From Dark Matter to Galaxies with Convolutional Neural Networks

Jacky H. T. Yip

CUHK Physics Student Conference

21st September 2019

advisor: Shirley Ho

FLATIRON INSTITUTE

CUHK 香港中文大學

NYU Center for Data Science

Today’s outline

- Motivation: Why the mapping; why convolutional networks?
- Method
  - Data: Problem with sparsity
  - Cascade model
- Result
- Conclusion and future work

Motivation

Cosmologists who are interested in galaxies

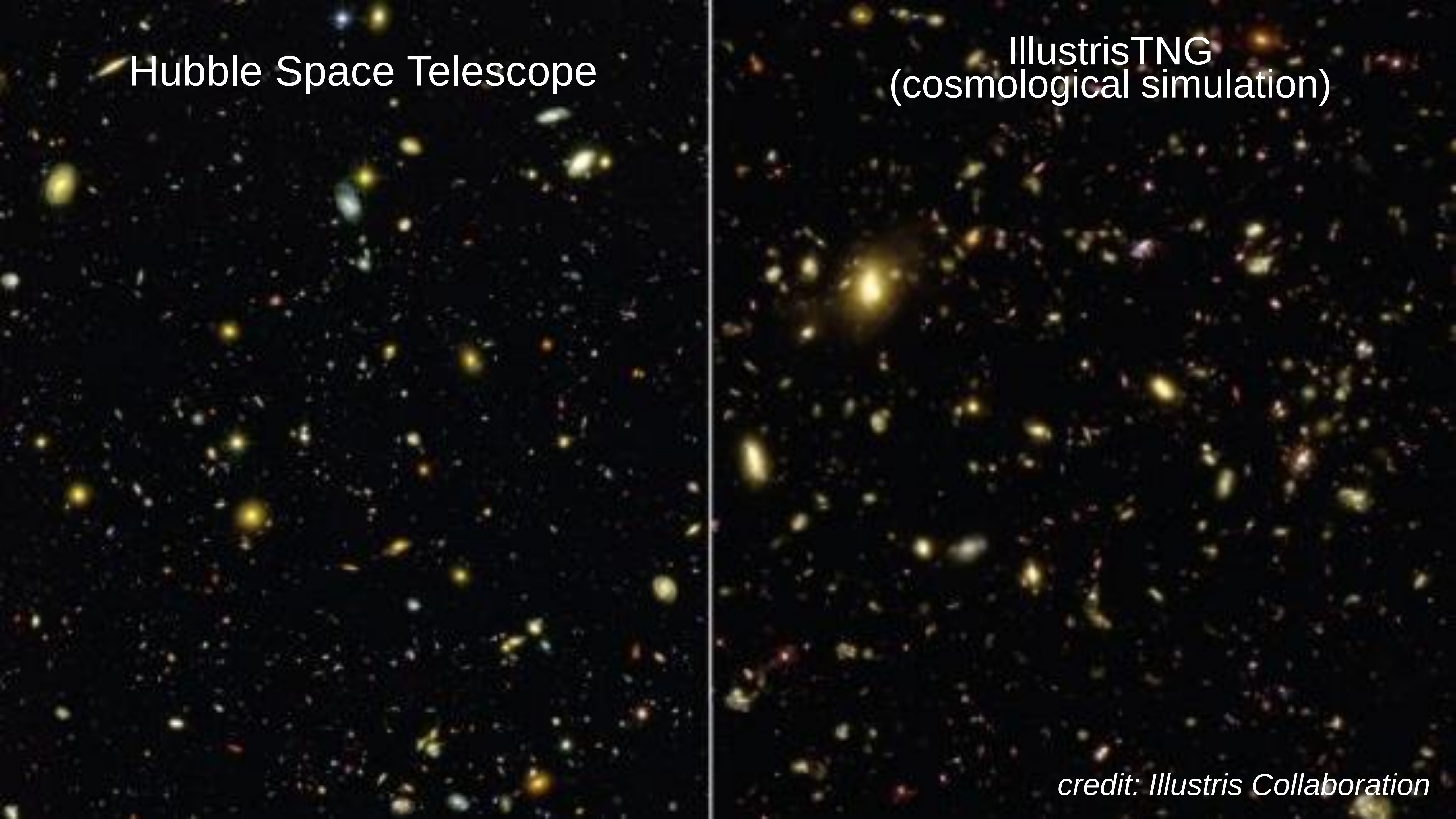
Observed data

compare ⟷

Theory predictions

from cosmological surveys
e.g. LSST, SDSS

from cosmological simulations
