Calibrated Deep Learning for Weak Lensing Cosmology: A Hybrid Dilated CNN Approach

NeurIPS 2025 FAIR Universe Weak Lensing ML Uncertainty Challenge
Phase 1 Submission Writeup

Hao-Chun Liang
Bo-Cheng Lai

November 17, 2025

Abstract

Weak gravitational lensing, observed as subtle, coherent distortions of galaxy images, is a key probe of the large-scale matter distribution and cosmological parameters. Deriving robust constraints from such data requires both accurate point estimators and reliable uncertainty quantification. In this work, in the context of the NeurIPS 2025 FAIR Universe Weak Lensing ML Uncertainty Challenge (Phase 1 on Codabench), we propose a compact, task-specific deep learning framework for inferring the matter density parameter Ω_m and the clustering amplitude S_8 from simulated convergence maps. Our approach uses a cosmology-tailored hybrid dilated convolutional neural network (HDC-CNN) with $\sim 2 \times 10^5$ trainable parameters, approximately fifty-fold fewer than commonly used convolutional backbones such as ResNet-18 or ConvNeXt-Tiny.

The architecture combines geometry-aware anisotropic downsampling with multi-scale context aggregation. It is designed to reflect the statistical properties of weak-lensing convergence fields and the challenge's uncertainty-focused evaluation protocol, enabling efficient training and fast inference on standard hardware. Probabilistic predictions are obtained via a compact ensemble-trained output head and further refined using post-hoc isotonic calibration, yielding well-calibrated predictive means and variances for the target cosmological parameters. On the official Phase 1 test set, our method attains a challenge score of 10.3752 (reported as 10.38 in the leaderboard), a mean squared error (MSE) of 0.1419, a coefficient of determination $R^2 = 0.8554$, and a 68% predictive-interval coverage of 0.6956. These results indicate that lightweight, task-specific architectures can deliver competitive cosmological constraints in weak-lensing analyses while maintaining reliable uncertainty quantification, and that substantial parameter compression is achievable when models are explicitly matched to the structure of the weak-lensing inference problem.

Team and Submission Details

Team name: NYCUPCS
Members: Hao-Chun Liang, Bo-Cheng Lai
Affiliations: Institute of Electronics, National Yang Ming Chiao Tung University
Submission ID: 421363

1 Introduction

Upcoming galaxy surveys such as LSST and Euclid [1, 2], together with cosmological simulations, are expected to deliver large data sets for weak gravitational lensing studies. These data will support inferences of the dark matter distribution and of cosmological parameters such as the total matter density Ω_m and clustering amplitude S_8 from lensing-induced distortions in galaxy shapes, provided that non-Gaussian information is exploited and observational and noise-related uncertainties are treated in a statistically principled manner.

Traditional cosmological analyses rely on summary statistics such as two-point correlation functions or power spectra, which capture only part of the lensing information and typically propagate uncertainties through analytical approximations. In contrast, convolutional neural networks (CNNs) trained on simulated lensing mass maps can learn complex non-Gaussian features, such as peak statistics and small-scale structures, and thereby improve parameter estimation beyond the power spectrum alone [3, 4]. On the FAIR-Universe Cosmology Challenge baselines, a power-spectrum method attains a mean squared error (MSE) of 0.349 and a coefficient of determination R^2 of 0.6446, whereas a CNN-only model improves these values to 0.1957 and 0.8006, respectively. The Hybrid Dilated CNN (HDC-CNN) introduced here further reduces the MSE to 0.1419 and increases R^2 to 0.8554, while maintaining coverage close to the nominal 68% level. Nevertheless, many existing approaches remain focused on point estimates and do not provide rigorously quantified predictive uncertainties, which limits their use in analyses that require calibrated confidence intervals.

To address this issue, the NeurIPS 2025 FAIR Universe Weak Lensing ML Uncertainty Challenge provides a weak-lensing benchmark for uncertainty-aware machine learning methods. In Phase 1, participants are given simulated convergence maps with known cosmological parameters and are tasked with predicting Ω_m and S_8 for unseen maps together with well-calibrated uncertainty estimates. The evaluation metric, described in Section 2.2, jointly rewards accurate point predictions and realistic confidence intervals.

We tackle Phase 1 with a lightweight Hybrid Dilated Convolutional Neural Network (HDC-CNN) with approximately 2.0×10^5 trainable parameters—over an order of magnitude fewer than standard CNN backbones. The architecture integrates dilated convolutions to obtain a large effective receptive field, anisotropic filtering to respect the elongated survey geometry, and coordinate-aware layers to encode positional information.

