NEURIPS 2025
Inferring Cosmological Parameters with CNN K-Fold Ensembling

An overview of the 6th place solution for the Weak Lensing ML Uncertainty Challenge

Andy Zhang
azhang81@jh.edu / fortybyte@gmail.com

What is Weak Lensing

- Dark matter does not interact with light directly but has mass, thus indirectly warping light like a black hole.

- Weak lensing is like a kaleidoscope

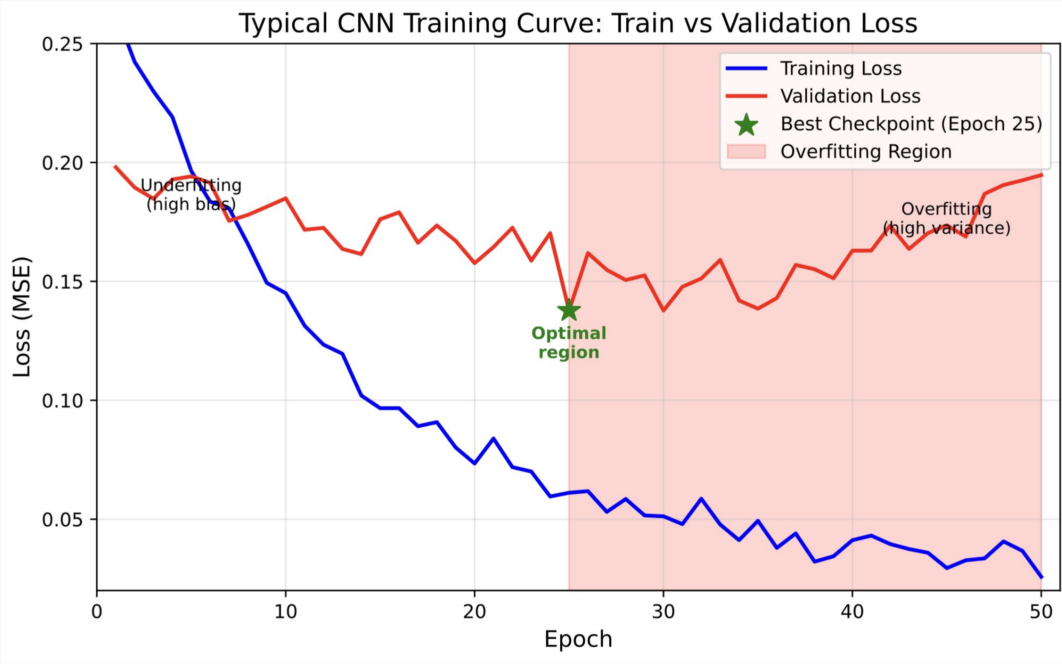
- **Different universes are kaleidoscopes with different shapes in it**

CNN

- Images classification means CNN
- **Not a lot of data**
- Perfect function from Weak Lensing Maps to Ω_m and S_8 is discrete