e.g. IllustrisTNG

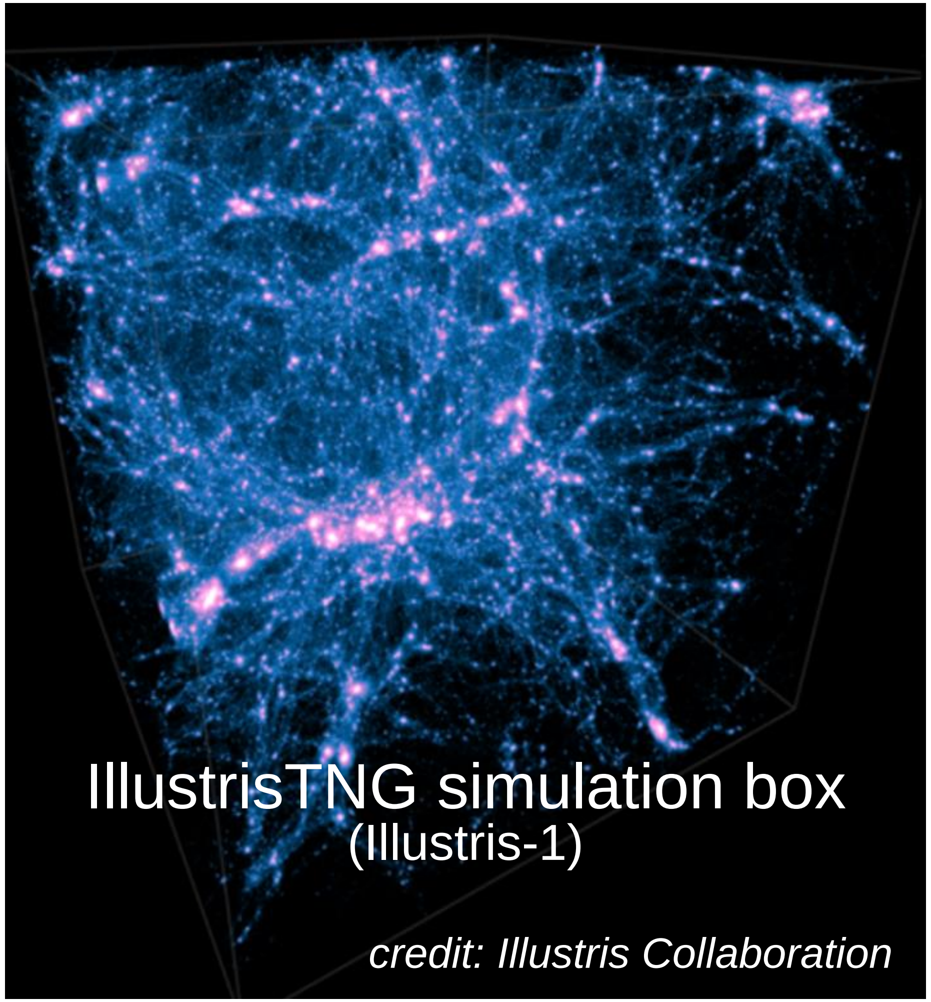

Hubble Space Telescope
IllustrisTNG
(cosmological simulation)
credit: Illustris Collaboration

Motivation – What are these simulations?

For state-of-the-art IllustrisTNG simulations:

- Evolving particles from soon after the big bang (z=127) until present day
  - ~10^{10} particles (dark matter, baryons)
  - Volume of (~10^2 Mpc/h)^3
  - Full physics: gravity, magneto-hydrodynamics…

- **Extremely computationally expensive!**
  - 19 million CPU hours (Illustris-1)



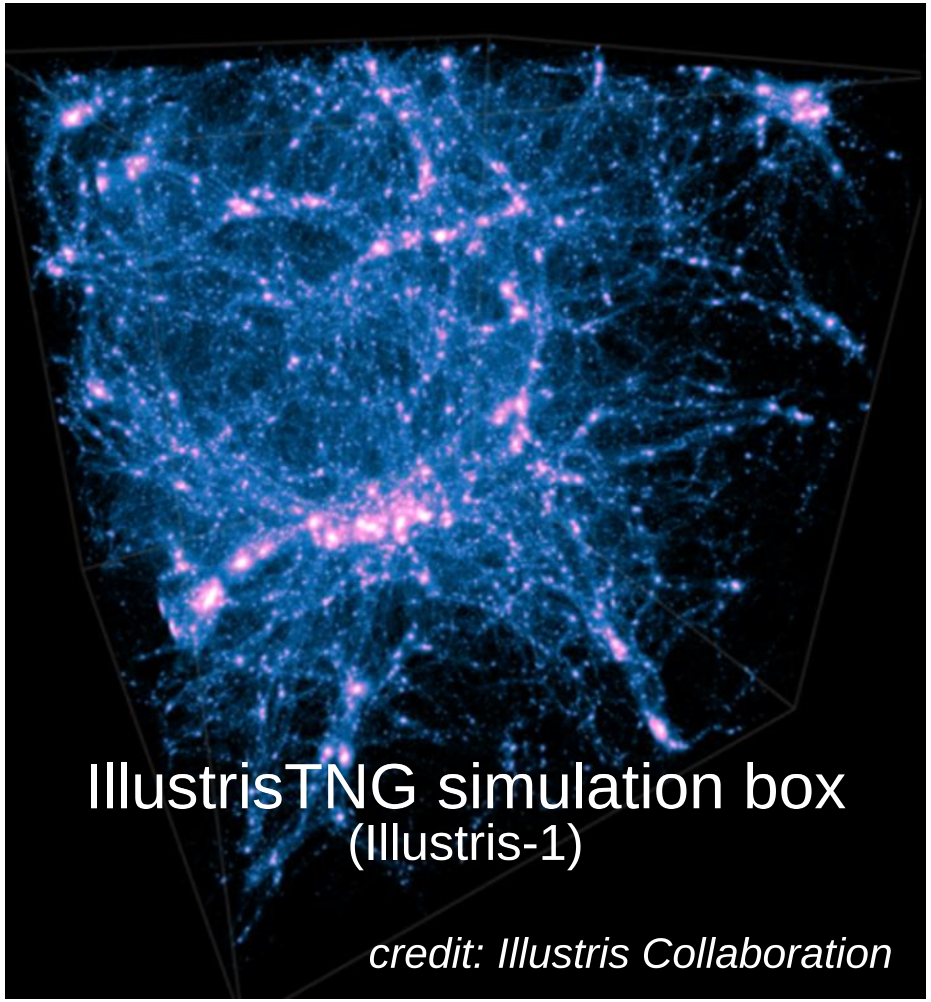
IllustrisTNG simulation box
(Illustris-1)