We equip the network with a probabilistic output head and model predictive uncertainty using a β-weighted negative log-likelihood loss to balance accuracy and uncertainty, and apply post-hoc isotonic-regression calibration of predictive variances on validation data so that nominal confidence intervals better match empirical error frequencies. An ensemble of models trained on different data splits and random initializations further enhances robustness and typically yields better-calibrated predictions than a single model.

In summary, this paper introduces a lightweight HDC-CNN architecture for weak-lensing convergence maps, proposes a probabilistic training and calibration procedure, and demonstrates calibrated predictive performance on the Phase 1 FAIR-Universe benchmark of simulated convergence maps. All results are obtained on simulated data from the FAIR-Universe challenge and thus do not address potential mismatches between the simulations and real survey systematics.

2 Problem Setup and Data

2.1 Weak Lensing Map Dataset

We study Phase 1 of a weak-lensing cosmology challenge, where models are trained and evaluated on simulated convergence (κ) maps. Each map is a pixelized image of the projected mass density in a sky patch; larger values of κ correspond to stronger gravitational lensing by foreground structures. Each training example is a pair $(x_i, \boldsymbol{\theta}_i)$, where $x_i \in \mathbb{R}^{1424\times 176}$ is a convergence map and $\boldsymbol{\theta}_i = (\Omega_{m,i}, S_{8,i})^\top$ are the cosmological parameters. The parameters are drawn from the ranges $\Omega_m \in [0.1, 0.6]$ and $S_8 \in [0.6, 1.0]$.

The training set comprises 101 cosmological models (Ω_m, S_8). For each model, we simulate 256 independent realizations of the large-scale structure and observational noise, yielding 25,856 maps in total. The held-out test set contains 4,000 maps drawn from the same distribution as the training set.

The simulations incorporate noise and systematics representative of real observations, including shape noise from intrinsic galaxy ellipticities, smoothing by the finite point-spread function (PSF), photometric redshift uncertainties, and intrinsic alignments. As a result, inferring $\boldsymbol{\theta}_i$ from a single noisy κ map x_i is a high-dimensional, heteroscedastic regression problem: the model must marginalize over stochastic large-scale structure and observational noise and extract low-amplitude but informative signatures of the underlying cosmology. Because each map covers only a limited survey area, finite-volume (cosmic) variance is large, so calibrated uncertainty estimates are essential.

2.2 Evaluation Metric (Uncertainty-Aware Score)

We evaluate methods using the official challenge metric, a Gaussian log-likelihood–based score augmented with an explicit penalty on point-estimate errors. For each test map x_i, we model the predictive distribution over $\boldsymbol{\theta}_i = (\Omega_{m,i}, S_{8,i})^\top$ as

$$p(\boldsymbol{\theta}_i \mid x_i) = \mathcal{N}\big(\hat{\boldsymbol{\theta}}_i, \hat{\Sigma}_i\big), \qquad \hat{\boldsymbol{\theta}}_i = (\hat{\Omega}_{m,i}, \hat{S}_{8,i})^\top,$$

with covariance

$$\hat{\Sigma}_i = \mathrm{diag}\big(\hat{\sigma}^2_{\Omega_m,i}, \hat{\sigma}^2_{S_8,i}\big).$$

The inference score is

$$\begin{aligned}\mathrm{Score}_{\mathrm{inference}} = -\frac{1}{N_{\mathrm{test}}} \sum_{i=1}^{N_{\mathrm{test}}} \Big[&\frac{(\hat{\Omega}_{m,i} - \Omega^{\mathrm{truth}}_{m,i})^2}{\hat{\sigma}^2_{\Omega_m,i}} + \frac{(\hat{S}_{8,i} - S^{\mathrm{truth}}_{8,i})^2}{\hat{\sigma}^2_{S_8,i}} \\ &+ \ln\big(\hat{\sigma}^2_{\Omega_m,i}\big) + \ln\big(\hat{\sigma}^2_{S_8,i}\big) \\ &+ \lambda\Big((\hat{\Omega}_{m,i} - \Omega^{\mathrm{truth}}_{m,i})^2 + (\hat{S}_{8,i} - S^{\mathrm{truth}}_{8,i})^2\Big)\Big],\end{aligned} \tag{1}$$

where $\Omega^{\mathrm{truth}}_{m,i}$ and $S^{\mathrm{truth}}_{8,i}$ are the true parameters, N_{test} is the number of test maps, and $\lambda = 10^3$ controls the strength of the point-estimate penalty. The overall minus sign ensures that larger scores correspond to better performance. Up to an additive constant and a global rescaling, the first four terms coincide with the negative log-likelihood of a factorized Gaussian predictive model, whereas the last term introduces an explicit ℓ_2 penalty on the mean-squared error of the predictions.

The score emphasizes three aspects:

1. **Accuracy of the mean predictions**: The χ^2 terms together with the additional quadratic penalty make the score sensitive to mean-squared error.