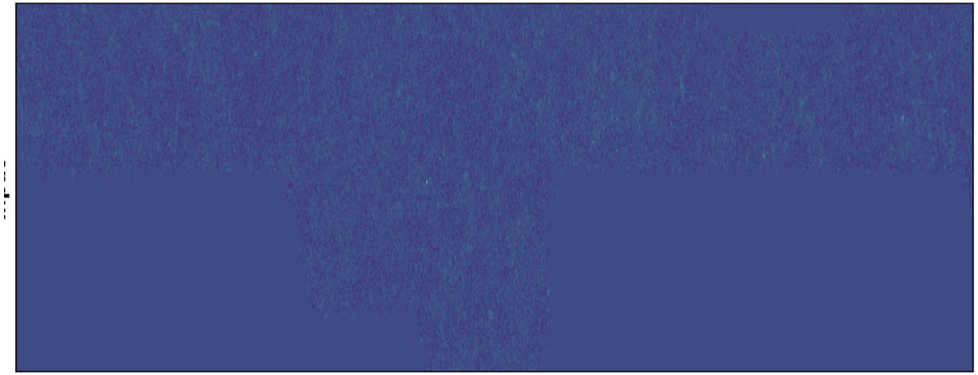
CNN

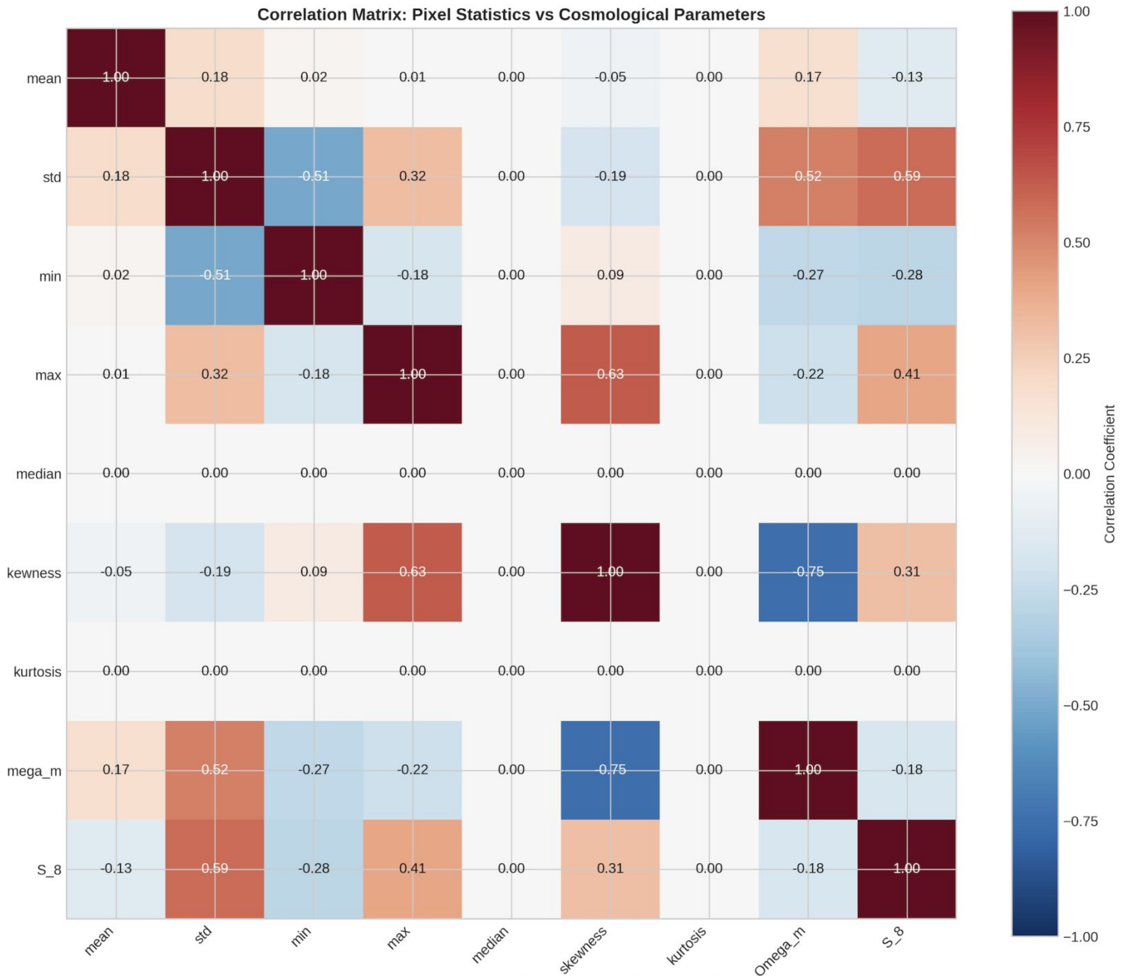
- RepVGG trained on four fifths of the data
- Obvious overfitting
- Model refused to guess exterior points



Generalization Challenge

- **Very Dense Features**
- Impossible task for human eyes
- Data analysis reveals some interesting findings

