Toward a deep learning approach for fast galaxy generation

Tanner Sether
PhD Student
Advisor: Elena Giusarma
Michigan Technological University
tmsether@mtu.edu

Michigan Technological University 1885

MICHIGAN SPACE GRANT CONSORTIUM NASA

ML4ASTRO2

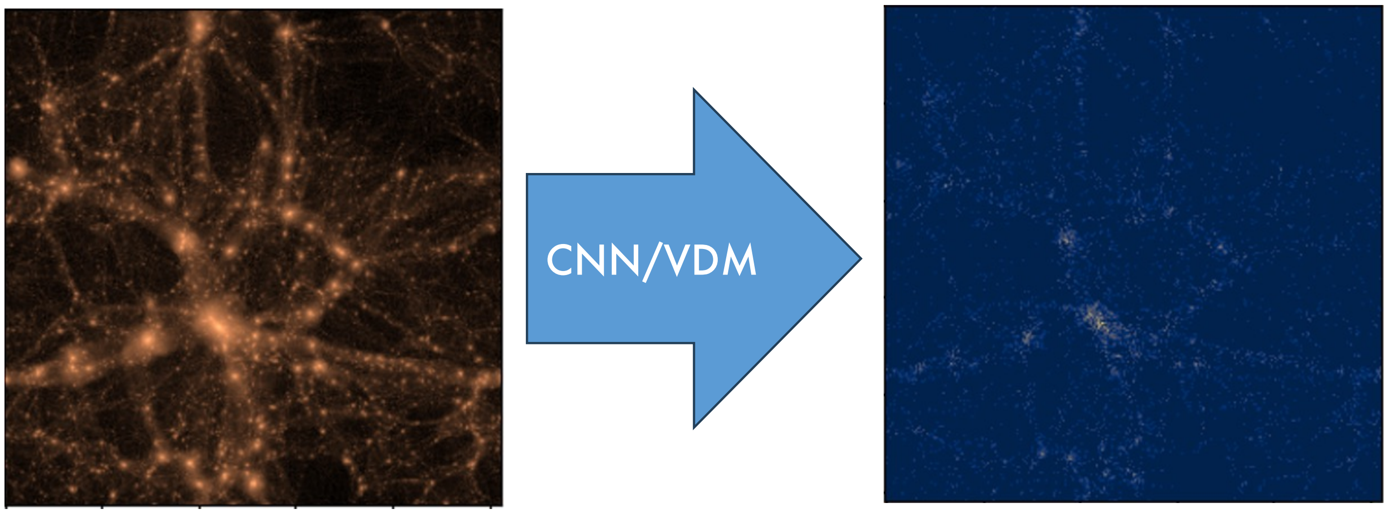
7/11/24

Introduction

- As missions collect data, simulations are needed to compare theory to prediction
- Use simulations to test parameters
  - Cosmology combined with astrophysics
- Accurate simulations can be incredibly expensive
- Cheap simulations + machine learning = success