2. **Calibration of predictive uncertainty**: Overconfident predictions (underestimated $\hat{\sigma}$) yield large χ^2 contributions when the true parameters deviate from the means, whereas underconfident predictions (excessively large $\hat{\sigma}$) are penalized by the $\ln(\hat{\sigma}^2)$ terms. The score rewards predictive distributions whose uncertainties are well calibrated, i.e., when approximately 68% of the true values lie within the nominal $\pm 1\sigma$ intervals for each parameter.

3. **Precision–calibration trade-off**: The metric favors uncertainty estimates that are both sharp and well calibrated: smaller variances improve the log-variance terms only if they remain consistent with the realized errors and do not induce large χ^2 penalties. Although our model predicts a full 2×2 covariance matrix, the scoring formula depends only on the marginal variances; the learned correlation primarily enters through the training loss, providing richer gradient information and potentially improving generalization.

We use $\text{Score}_{\text{inference}}$ as our primary metric and design the architecture and training objective to perform well under it, although it is not used directly as the training loss; the training objective is detailed in Section 3.2.

3 Methodology

3.1 Model architecture: hybrid dilated CNN

We use a Hybrid Dilated Convolutional Neural Network (HDC-CNN) to predict cosmological parameters from weak-lensing convergence maps. The network is designed to be parameter- and computationally efficient while capturing both small-scale peaks and large-scale modes in elongated 1424×176 κ maps (Figure 1).

Anisotropic downsampling with BlurPool. The input κ map is downsampled only along its long axis using a two-stage anti-aliased BlurPool-style scheme [5]: a fixed binomial low-pass filter with coefficients $[1, 4, 6, 4, 1]/16$ along the long axis, followed by a learnable stride-2 depthwise convolution with a 1×5 kernel. This preserves transverse resolution where detail is most critical while aggregating information along the strip direction and reducing aliasing.

Dilated convolutions and large kernels. Multi-scale context is captured by a sequence of Hybrid Dilated Convolution with Smoothing (HDC-S) blocks with dilation pairs (1, 2), (2, 5), and (3, 7). Each block uses depthwise separable convolutions, and blocks that include dilation rates ≥ 5 are preceded by a 1×3 or 1×5 depthwise pre-smoothing convolution to mitigate grid artifacts. A large-kernel residual branch with a 1×63 depthwise convolution at dilation 4 yields an effective receptive field of 249 pixels along the long axis, comparable to the post-downsampling feature-map extent. An ASPP-lite module with parallel dilated convolutions, including a branch spanning the full transverse dimension, concatenates their outputs to provide a compact multi-scale representation.

Compactness, normalization, and attention. The network has $\sim$202k trainable parameters, about $100\times$ fewer than a ResNet-50, achieved by extensive use of depthwise separable convolutions and at most 96 feature channels in intermediate layers. We apply Weight Standardization [6] to depthwise and separable convolutions, use Group Normalization with 8 groups instead of Batch Normalization, incorporate DropPath with linearly increasing drop probabilities up to 0.05 across six stages, and apply LayerScale with small, stage-dependent initialization factors. Efficient Channel Attention (ECA) modules [7] after Down-C and ASPP recalibrate channels via 1D convolutions with kernel size

$$k = \left\lfloor \frac{\log_2 C}{\gamma} + \frac{b}{\gamma} \right\rfloor_{\text{odd}},$$

with $\gamma = 2$ and $b = 1$. Here $\lfloor \cdot \rfloor_{\text{odd}}$ denotes rounding to the nearest odd integer. A 96→48→96 bottleneck in Down-B further improves expressivity at low parameter cost. We train on maps in their original orientation without geometric data augmentation.

CoordConv and output head. To expose position-dependent systematics from elongated survey masks, we add CoordConv-style channels [8]: after a 3×3 stem convolution producing 32 feature maps, we concatenate two channels containing normalized x and y coordinates in $[-1, 1]$. After global average pooling, a 64-dimensional feature vector feeds two fully connected heads: one predicts $(\hat{\Omega}_m, \hat{S}_8)$, and one outputs a Cholesky parameterization of a 2×2 covariance matrix $\Sigma = LL^\top$, from which we retain only the diagonal variances to match the evaluation metric. The heads use FP32 precision even when the backbone is trained in mixed-precision FP16, and gradients are clipped to a max norm of 1.0 for numerical stability.

In summary, we deliberately avoid very deep architectures or transformer-based attention: the inductive biases of CNNs (translation equivariance and locality) are well aligned with physical fields such as convergence maps, and preliminary experiments adding a self-attention layer on top of the CNN features did not yield observable performance gains, suggesting that the dilated convolutions already capture the relevant long-range interactions.