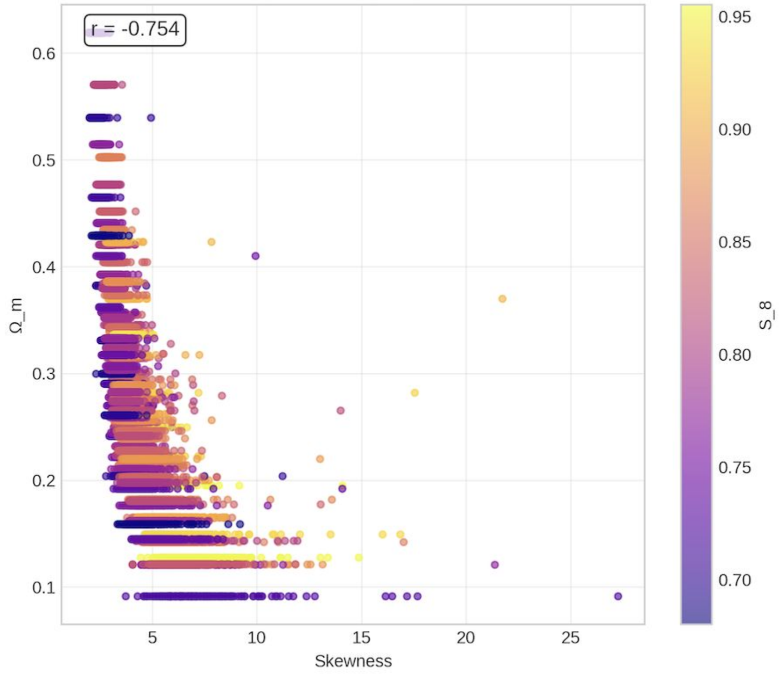




Correlation Matrix: Pixel Statistics vs Cosmological Parameters
mean
std
min
max
median
kewness
kurtosis
mega_m
S_8
1.00 0.18 0.02 0.01 0.00 -0.05 0.00 0.17 -0.13
0.18 1.00 -0.51 0.32 0.00 -0.19 0.00 0.52 0.59
0.02 -0.51 1.00 -0.18 0.00 0.09 0.00 -0.27 -0.28
0.01 0.32 -0.18 1.00 0.00 0.63 0.00 -0.22 0.41
0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
-0.05 -0.19 0.09 0.63 0.00 1.00 0.00 -0.75 0.31
0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.17 0.52 -0.27 -0.22 0.00 -0.75 0.00 1.00 -0.18
-0.13 0.59 -0.28 0.41 0.00 0.31 0.00 -0.18 1.00
mean
std
min
max
median
skewness
kurtosis
Omega_m
S_8
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Correlation Coefficient

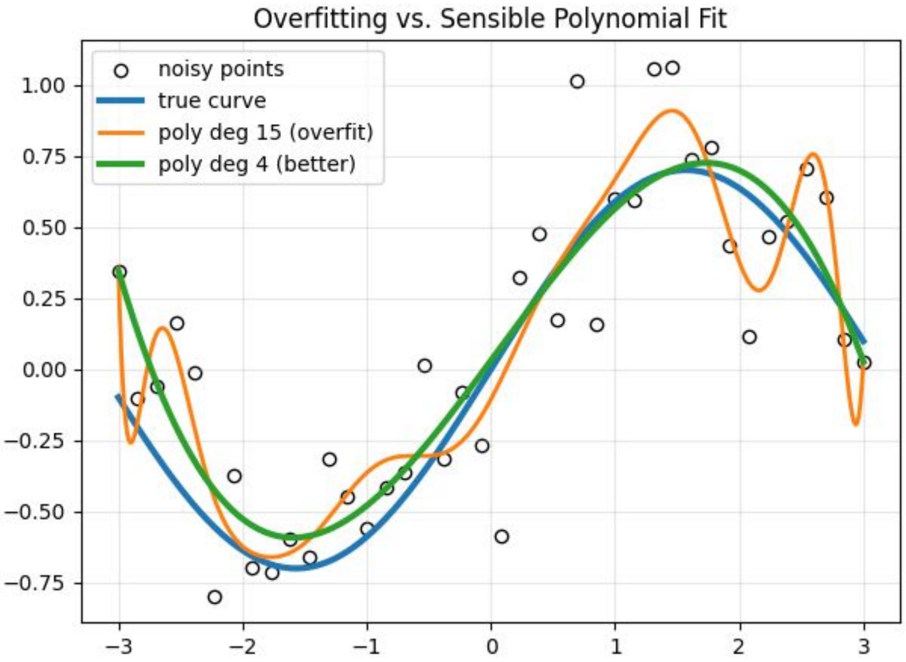
Generalization Challenge

- Ω_m is most highly correlated to skewness
- S_8 is most highly correlated to std
- Very few outliers