credit: Illustris Collaboration

Motivation – What are these simulations?

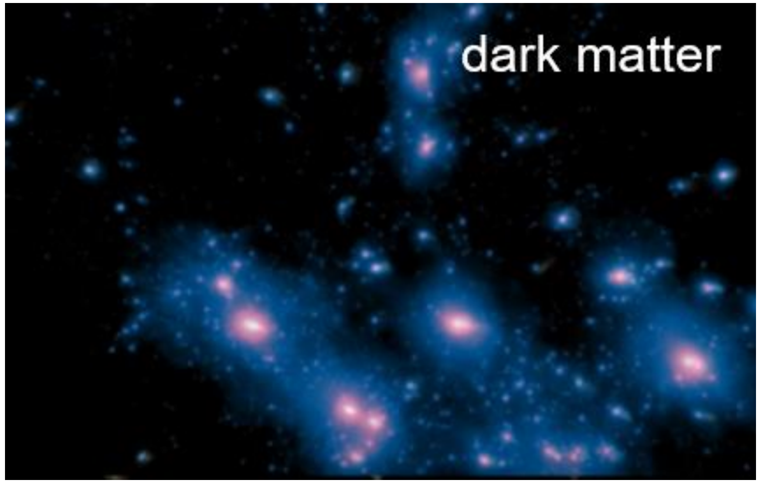
For state-of-the-art IllustrisTNG simulations:

- Instead of full-hydro sims with all matters, consider dark-matter-only sims
- Gravity-only N-body sims
  - **Computationally cheaper!**



Motivation – What if?

Input

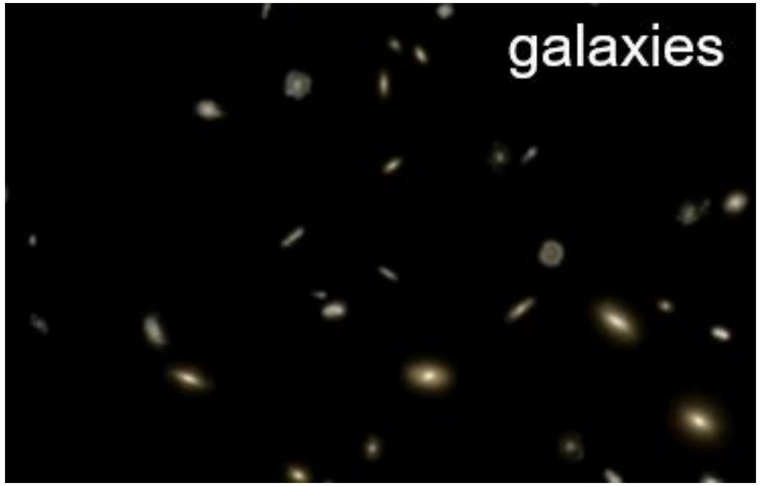


from dark-matter-only N-body sims (cheaper)

Machine learning

Convolutional neural networks (CNNs)