3.2 Training Strategy and Loss Function

Reliable uncertainty quantification in probabilistic models depends on both the training objective and the validation protocol. Our pipeline comprises three components: (i) a modified loss that re-weights the contribution of mean and variance errors, (ii) a repeated cross-validation ensemble, and (iii) a post-hoc calibration step applied to out-of-fold predictions.

Loss Function (β-NLL). We train the network by minimizing a Gaussian negative log-likelihood (NLL) over the target parameters. To mitigate the tendency of heteroscedastic Gaussian NLL training to underestimate predictive variances, we introduce a variance-dependent weight. For each sample i and parameter $d \in \{\Omega_m, S_8\}$, the loss is

$$\mathcal{L}_{\beta\text{-NLL}} = \sum_{d \in \{\Omega_m, S_8\}} \frac{1}{2} \left(\frac{(y_d - \hat{\mu}_d)^2}{\hat{\sigma}_d^2} + \ln \hat{\sigma}_d^2 \right) \left(\text{sg}(\hat{\sigma}_d^2) \right)^\beta, \tag{2}$$

where y_d is the true value, $\hat{\mu}_d$ and $\hat{\sigma}_d^2$ are the predicted mean and variance, and $\text{sg}(\cdot)$ denotes a stop-gradient operator applied to the weighting term. The hyperparameter $\beta > 0$ controls how strongly high-uncertainty samples are up-weighted; in all experiments we set $\beta = 0.3$, selected based on validation performance. For $\beta = 0$, Eq. (2) reduces to the standard diagonal Gaussian NLL.

Because gradients do not flow through the weighting factor, it does not introduce additional gradient terms beyond those of the standard Gaussian NLL: gradients with respect to $\hat{\mu}_d$ and $\hat{\sigma}_d^2$ are rescaled by a positive scalar that depends on the predicted variance but is treated as constant during backpropagation. Empirically, $\beta = 0.3$ improves both pointwise accuracy and probabilistic calibration, as assessed by the challenge metric and PIT-based diagnostics, compared to $\beta = 0$.

Implementation Note. The prediction head is implemented using a Cholesky-parameterized covariance computation, but in this work we restrict ourselves to a diagonal covariance structure for consistency with the evaluation protocol and the scalar uncertainty metric used in the benchmark. The network outputs one mean and one variance for each parameter (Ω_m and S_8), and the off-diagonal terms are set to zero. The β-weighting in Eq. (2) is applied independently per dimension.

Data Preprocessing and Normalization. Before training, we add Gaussian shape noise to

each convergence map according to the usual weak-lensing noise model

$$\sigma_{\text{noise}} = \frac{\texttt{shape\_noise}}{\sqrt{2\, n_g\, \texttt{pixel\_size}^2}}, \tag{3}$$

with $\texttt{shape\_noise} = 0.4$, galaxy density $n_g = 30/\text{arcmin}^2$, and $\texttt{pixel\_size} = 2.0$ arcmin. Noise is applied only to unmasked regions, i.e., within the survey footprint.

We compute normalization statistics on the full noisy training set using Welford's online algorithm for numerical stability and memory efficiency. The resulting mean and standard deviation are used to normalize all input images to zero mean and unit variance across all folds, ensuring consistent preprocessing. Target cosmological parameters (Ω_m, S_8) are standardized to zero mean and unit variance using statistics computed on the training portion in each cross-validation split (implemented with `StandardScaler`); this uses only training labels and therefore does not introduce label leakage from validation or test folds.

Ensemble and Cross-Validation. We adopt a repeated K-fold cross-validation scheme with $K = 4$ folds and $R = 2$ independent repeats, yielding 8 models in total. The 25,856 training maps are partitioned into 4 folds, each containing an equal mix of all 101 cosmologies but disjoint subsets of the 256 noise realizations per cosmology. This "split by systematic realizations" yields validation folds whose map realizations are distinct from those used for training while preserving the same underlying cosmology distribution. This provides a validation setting that is closer to an i.i.d. scenario, which is important for assessing uncertainty calibration; in contrast, splitting by cosmology would induce a distribution shift that is not considered in Phase 1 of the challenge.

For each repeat, we train a 4-fold cross-validation ensemble by using 3 folds (≈19,392 maps) for training and 1 fold (6,464 maps) for validation. The two repeats use different random seeds for weight initialization and data shuffling, providing additional diversity in the ensemble.

Training is performed for up to 40 epochs with early stopping based on the validation challenge metric, using a patience of 17 epochs and a minimum relative improvement of 10^{-4}. We optimize with AdamW using per-parameter-group learning rates and a OneCycleLR schedule ($\texttt{pct\_start} = 0.15$, head learning rates scaled by $1.5\times$ for the mean head and $0.7\times$ for the covariance head relative to the backbone). We use a batch size of 36, gradient clipping with max norm 1.0, and mixed precision (FP16 for the backbone, FP32 for the heads). The 8 models are trained in parallel on 8 GPUs.

