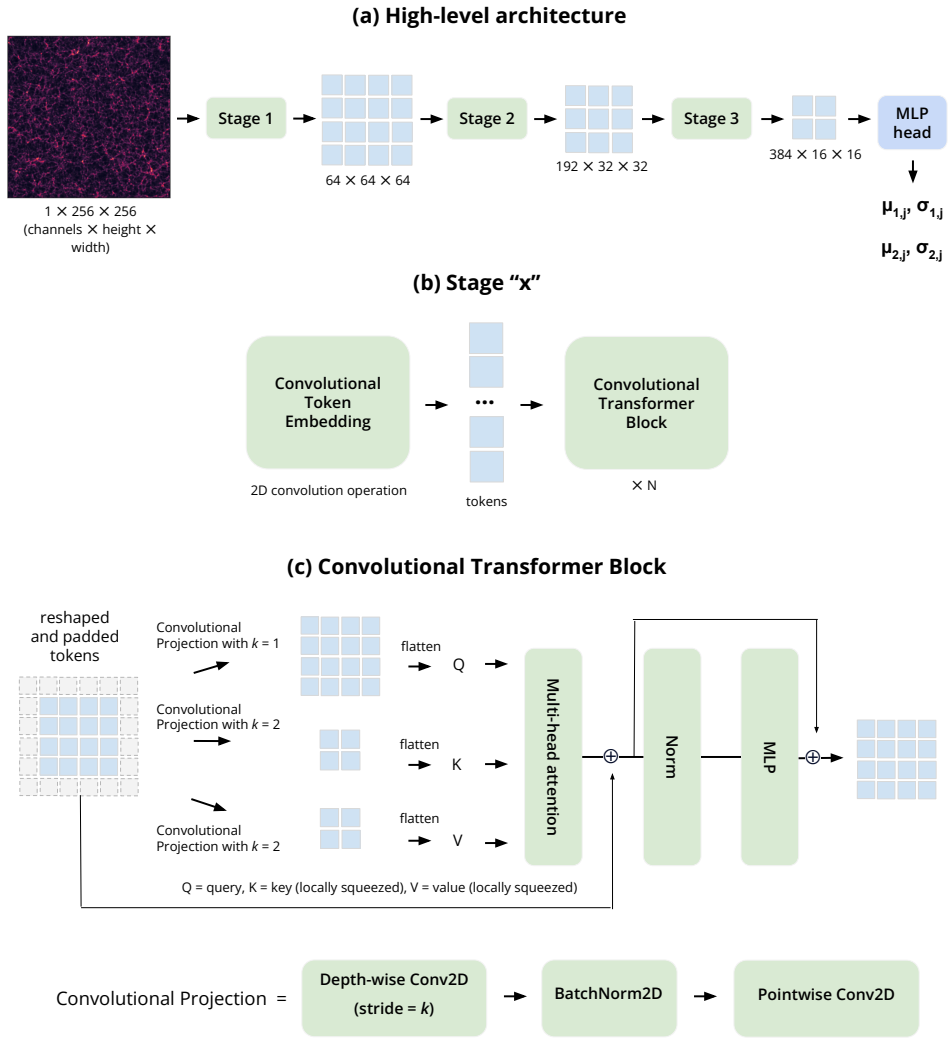

Figure 3: Architecture of the CvT network, demonstrating how convolutions and vision transformers are integrated in CvT. (a) shows CvT's hierarchical multi-stage pipeline, allowing spatial downsampling while increasing the no. of feature maps. The MLP head performs the regression to output the mean and standard deviation of the marginal posteriors of the two cosmological parameters (see "Training details" for notation). (b) shows each stage's pipeline, consisting of a convolutional token embedding layer followed by N convolutional transformer blocks. (c) details the architecture of the convolutional transformer block, which contains convolutional projection to project the query, key, and values as the first step, which is consequently passed to the multi-head self-attention module, and then normalization layer and MLP. No regression token is used.