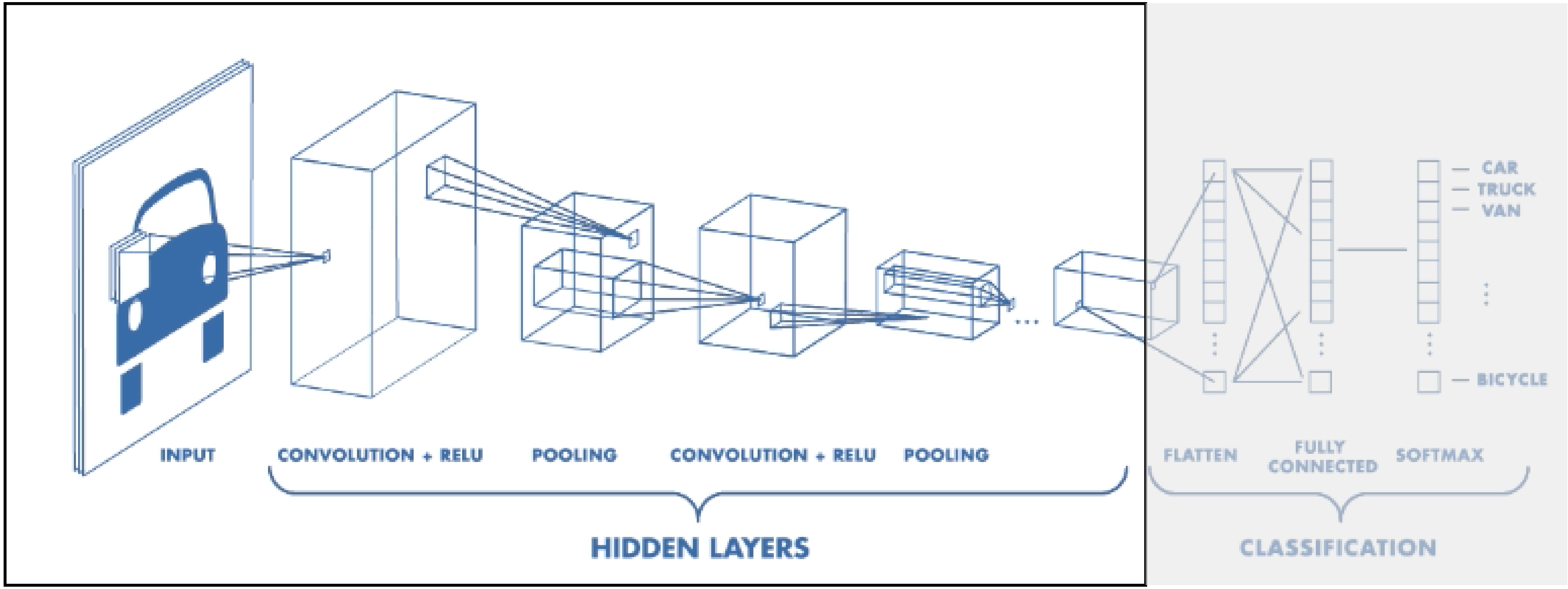
Target



from full-hydrodynamical sims (very expensive)

credit: Illustris Collaboration

Motivation – Why CNNs?



credit: towardsdatascience.com/basics-of-the-classic-cnn-a3dce1225add

Motivation – Why CNNs?

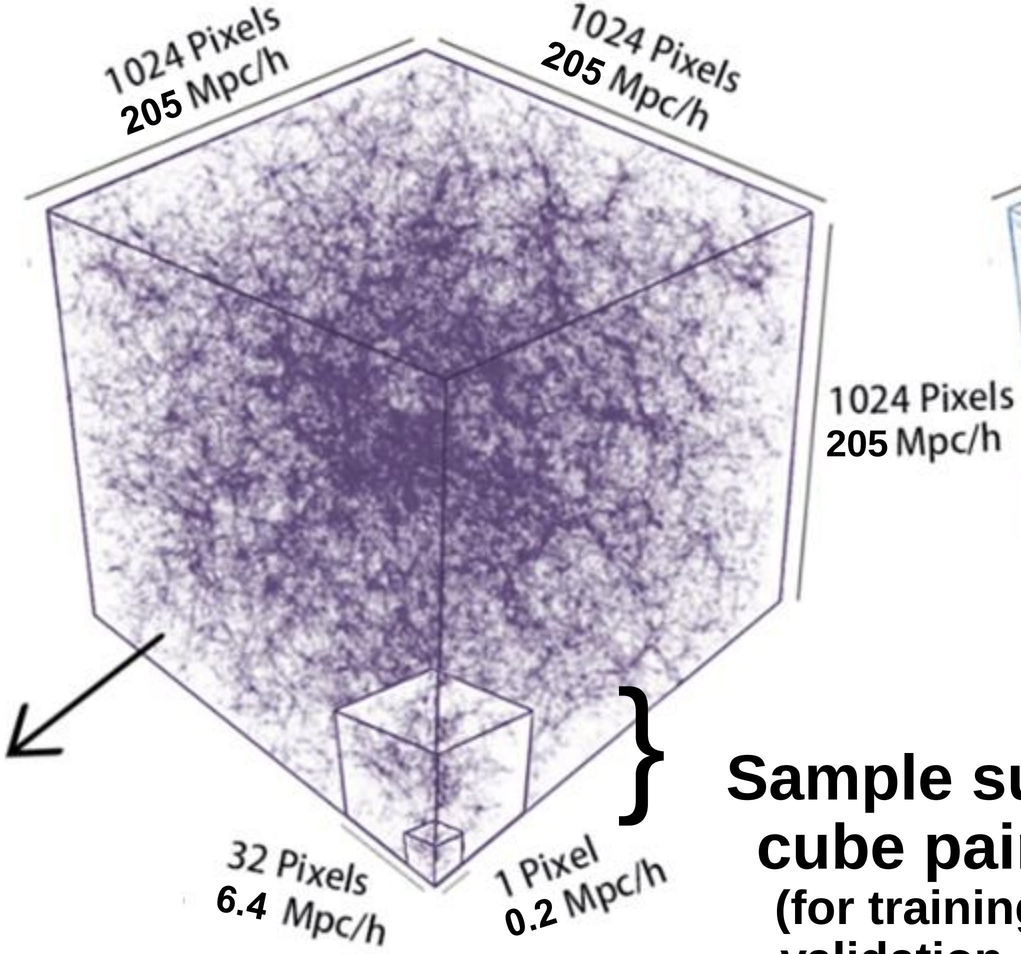
- CNNs are used in image classification and segmentation
  - Capture local information
  - Translational invariant
- Galaxies form as baryonic matter collapses in dark matter halos
  - Expect to depend on local properties of the dark matter halos
  - Independent on the galaxy’s absolute position