Post-hoc Calibration. Probability Integral Transform (PIT) diagnostics on out-of-fold predictions for the raw 8-model ensemble reveal residual miscalibration: the PIT histograms deviate from uniformity and Kolmogorov–Smirnov (KS) tests yield p-values < 0.05 for both parameters. The PIT histograms exhibit a U-shaped pattern, consistent with under-dispersion and hence underestimated uncertainties.

To address this, we recalibrate the predicted variances using isotonic regression rather than parametric alternatives such as Platt scaling, Beta calibration, or temperature scaling, for three reasons. First, isotonic regression only assumes monotonicity—larger predicted variances should correspond to larger empirical errors—without imposing a specific functional form, which helps avoid model misspecification when the true reliability curve deviates from simple parametric families. Second, our large validation set mitigates overfitting concerns associated with non-parametric methods. Third, isotonic regression has been found to perform well in closely related settings (e.g. the FAIR Universe HiggsML Uncertainty Challenge [9] and recent strong-lensing pipelines in Euclid [10]).

We calibrate at the ensemble level, using the same aggregation scheme as at test time. For each out-of-fold validation sample, we compute the ensemble mean $\bar{\mu}_d$ and variance $\bar{\sigma}_d^2$ and record the squared residual $(y_d - \bar{\mu}_d)^2$. For each parameter $d \in \{\Omega_m, S_8\}$, we then fit an isotonic regressor g_d

that maps predicted variance to empirical squared error,

$$\tilde{\sigma}_d^2 = g_d(\bar{\sigma}_d^2), \tag{4}$$

under the constraint that g_d is non-decreasing. At test time, we apply the same mappings g_d to ensemble variances. This calibration step improves the Phase 1 challenge metric on out-of-fold validation data from 10.24 to 10.27 and can be updated or replaced without retraining the neural network.

Prediction and Ensemble Aggregation. At inference time, each of the 8 models produces a predictive mean $\mu_d^{(m)}$ and variance $\sigma_d^{2(m)}$ for each parameter d. We aggregate these predictions with uniform weights. The ensemble predictive mean is

$$\bar{\mu}_d = \frac{1}{M} \sum_{m=1}^{M} \mu_d^{(m)}, \qquad M = 8, \tag{5}$$

and the ensemble predictive variance is approximated by the average of the per-model variances,

$$\bar{\sigma}_d^2 \approx \frac{1}{M} \sum_{m=1}^{M} \sigma_d^{2(m)}. \tag{6}$$

In a full mixture-of-Gaussians treatment, the predictive variance would also include the dispersion of the means,

$$\bar{\sigma}_d^2 = \frac{1}{M} \sum_{m=1}^{M} \left(\sigma_d^{2(m)} + (\mu_d^{(m)})^2 \right) - \bar{\mu}_d^2. \tag{7}$$

In our case, however, the models have very similar performance and their predictive means for a given input are tightly clustered; we empirically observe that using the full expression instead of the approximation produces no appreciable differences in the reported metrics. We therefore adopt the simpler approximation for computational efficiency. We also experimented with score-weighted ensembling, but obtained results very similar to the uniform-weighted ensemble.

4 Results and Discussion

4.1 Performance on Challenge Metrics

Our calibrated HDC-CNN ensemble achieved a Phase 1 test score of **10.38**, as reported by the official challenge evaluation server. The corresponding calibrated out-of-fold (OOF) validation score is 10.27, close to the test score and suggesting limited overfitting in this in-distribution setting. Table 1 summarizes the performance of individual models and the final ensemble at different stages of the pipeline.

Although the best single model attains a validation score of 10.64 on its own fold, it underperforms the ensemble on the public test set. The calibrated ensemble reaches 10.38 on 4,000 test maps and provides better-calibrated predictions across folds, so we adopt it as our primary model, favoring test-set performance and robustness over fold-wise peak scores.

The ensemble yields probabilistic constraints with mean 1σ width of approximately 0.03 in S_8, which is small compared to the range of S_8 in the data (0.6–1.0), corresponding to roughly 7.5% of the dynamic range. Uncertainty estimates adapt to the information content of each map: maps with limited informative structure (e.g., dominated by noise or lacking strong nonlinear clustering features) receive larger uncertainties, whereas information-rich maps yield tighter credible intervals. This empirical behavior is consistent with the model assigning higher uncertainty to harder instances.

Table 1: Performance summary of the HDC-CNN ensemble across pipeline stages. Scores correspond to the log-likelihood–based challenge metric in Eq. (1), which is the negative of a Gaussian negative log-likelihood with an additional MSE penalty (higher is better). Fold indices are denoted as `fold.repeat`.