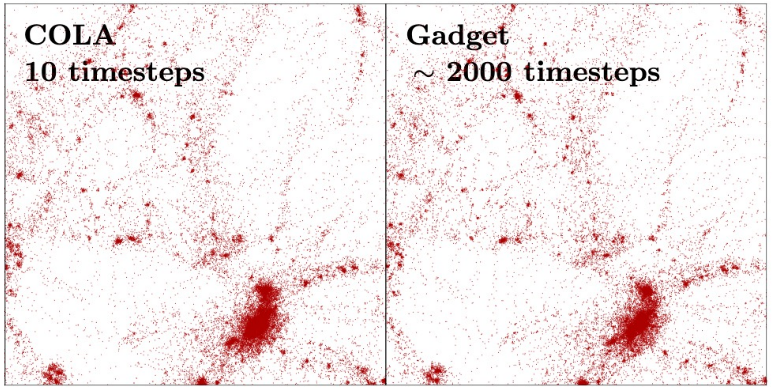
Objective

- Goal: Use NNs to map from DM to Galaxy distributions



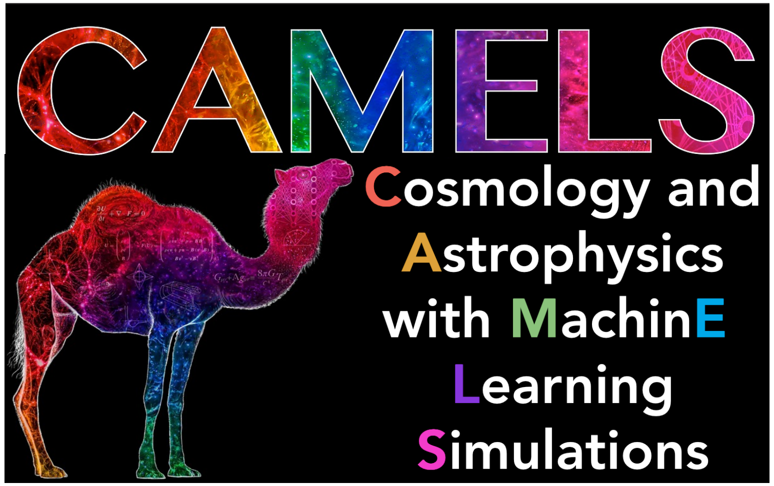
Methods – Data

- CAMELS simulations
  - N-Body and Hydrodynamic Simulations
- COLA (COmoving Lagrangian Acceleration)
  - Fast approximations to N-Body simulations
  - More computationally accessible



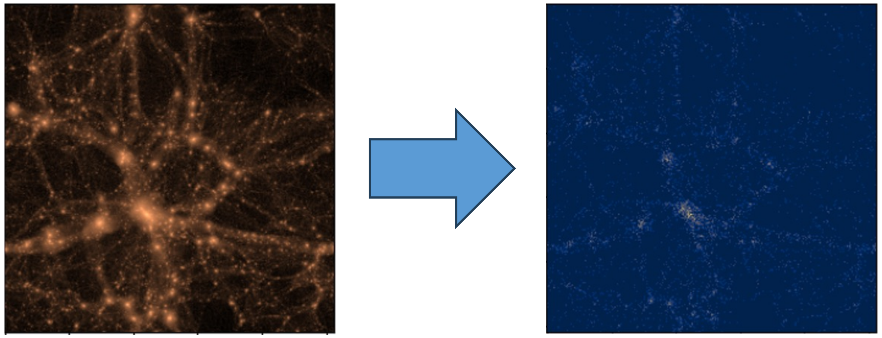
Tassev, S. et al.
arXiv:1301.0322

https://www.camel-simulations.org/



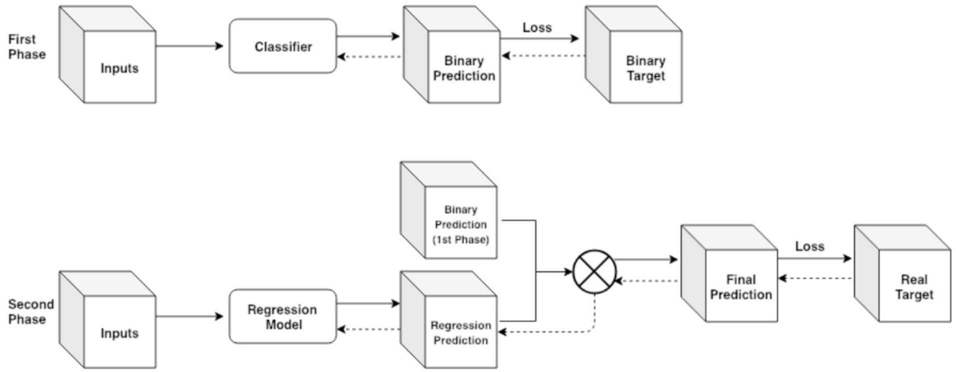
Methods – Data

- Input: DM density fields from N-Body simulations
- Target: Galaxy fields from hydrodynamic simulations
- Our data is heavily imbalanced – about 17 million particles in the input, about 18000 in the target
  - Using 256^3 voxels
- >99% accuracy possible by predicting 0 galaxies!



Methods – Network Architecture

- Use a two-phase architecture – binary classification followed by regression
  - Use classifier to determine if each voxel is likely to contain a galaxy or not
  - Regression on voxels likely to contain galaxies
- Weighted cross-entropy loss, to minimize false negatives
- Classifier: Inception or R2U-Net
- Regressor: R2U-Net