Generalization Challenge

- **Requires Stable Validation Loss**
- Larger Models
- Data Augmentation

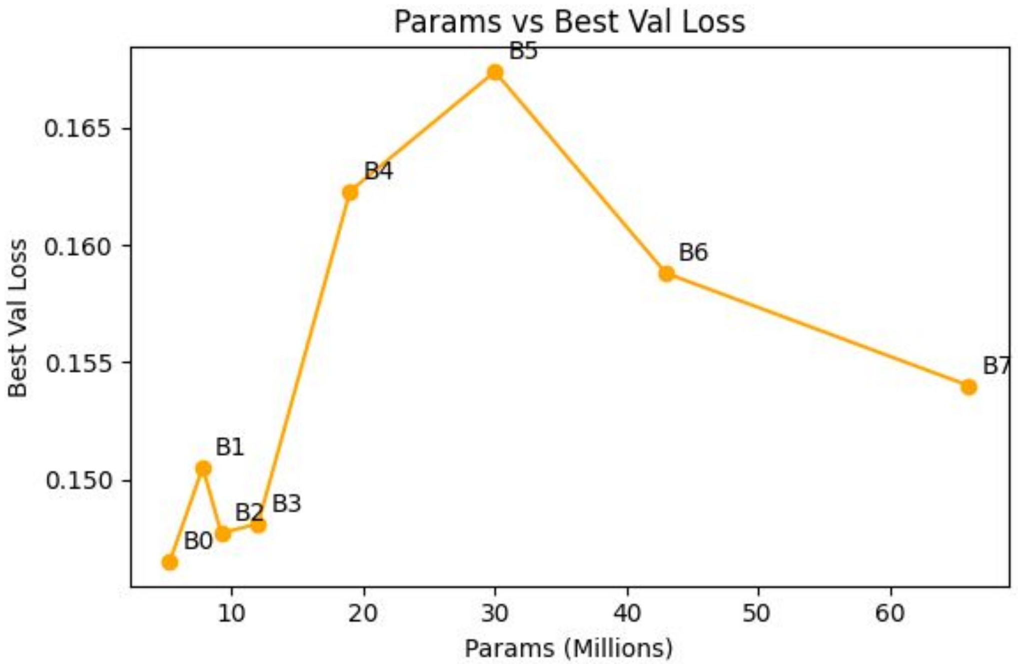


Double Descent

Model	Params	Time (min)	Best Val Loss	Best Epoch	Train Loss	Gap
B0	5.3M	19.9	0.1465	6	0.1289	0.059
B1	7.8M	24.1	0.1505	15	0.1728	-0.022
B2	9.2M	25.0	0.1477	8	0.1303	0.03
B3	12M	32.4	0.1481	6	0.1429	0.023
B4	19M	41.5	0.1623	9	0.2065	-0.037
B5	30M	56.4	0.1674	3	0.1177	0.051
B6	43M	72.0	0.1588	12	0.1172	0.042
B7	66M	96.4	0.154	6	0.0952	0.073

Double Descent

- Performance implies efficient net has double descent behavior on this dataset
- Not feasible to train models larger than B7
- **B7 train time 5 times of B0 train time**



Data Augmentation

Experimental Setup:

- EfficientNet-B3
- 15 epochs
- Batch size 16
- AdamW with 1e-3 learning rate and 1e-4 weight decay
- MSE Loss

Tested Augmentations:

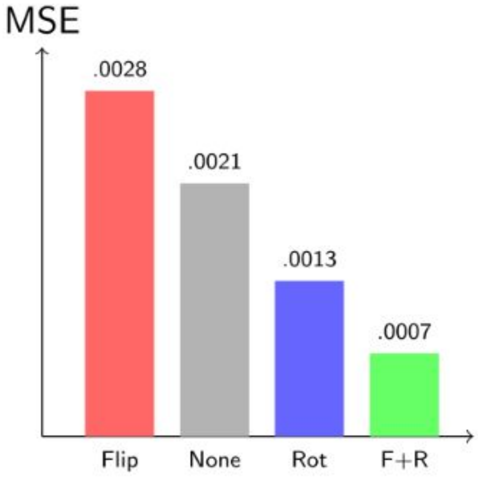
- Random Horizontal and Vertical Flips
- Random Rotation
- Dropout
- Random Noise
- Mix-up
- Coordinate Channels

Data Augmentation

- Dropout, mixup, and random noise all had negative impact

- Extremely dense problem so any slight change to individual pixel values will deteriorate prediction ability

Augmentation	Val MSE	Best Ep.	vs None
None	0.002051	11	1.0×
Flip only	0.002802	11	1.37× ↓
Rotation only	0.001261	5	0.61× ↑
Flip + Rotation	**0.000675**	15	**0.33×** ↑