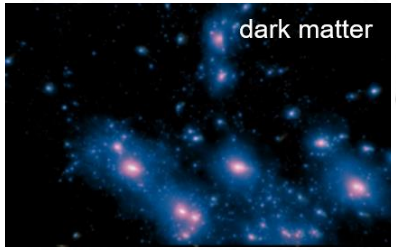
Today’s outline

- Motivation: Why the mapping; why convolutional networks?
- **Method**
  - Data: Problem with sparsity
  - Cascade model
- Result
- Conclusion and future work

Method – Data

Input

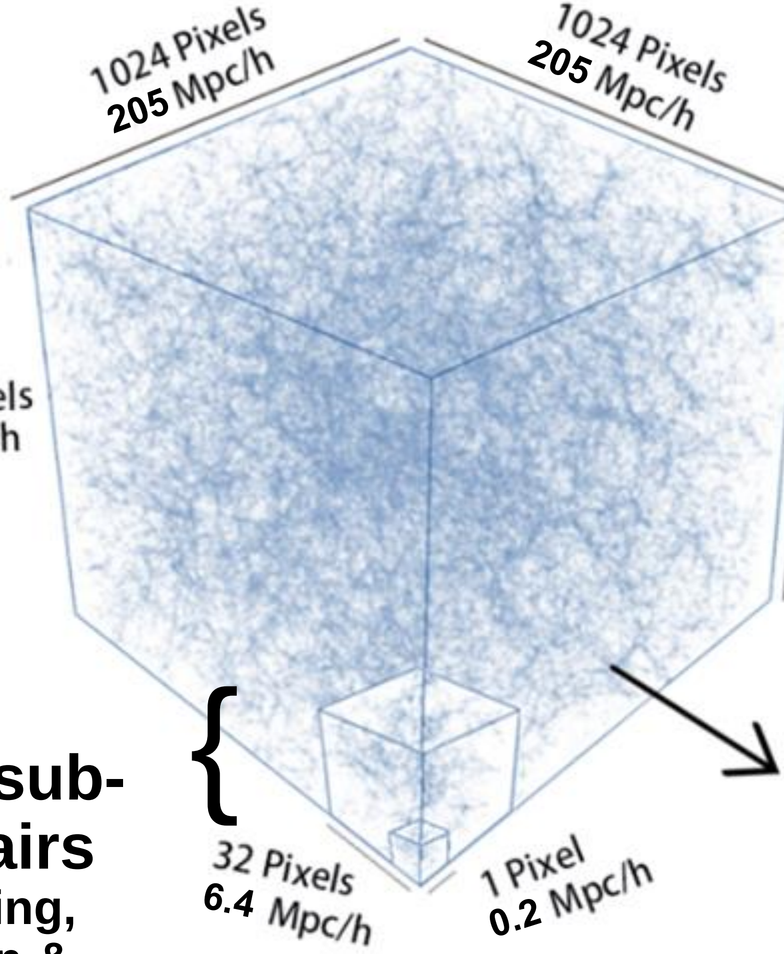


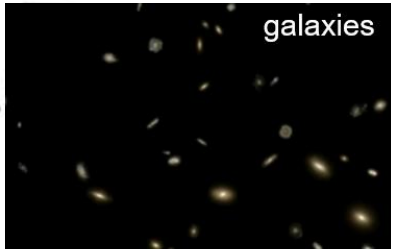


Dark matter mass density

Sample sub-cube pairs
(for training, validation & testing)