Stage	**Score**	**Notes**
Individual Models (Validation)		
Best single model (uncalibrated)	10.64	Fold 2.1 (EMA)
Worst single model (uncalibrated)	9.75	Fold 3.0 (RAW)
Mean individual score	10.24	Std. dev.: 0.297 (coefficient of variation 2.9%)
Ensemble (Out-of-Fold)		
Uncalibrated ensemble (OOF)	10.24	8 models, uniform weights
Calibrated ensemble (OOF)	10.27	+0.03 vs. uncalibrated OOF
Final Test Set		
Calibrated ensemble (test)	**10.38**	4,000 test maps
Improvement over calibrated OOF	+0.11	positive test–OOF gap
Variant Selection		
EMA selected	5/8	Folds 0.0, 0.1, 1.0, 2.1, 3.0
RAW selected	3/8	Folds 1.1, 2.0, 3.1

4.2 Ablation Insights

We conducted ablation studies on key architectural and training components; here we summarize the main findings.

Architectural design choices. Our HDC-CNN incorporates several domain-motivated features: dilated convolutions for multi-scale feature extraction, anisotropic downsampling to preserve vertical resolution in the elongated 1424×176 input maps, an ASPP-lite module for parallel multi-scale processing, and a Cholesky-parameterized covariance head to ensure numerical stability. These choices are motivated by the data geometry (the elongated survey strip) and cosmological domain knowledge.

Loss function and training. The β-NLL loss with $\beta = 0.3$ provided better uncertainty calibration than the standard NLL ($\beta = 0$). The β-weighting mechanism, which up-weights high-uncertainty samples by a factor of $(\sigma^2)^{0.3}$, was observed to mitigate variance collapse while maintaining stable gradient flow through the uncertainty head. The internal Cholesky parameterization of the covariance ensures positive-definite variance predictions throughout training and avoids numerical pathologies.

Ensembling and EMA. Individual fold models achieved validation scores ranging from 9.75 to 10.64 (uncalibrated), with most models in the 10.2–10.6 range. Ensembling the 8 models (4 folds $\times$ 2 repeats) improved both accuracy and calibration by averaging out individual model idiosyncrasies. Each model was trained with an exponential moving average (EMA) of the weights with decay 0.999; for every fold we compared the raw and EMA weights on validation and retained the better one. EMA was selected for 5 out of 8 models in the final ensemble and yielded small but consistent improvements in OOF score and calibration metrics.

Calibration impact. Isotonic calibration improved the OOF validation score from 10.24 (uncalibrated ensemble) to 10.27, a gain of +0.03. Although small in absolute terms, this gain is for a probabilistic metric such as the log-likelihood–based score in Eq. (1). The success of post-hoc calibration indicates that the model's base uncertainty estimates are already reasonable and primarily benefit from a monotonic rescaling, supporting the strategy of separating uncertainty learning (via β-NLL during training) from uncertainty calibration (via isotonic regression post-hoc).

4.3 Generalization and Robustness

The ensemble exhibits a "positive generalization gap": the score on the unseen test set (10.38) exceeds the calibrated OOF validation score (10.27) by 0.11, corresponding to an improvement of roughly 1%. This behavior could arise from favorable sampling in the test set (each dataset split can differ slightly in difficulty) or from the specific regularization and ensembling choices, which may limit overfitting. Cross-validation and the relatively large test set of 4,000 maps provide a stable estimate of performance. Although we cannot fully disentangle sampling effects from true generalization gains without further experimentation, the absence of degradation from validation to test is consistent with limited overfitting, which is plausible given the relatively low capacity of our architecture and the use of cross-validation and ensembling.

Phase 1 evaluates in-distribution performance: both training and test data are drawn from the same simulation setup. In Phase 2 of the challenge, models will encounter out-of-distribution (OoD) situations where certain systematic effects or cosmological parameters may differ. Our current model is not specifically tuned for unknown OoD shifts, but calibrated uncertainty estimates may increase when the input deviates from patterns seen during training; we hypothesize that this could serve as a weak indicator of distribution shift, but verifying this requires dedicated experiments. We therefore regard our Phase 1 results as evidence of reliability within the simulated in-distribution setting and leave a systematic OoD evaluation (e.g., varying baryonic feedback prescriptions, masking patterns, or noise models) and tests on real survey data to future work.

4.4 Comparison to Traditional Methods

Although a direct head-to-head comparison was outside the scope of the competition, it is useful to relate our deep-learning approach to classical cosmological inference pipelines. A typical traditional analysis computes the power spectrum of each map and then inverts a theoretical model to estimate (Ω_m, S_8), focusing on two-point statistics. In contrast, deep learning can exploit higher-order statistics and non-Gaussian information contained in the full pixel-level data.