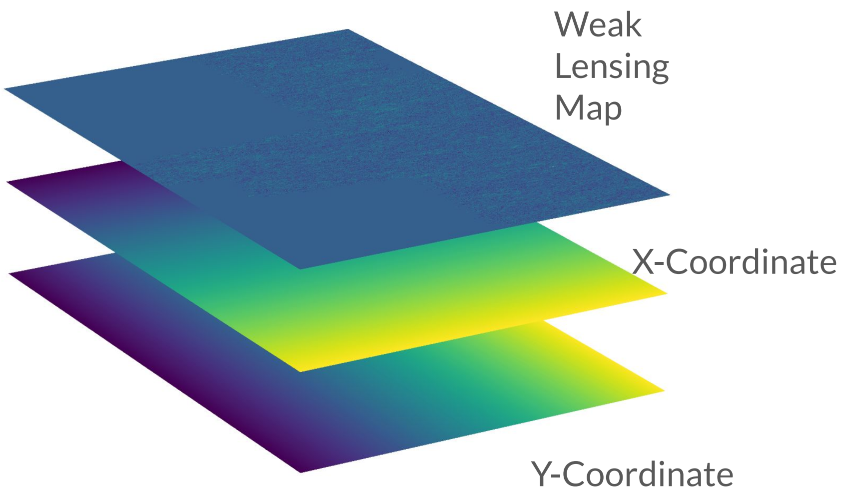


Coordinate Channels

- Second and third channel inputs became coordinates

- Specific areas may have specific indicators

- Sinusoidal or radial may perform better

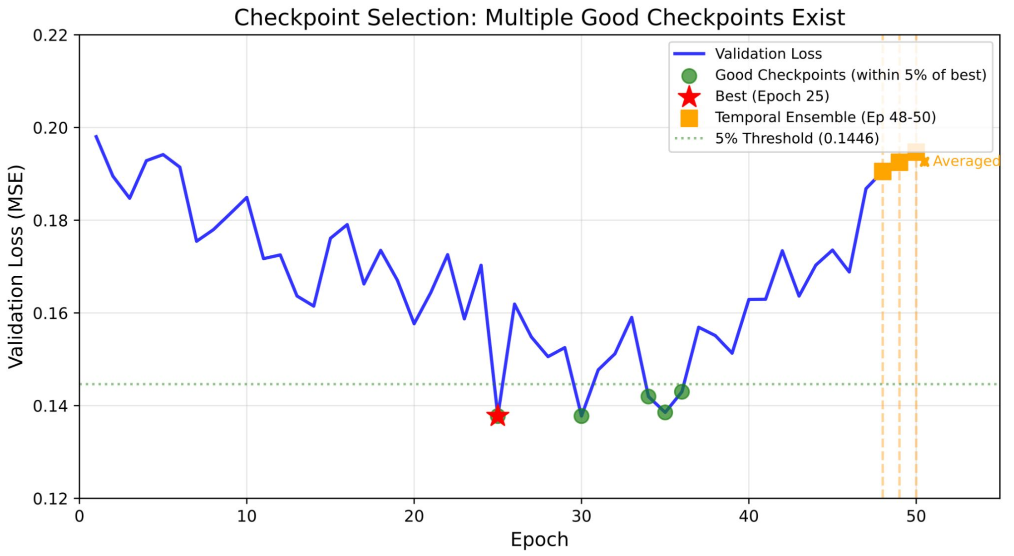


K-Fold & Ensembling

- **Does not require perfect convergence**

- Decreases variance while keeping bias

- Uses full training data



K-Fold & Ensembling

- **Does not require perfect convergence dynamics**

- Decreases variance while keeping bias

- Uses full training data

For a single model f, the expected prediction error decomposes as:

$$\mathbb{E}[(y - f(x))^2] = \underbrace{(\mathbb{E}[f(x)] - f^*(x))^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2]}_{\text{Variance}} + \underbrace{\sigma_\epsilon^2}_{\text{Irreducible}}$$

where $f^*(x)$ is the true function and σ_ϵ^2 is irreducible noise.

Submission

Core Model: RepVGG-D2se CNN regressor → predicts (Ω_m, S_8)

- Split Across 5 Folds each with equal distribution of cosmologies
- Stability first training:
  - batch size 80
  - Grad accumulation
  - Ghost BatchNorm
- Two Phase Augmentation:
  - Flips and Rotations: Epochs 1-45
  - Just Flips: epochs 45-50
  - Test Time Augmentations
- MCMC Error Bar Estimation
  - **1.2x final scaling**

K-Fold & Ensembling

Fold	Val Size	Train Size	Val Index Range
0	5,252	20,604	0–52
1–4	5,151	20,705	52–256

- Split each 256 systematics into 5 Folds
- Select top 5 checkpoints among each fold compute inference with TTA
    - All 4 combinations of horizontal and vertical flips
- Throw out the farthest 30% of inferences to the average among all inferences in the fold
- Compute the final predictions for this fold by averaging the remaining inferences

K-Fold & Ensembling

- Compute predicted points by averaging the cosmological inferences from the 5 folds

- Compute the error bars on the final model prediction using a weighted average
    - Since all folds are weighted equally w_k is 0.2