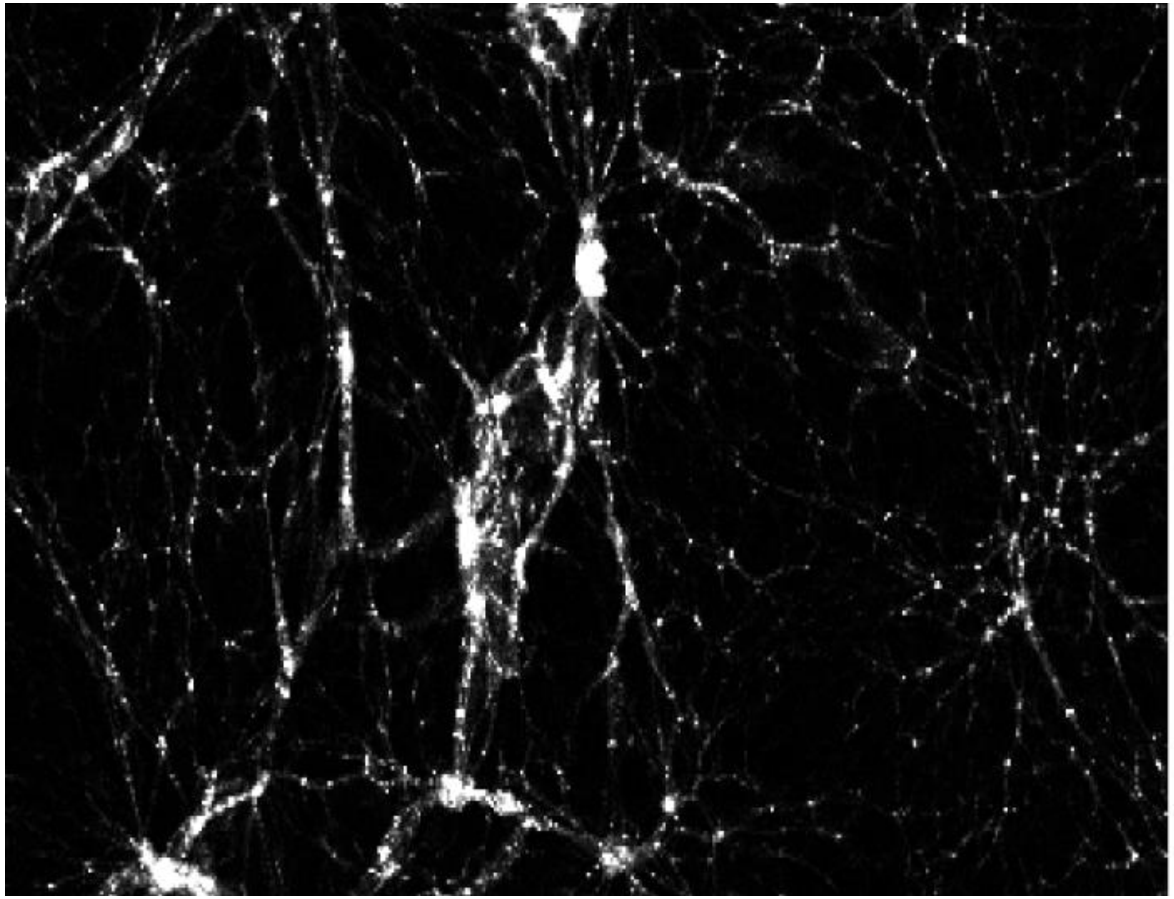
Target





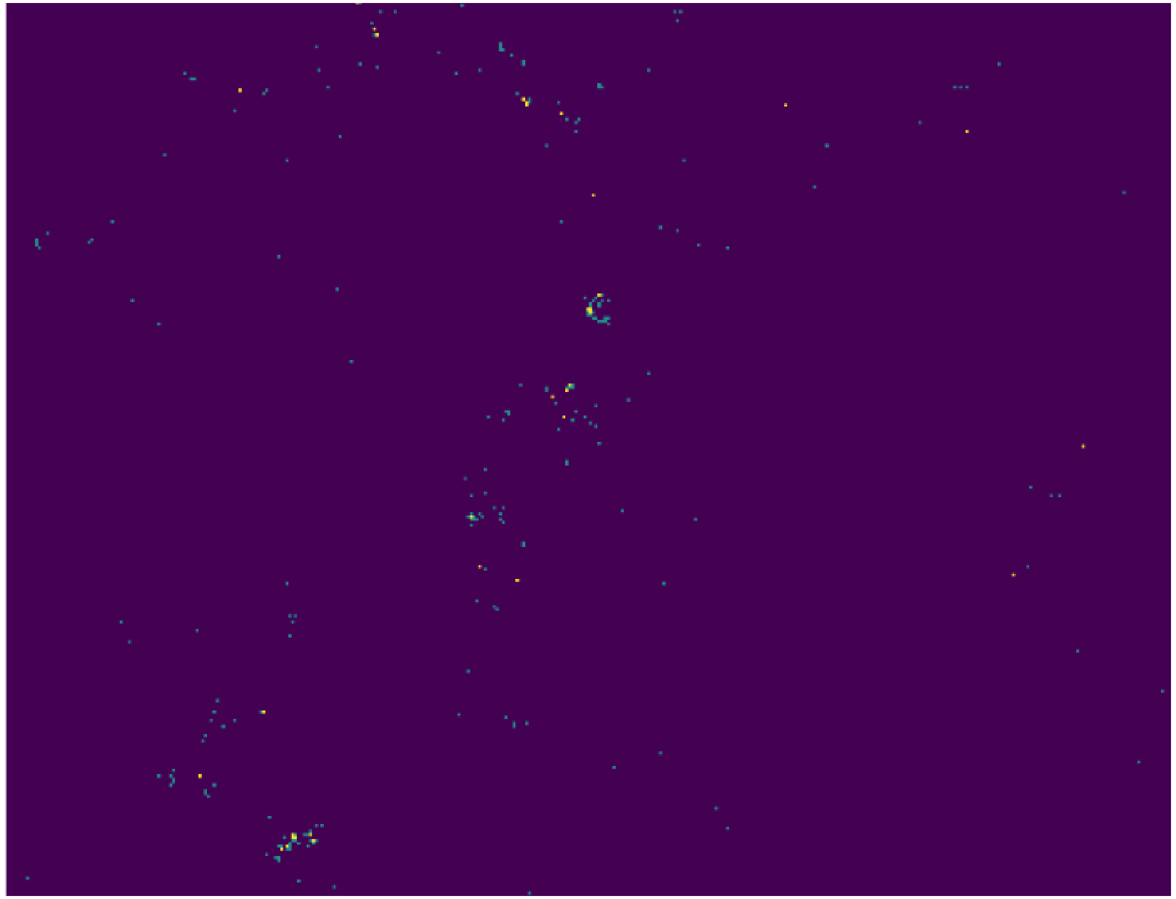
Galaxy number density

Input



Dark matter mass density

Target



Galaxy number density

Target galaxy number density

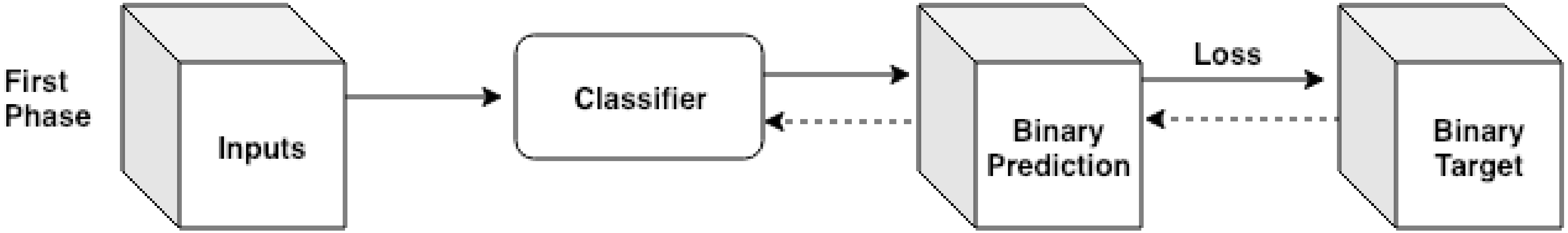
- **Highly sparse:** only **0.15%** of voxels have galaxies
- **Problem: Hard to regress; predicting zero galaxy in all voxels still results in 99.85% accuracy!**