Zhang et al.
https://doi.org/10.48550/arXiv.1902.05965

Results – Classification

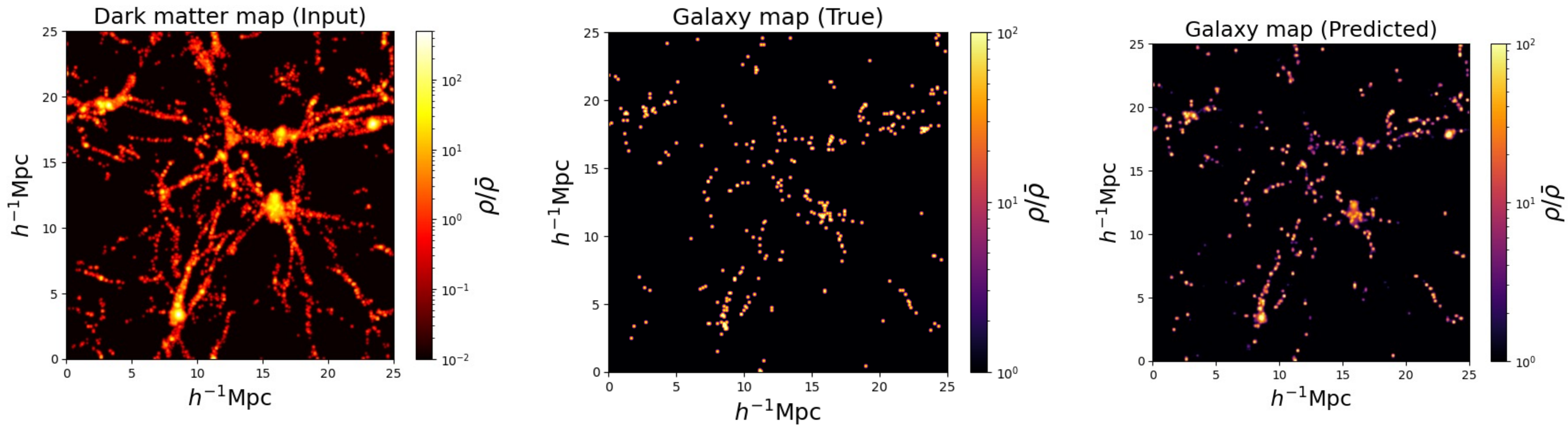
- Want to select model with highest recall with high accuracy
    - High recall is important to avoid false negatives

Model	Accuracy	Recall	Precision	Training Time (RTX A6000)
Inception	97.34	95.38	3.537	6 minutes
R2U-Net	98.24	94.16	5.196	12 minutes

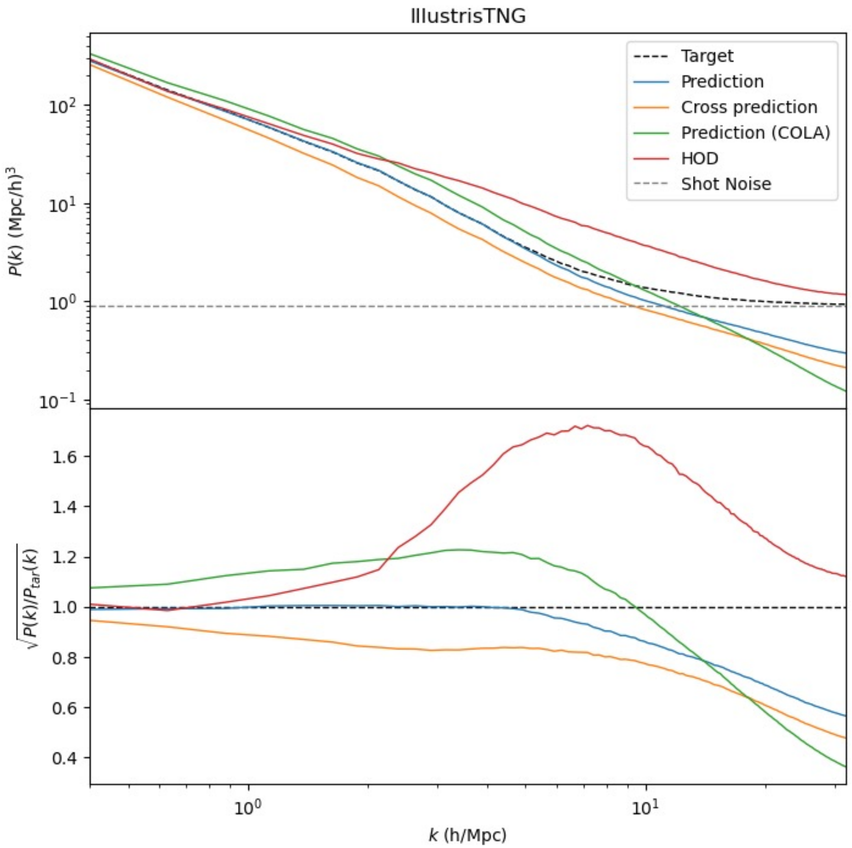
Methods – HOD

- Need benchmark model to compare our model to
- HOD (Halo Occupation Distribution)
- Can range in complexity
- We use a model with 3 free parameters
  - M_{min}, M_1, α
  - Place galaxies in halos with mass $> M_{min}$ with a Poisson distribution with mean $\left(\frac{M}{M_1}\right)^{\alpha}$