Our model learns a mapping from convergence maps to cosmological parameters, together with uncertainty estimates that are empirically validated through cross-validation and post-hoc calibration. The Phase 1 performance indicates that such deep-learning-based pipelines can attain high-precision parameter estimates within this benchmark while automatically leveraging the full field information.

From a scientific interpretability standpoint, understanding which features the CNN uses for its predictions is important for cosmological applications. The architecture includes dilated convolutions and multi-scale feature extraction, suggesting sensitivity to both large-scale clustering patterns and small-scale structural details. Future work could incorporate interpretability studies—for example, feature visualization, occlusion tests, or Fourier-mode sensitivity analysis—to connect the network's learned representations back to physical concepts and to identify modes of information that are not captured by standard summary statistics.

5 Conclusion

We presented a solution for probabilistic inference of cosmological parameters from weak lensing maps, developed for the NeurIPS 2025 Weak Lensing Uncertainty Challenge (Phase 1). A compact 202k-parameter Hybrid Dilated CNN, tuned to the anisotropic survey geometry and multiscale lensing maps, achieves competitive performance on the benchmark when combined with our training

and calibration scheme. The model produces predictive distributions for Ω_m and S_8, which are calibrated with post-hoc isotonic regression.

To obtain reliable uncertainty estimates, we depart from conventional maximum-likelihood training. The combination of β-NLL loss, cross-validation ensembling, and isotonic regression effectively aligns predicted credible intervals with empirical error distributions. Rather than relying directly on uncalibrated network outputs, we calibrate the predictive distributions using held-out data. This suggests that, in scientific applications, ML models should be assessed not only by point-prediction accuracy but also by the quality of their uncertainty quantification.

Our model attains a challenge score of 10.38 on the test set and 10.27 on out-of-fold validation, indicating close agreement between validation and test performance. Code, trained models, and reproduction instructions are available at https://github.com/noyaboy/Weak-Lensing-Uncertainty-Challenge to facilitate verification and reuse.

5.1 Future Outlook

While Phase 1 provides a controlled setting, an important next step is to apply these techniques to real survey data and out-of-distribution scenarios. A natural extension is to adapt the pipeline to Phase 2, where potential distribution shifts (e.g., additional systematics or cosmologies outside the training range) will probe the robustness of the method. Possible improvements include training on a broader set of simulations to cover survey conditions, and exploring Bayesian neural networks or other approximate inference techniques to characterize epistemic uncertainty under extrapolation.

Another promising direction is to integrate domain knowledge into the architecture and training objectives. For instance, one could enforce physical constraints or known scaling relations (a form of scientific inductive bias), or augment training with summary statistics as auxiliary targets (hybrid ML + physics approaches). Such extensions may help ensure that the learned representations remain physically interpretable and robust to changes in survey conditions.

The methodology is, in principle, applicable beyond cosmology. Many scientific and engineering problems involve spatially distributed data and require both predictions and uncertainty estimates, including climate modeling, medical imaging diagnostics, and materials microscopy. Combining a CNN tailored to the data modality with rigorous uncertainty quantification via ensembling and calibration offers a practical blueprint for such settings. Overall, these results provide a step towards deploying uncertainty-aware deep learning methods in high-precision cosmology and related scientific domains where quantitative control over predictive uncertainties is essential.

References

[1] Ž. Ivezić, S. M. Kahn, J. A. Tyson, et al. LSST: From Science Drivers to Reference Design and Anticipated Data Products. *Astrophysical Journal*, 873:111, 2019.

[2] R. Laureijs, J. Amiaux, S. Arduini, et al. Euclid Definition Study Report. ESA/SRE(2011)1, 2011.

[3] A. Gupta, J. M. Zorrilla Matilla, D. Hsu, and Z. Haiman. Non-Gaussian information from weak lensing data via deep learning. *Physical Review D*, 97:103515, 2018.

[4] D. Ribli, B. A. Pataki, J. M. Zorrilla Matilla, et al. Weak lensing cosmology with convolutional neural networks on noisy data. *Monthly Notices of the Royal Astronomical Society*, 490:1843–1860, 2019.

[5] R. Zhang. Making convolutional networks shift-invariant again. In *International Conference on Machine Learning (ICML)*, pages 7324–7334, 2019.

[6] S. Qiao, H. Wang, C. Liu, W. Shen, and A. Yuille. Micro-batch training with batch-channel normalization and weight standardization. *arXiv preprint arXiv:1903.10520*, 2019.

[7] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu. ECA-Net: Efficient channel attention for deep convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11534–11542, 2020.

[8] R. Liu, J. Lehman, P. Molino, F. P. Such, E. Frank, A. Sergeev, and J. Yosinski. An intriguing failing of convolutional neural networks and the CoordConv solution. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 9605–9616, 2018.