→ **Cascade model**

Method – Cascade model

Cascade model of two phases:

- **1st phase: Binary classification**
  - *Inception Network*
  - **Classify whether a voxel has a galaxy or not (empty or not)**
  - **Successfully reduces the sparsity by ~50 times effectively**

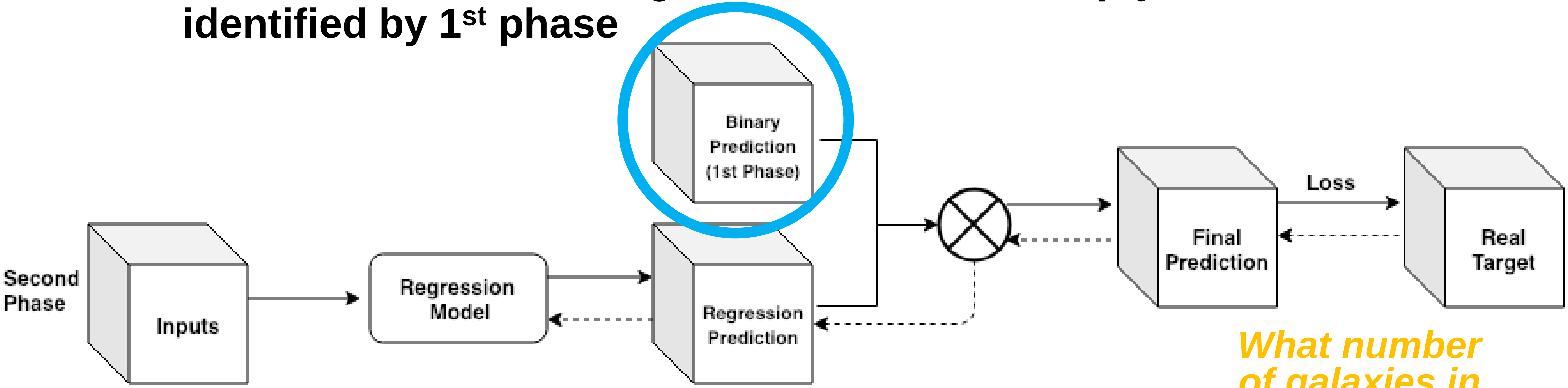


Voxel empty or not?