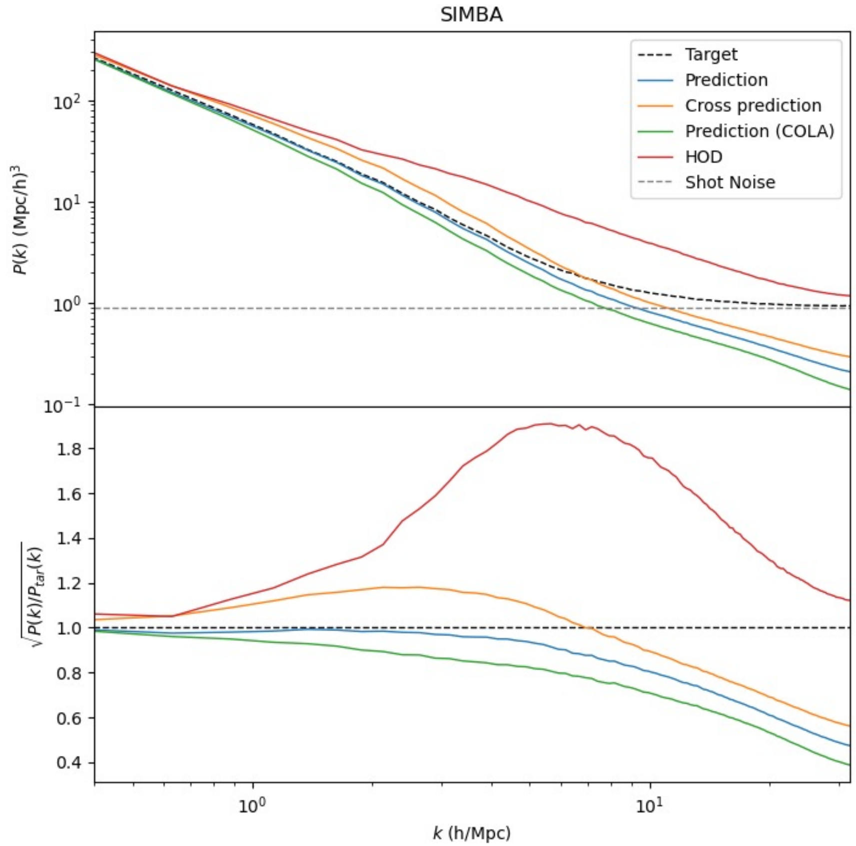
Results – Regression



Results – Regression, Power spectra



Cross prediction = trained on SIMBA



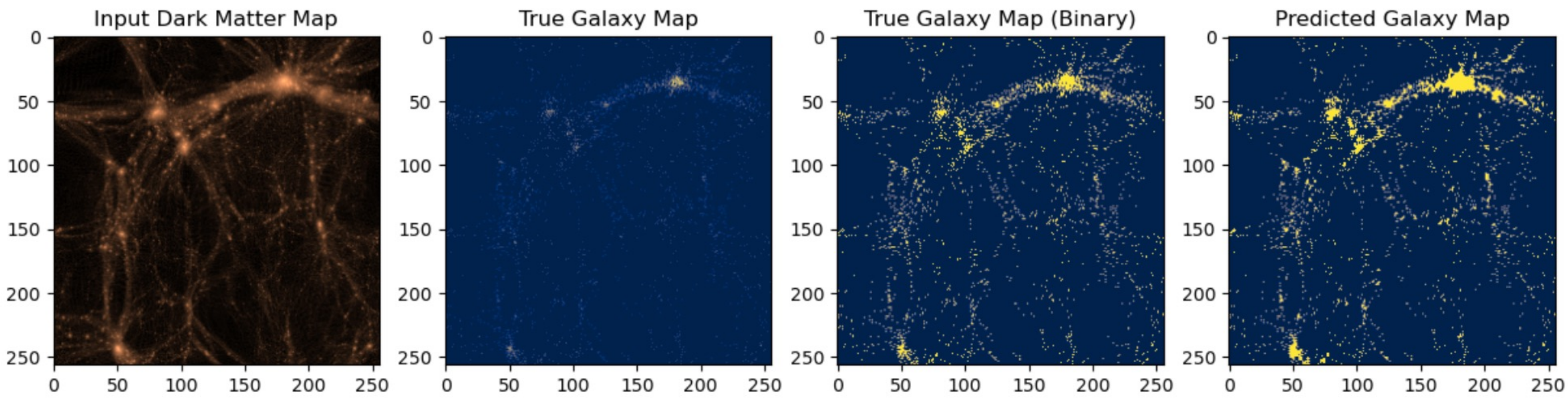
Cross prediction = trained on IllustrisTNG