[9] FAIR Universe Collaboration. The FAIR Universe HiggsML Uncertainty Challenge: Reconstructing Higgs-Boson Decay Channels with Model Uncertainty. *arXiv preprint arXiv:2410.02867v5*, 2024.

[10] Euclid Collaboration. Strong gravitational lensing and star-galaxy classification in Euclid and DES using isotonic regression calibration. *arXiv preprint arXiv:2503.15328v1*, 2025.

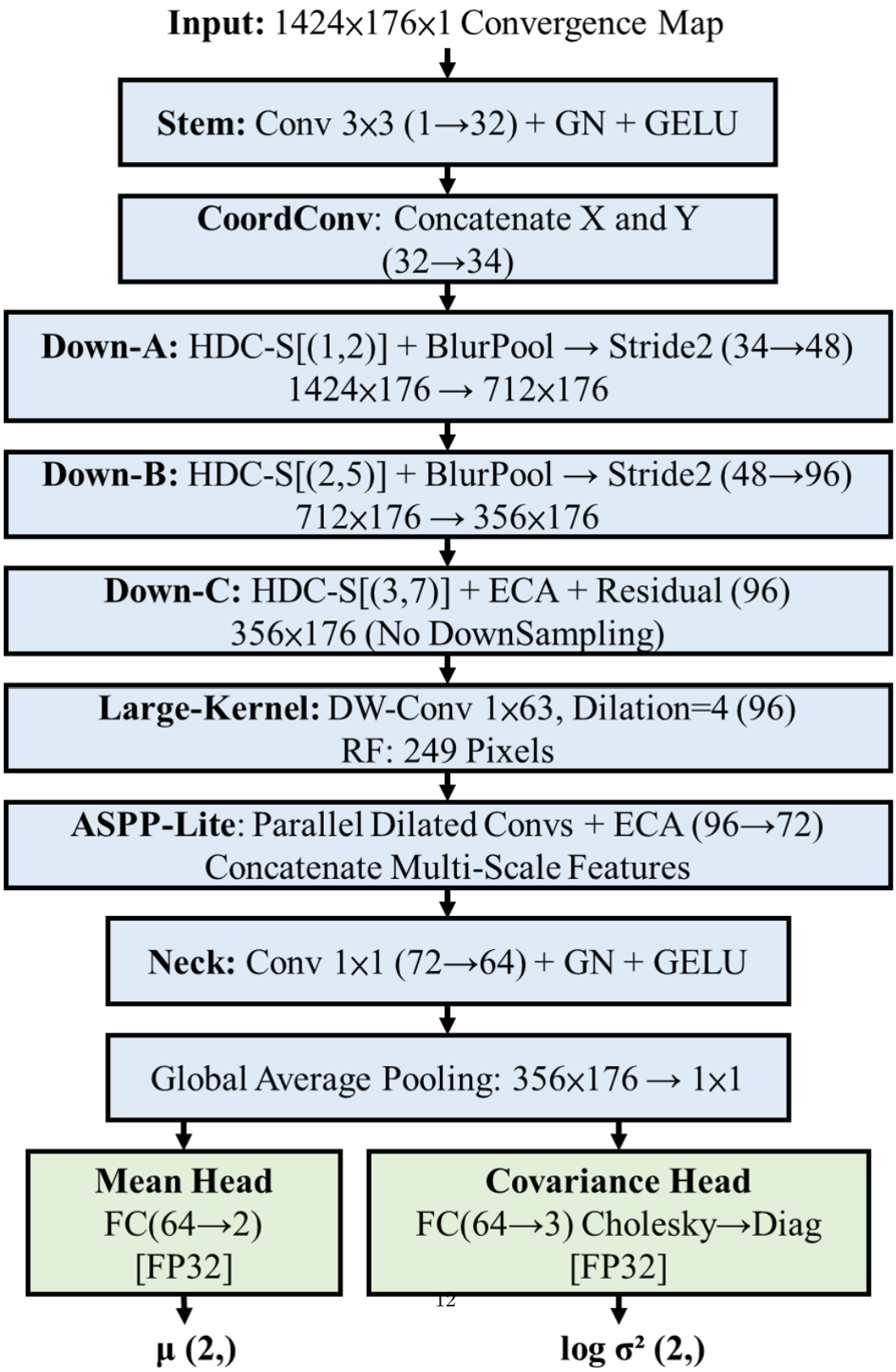

Figure 1: Overview of the HDC-CNN architecture for 1424×176 convergence maps. The network uses anisotropic downsampling, hybrid dilated convolution blocks with increasing dilation rates, a large-kernel residual branch, and an ASPP-lite module for multi-scale feature aggregation. Two FP32 heads output parameter means and variances. The model has 202k trainable parameters.