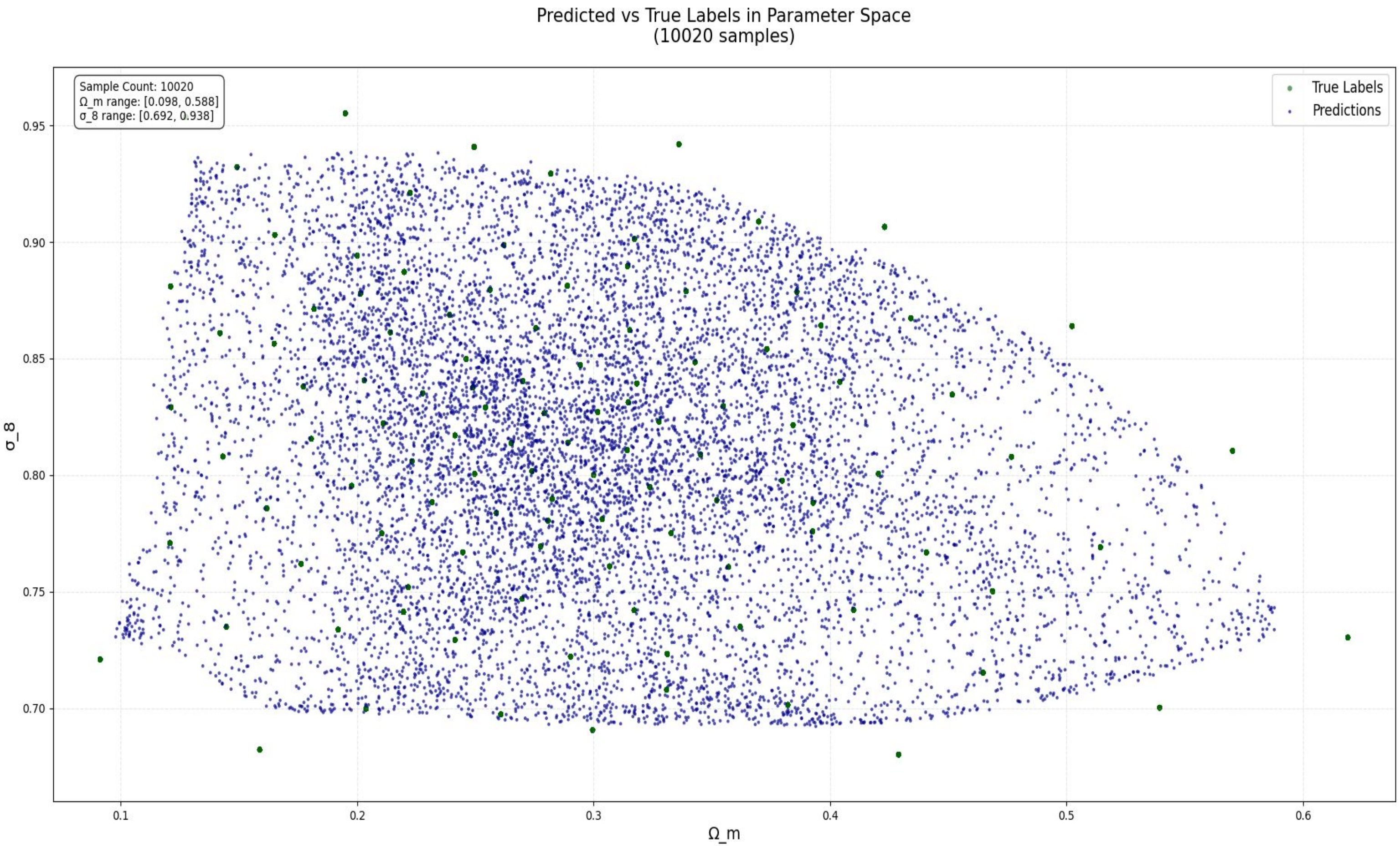
For K model predictions (μ_k, σ_k):

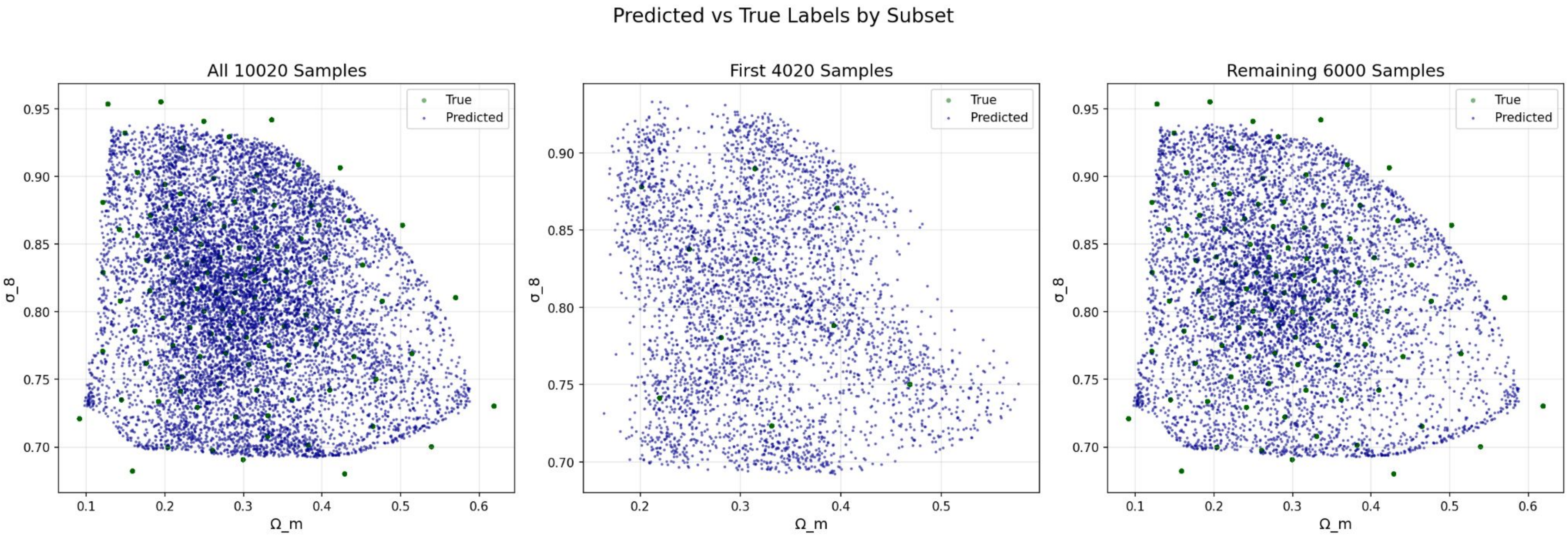
$$\sigma_{\text{mix}}^2 = \sum w_k \sigma_k^2 + \sum w_k (\mu_k - \mu_{\text{mix}})^2$$