Variational Diffusion Model (VDM)

- **Goal:** Predict conditional probability distributions of galaxies given DM
- **Forward Diffusion Process:**
  - Noise is progressively added to the galaxy field
  - During training, U-Net learns noise added: (Noisy image, DM, t)
- **Denoising Process:** Reverse process, generate samples
- Loss: VLB of $p_\theta(x_{gal}, x_{DM})$
- Our model works on 2D data, but can be configured for 3D
- Very easy to add parameter conditioning

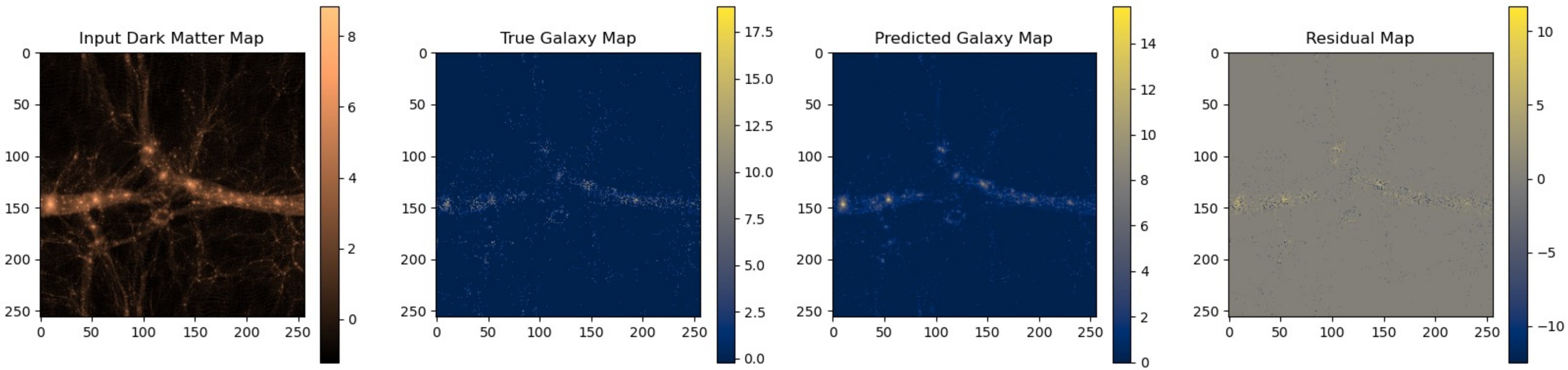
Kingma, D. et al., arXiv:2107.00630

Results – Classification



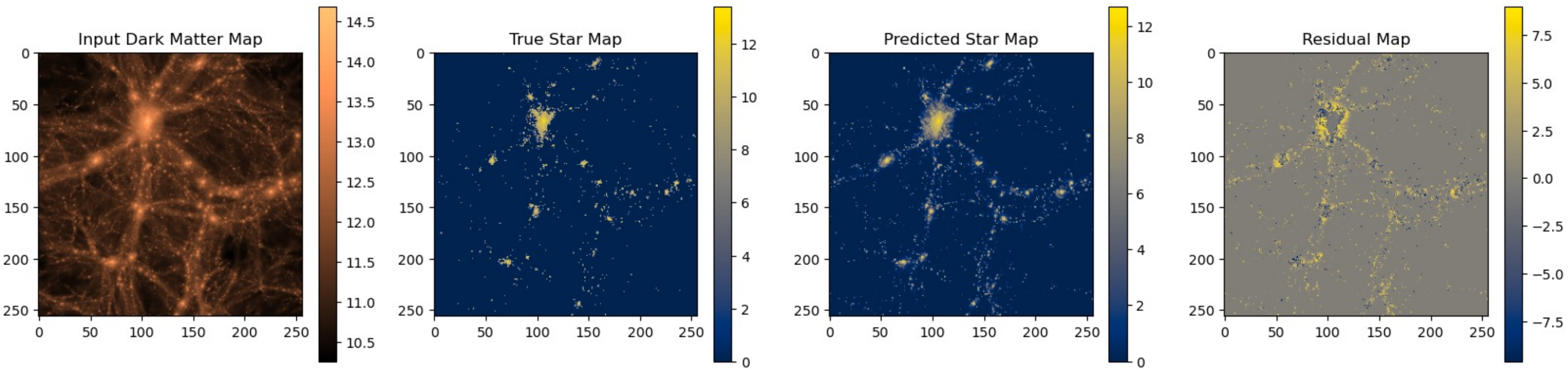
Accuracy: 0.9907
Precision: 0.8370
Recall: 1.0000

Results – Regression



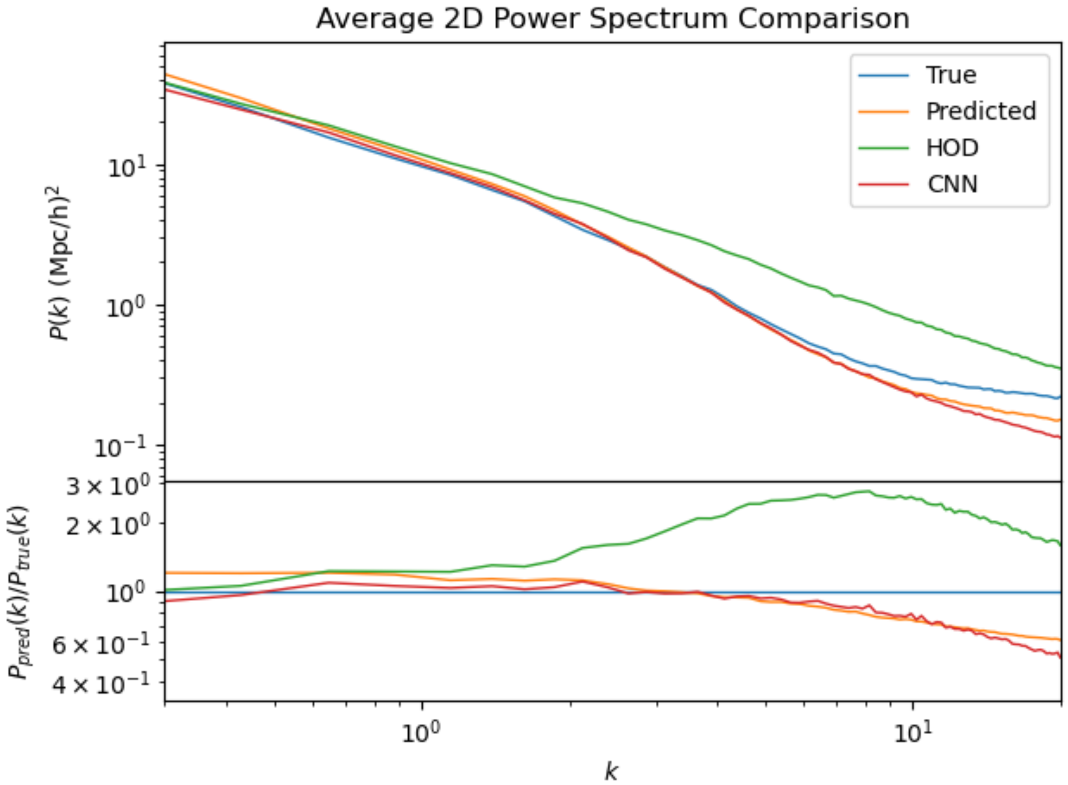
*preliminary results

Results – Regression



**pre-preliminary results

Results – Galaxy ower spectra



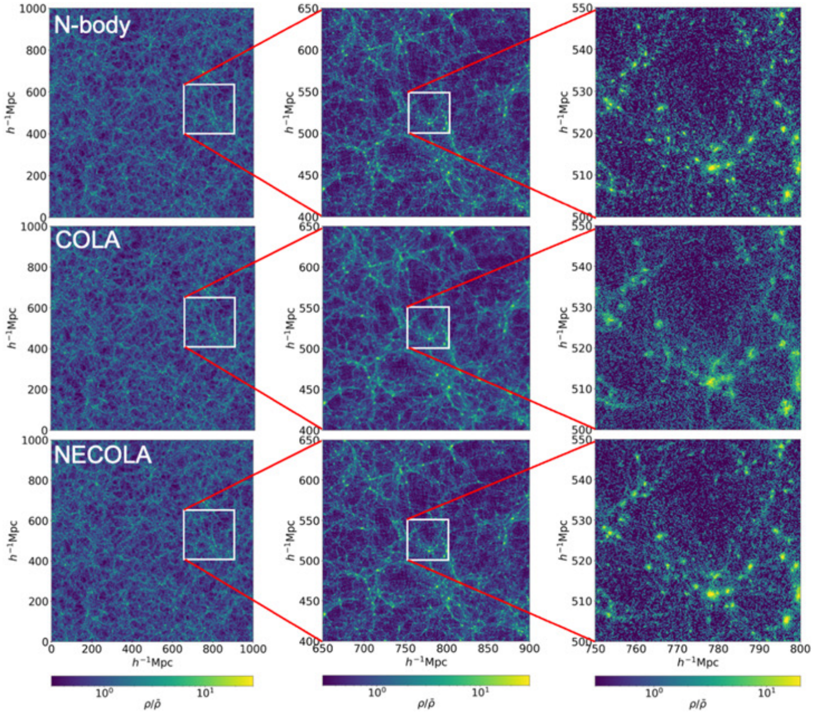
*preliminary results

Conclusions

- Model is able to predict locations of galaxies down to reasonably small scales
  - After training, little computational cost is needed
- Setup is generic and many networks can be used
- Further work is needed to reach smaller scales
- Parameter conditioning is needed to be useful across parameter space

Future Work

- To improve the results and utility of our model, we have a few changes we plan on implementing:
- Use LH data to make more robust
- Expand using other hydro models
- Replace N-Body simulations in all models with COLA/NECOLA
- Train using velocities
- Focus training on halos

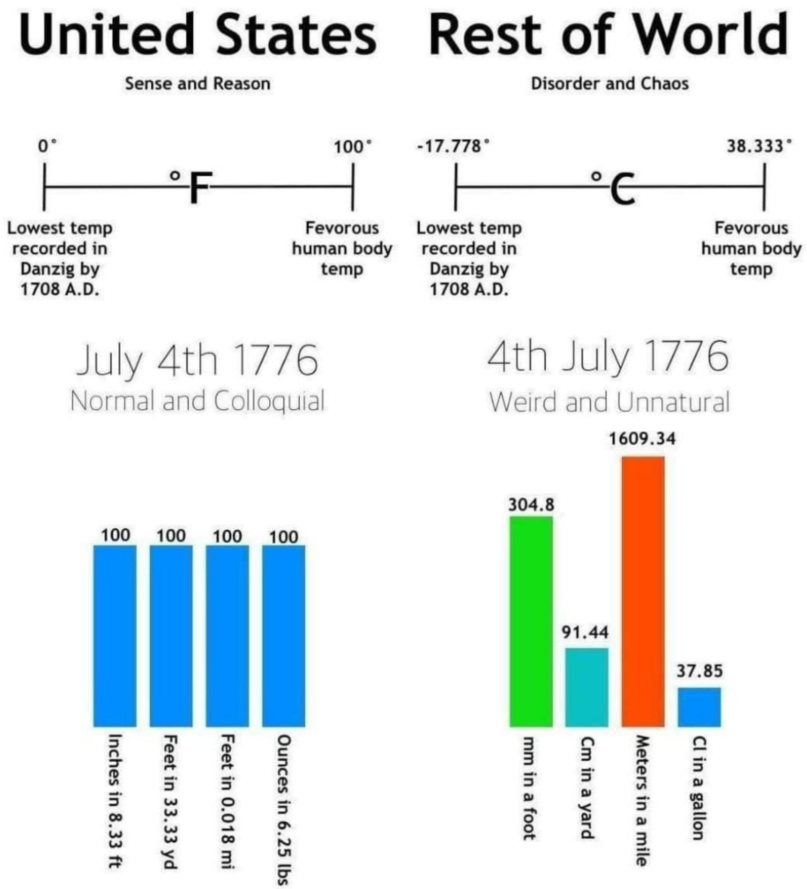


Kaushal et al. arXiv:2111.02441

Acknowledgements

- Elena Giusarma
- Mauricio Reyes Hurtado
- Francisco Villaescusa-Navarro and CAMELS
- Neerav Kaushal
- Michigan Tech Physics Department
- Michigan Tech Graduate Student Government
- Research reported in this publication was supported in part by funding provided by the National Aeronautics and Space Administration (NASA), under award number **80NSSC20M0124**, Michigan Space Grant Consortium (MSGC).