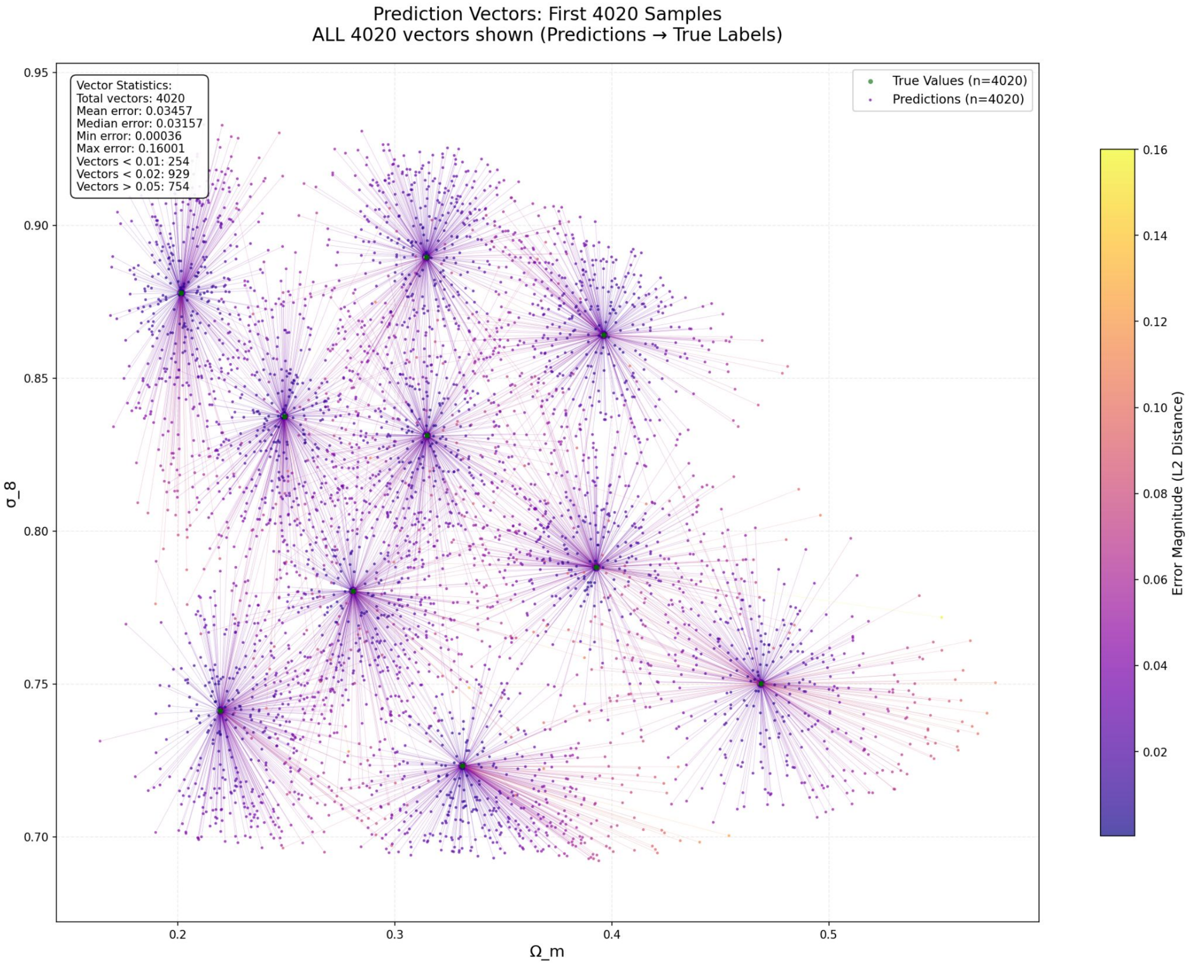
K-Fold & Ensembling

- Don’t want to waste all the previous checkpoints
- Combine error bars using the aforementioned weighted average
- At least 0.2 score increase from singular model

Model Family	Weight	Role
RepVGG-D2se	**0.50**	Primary
EfficientNet-V2-L variants	0.15	Diversity
RepVGG variants	0.15	Diversity
EfficientNet-V2-L additional	0.10	Support
EfficientNet-B7 batch80	0.05	Support
EfficientNet-B7 batch64	0.05	Support

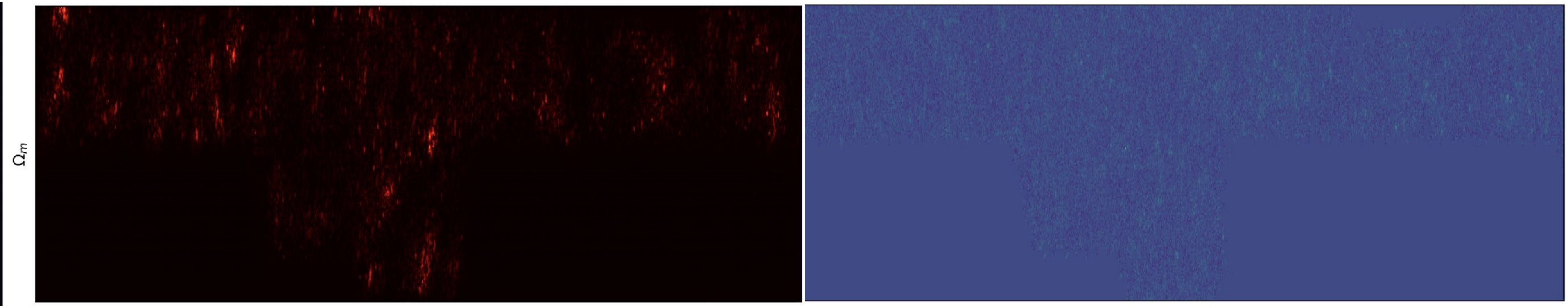

Predicted vs True Labels in Parameter Space
(10020 samples)
Sample Count: 10020
Ω_m range: [0.098, 0.588]
σ_8 range: [0.692, 0.938]
True Labels
Predictions
0.95
0.90
0.85
0.80
0.75
0.70
σ_8
0.1
0.2
0.3
0.4
0.5
0.6
Ω_m


Predicted vs True Labels by Subset
All 10020 Samples
First 4020 Samples
Remaining 6000 Samples
True
Predicted
σ_8
Ω_m
0.70
0.75
0.80
0.85
0.90
0.95
0.1
0.2
0.3
0.4
0.5
0.6



Prediction Vectors: First 4020 Samples
ALL 4020 vectors shown (Predictions → True Labels)
Vector Statistics:
Total vectors: 4020
Mean error: 0.03457
Median error: 0.03157
Min error: 0.00036
Max error: 0.16001
Vectors < 0.01: 254
Vectors < 0.02: 929
Vectors > 0.05: 754
True Values (n=4020)
Predictions (n=4020)
σ_8
Ω_m
Error Magnitude (L2 Distance)
0.95
0.90
0.85
0.80
0.75
0.70
0.2
0.3
0.4
0.5
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02

Activation: Ω_m =0.3, S_8=0.8

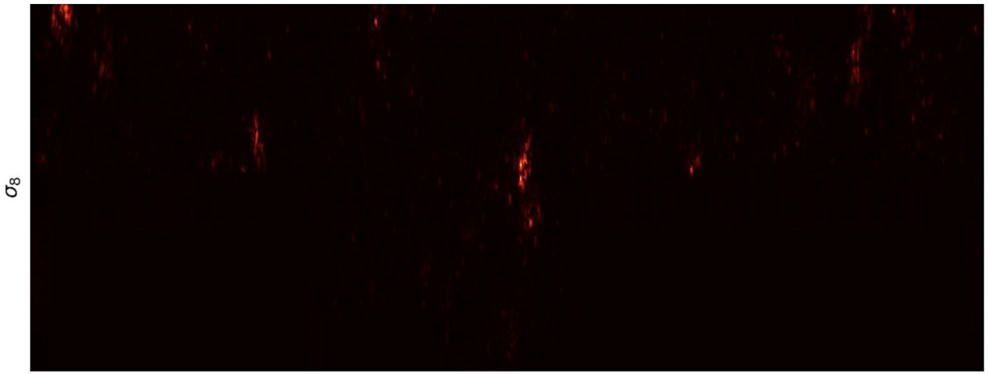


Activation: Ω_m =0.3, S_8=0.8

- Ω_m tends to depend on aspects of the entire weak lensing map

- S_8 tends to depend on a specific few aspects of the weak lensing map

- **Requires entire map to make an optimal predictions**



Submission

- 8th place on public test data

- 6th place on private test data

Score	MSE	Coverage
1. 11.4163	0.1102	0.6951
2. 10.8722	0.1121	0.6775
3. 11.1442	0.1111	0.6863

Conclusion

Traditional ML Techniques are able to go a long way, but don't get too demotivated if a novel approach does not work.

Questions & Comments

Welcome!