Thank you!

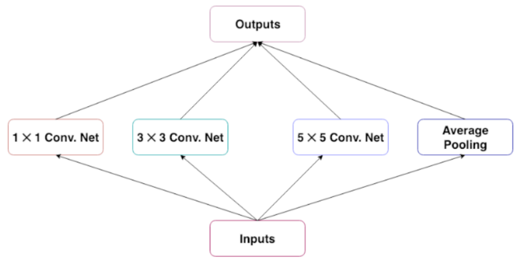


United States
Sense and Reason
0°
100°
°F
Lowest temp recorded in Danzig by 1708 A.D.
Fevorous human body temp
Rest of World
Disorder and Chaos
-17.778°
38.333°
°C
Lowest temp recorded in Danzig by 1708 A.D.
Fevorous human body temp
July 4th 1776
Normal and Colloquial
4th July 1776
Weird and Unnatural
100
100
100
100
Inches in 8.33 ft
Feet in 33.33 yd
Feet in 0.018 mi
Ounces in 6.25 lbs
304.8
91.44
1609.34
37.85
mm in a foot
Cm in a yard
Meters in a mile
Cl in a gallon

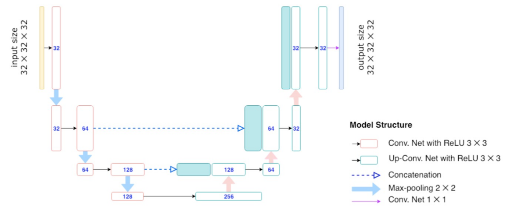
References

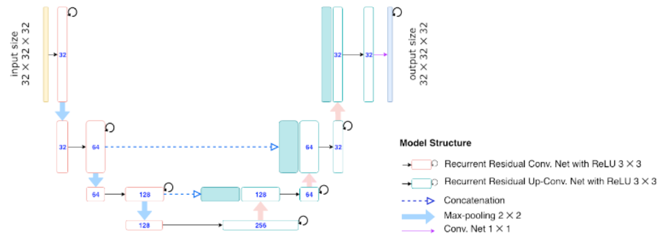
1. Alom, Md. et al, "Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation", arXiv:1802:06955 (2018).
2. Kingma, D. et al, "Variational Diffusion Models", arXiv:2107.00630 (2021)
3. Ono, V. et al, "Debiasing with Diffusion: Probabilistic reconstruction of Dark Matter fields from galaxies with CAMELS", arXiv:2403.10648
4. Tassev, S. et al, "Solving Large Scale Structure in Ten Easy Steps with COLA," *Journal of Cosmology and Astroparticle Physics*, vol. 2013, no. 06, p. 036 (2013).
5. Villaescusa-Navarro F. et al, "The CAMELS project: Cosmology and Astrophysics with MachinE Learning Simulations," arXiv:2010.00619 (2021).
6. Zhang, X. et al, "From Dark Matter to Galaxies with Convolutional Networks," arXiv:1902.05965 (2019).

Methods – Network models



- Inception model
  - CNN with multiple kernel sizes to capture info at select scales
- U-Net
  - CNN that effectively captures spatial relations and context at multiple scales
- R2U-Net
  - Residual Recurrent U-Net
  - Adds residual blocks to "deepen" network
  - Recurrent nets "remember" earlier states
  - Learns special dependencies better





Zhang et al.
arXiv:1902.05965

Data

- We make use of two types of simulations from the CAMELS project:
  - 27 DM-only N-Body simulations
  - 27 hydrodynamic simulations, with DM, gas particles, stars, black holes, etc.
- Each simulation evolves 256^3 DM particles, plus 256^3 gas particles (hydrodynamic only) in a box size of $(25\,h^{-1}\mathrm{Mpc})^3$ from $z = 127$ to $z = 0$.
  - For the N-Body sims, we create a 3D field of the counts of particles within each of 256^3 voxels
  - For the hydrodynamic sims, we create a similar field, but instead with counts of galaxies