Method – Cascade model

Cascade model of two phases:

- **2^{nd} phase: Regression model** **(masked by 1^{st} phase's prediction)**
  - *Recurrent Residual U-net*
  - **Predict the numbers of galaxies in the non-empty voxels which were identified by 1^{st} phase**



What number of galaxies in voxel?

Today's outline

- Motivation: Why the mapping; why convolutional networks?
- Method
  - Data: Problem with sparsity
  - Cascade model
- **Result**
- Conclusion and future work

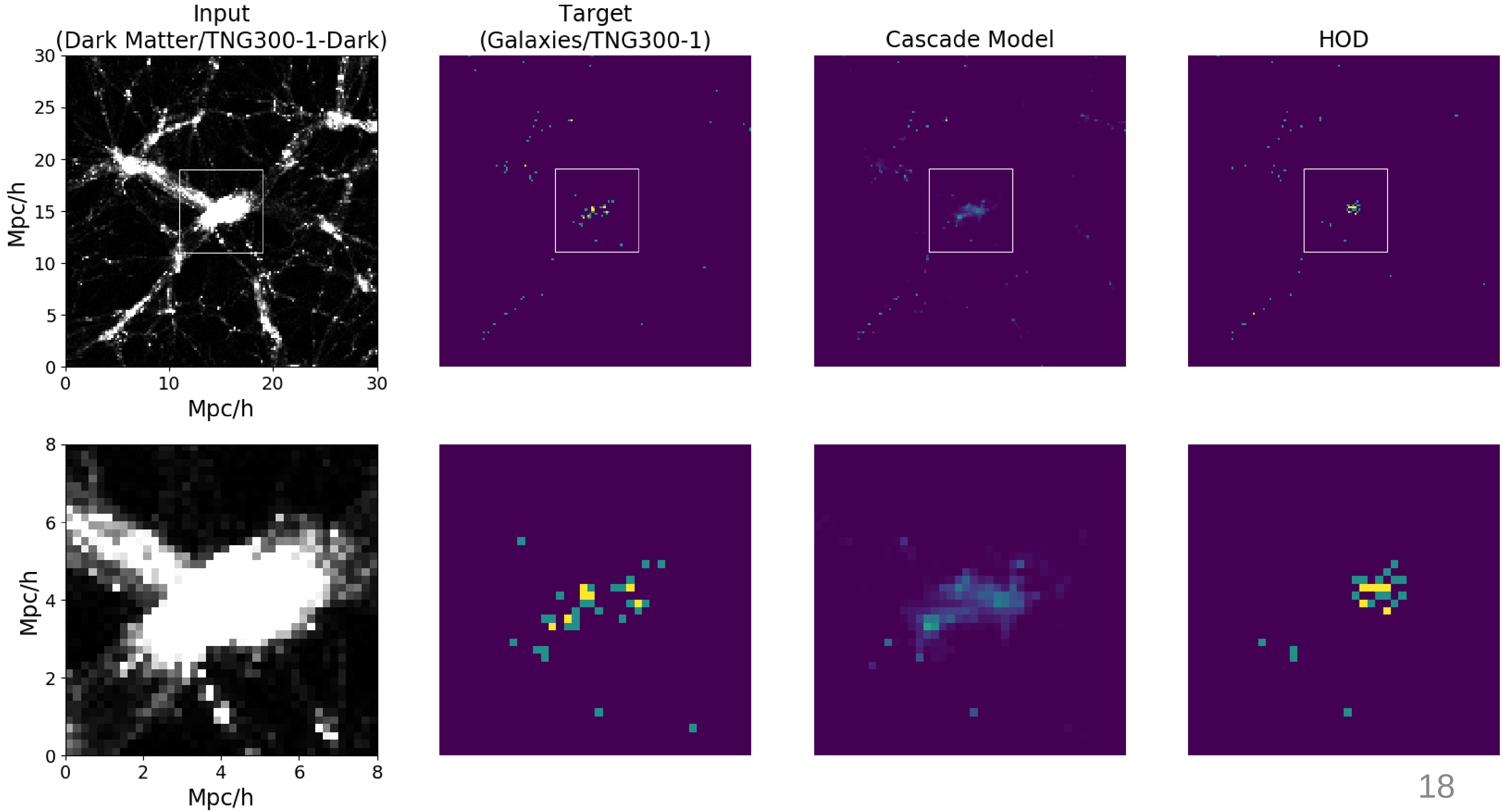
Result – Benchmark model: HOD

Halo Occupation Distribution (HOD) model

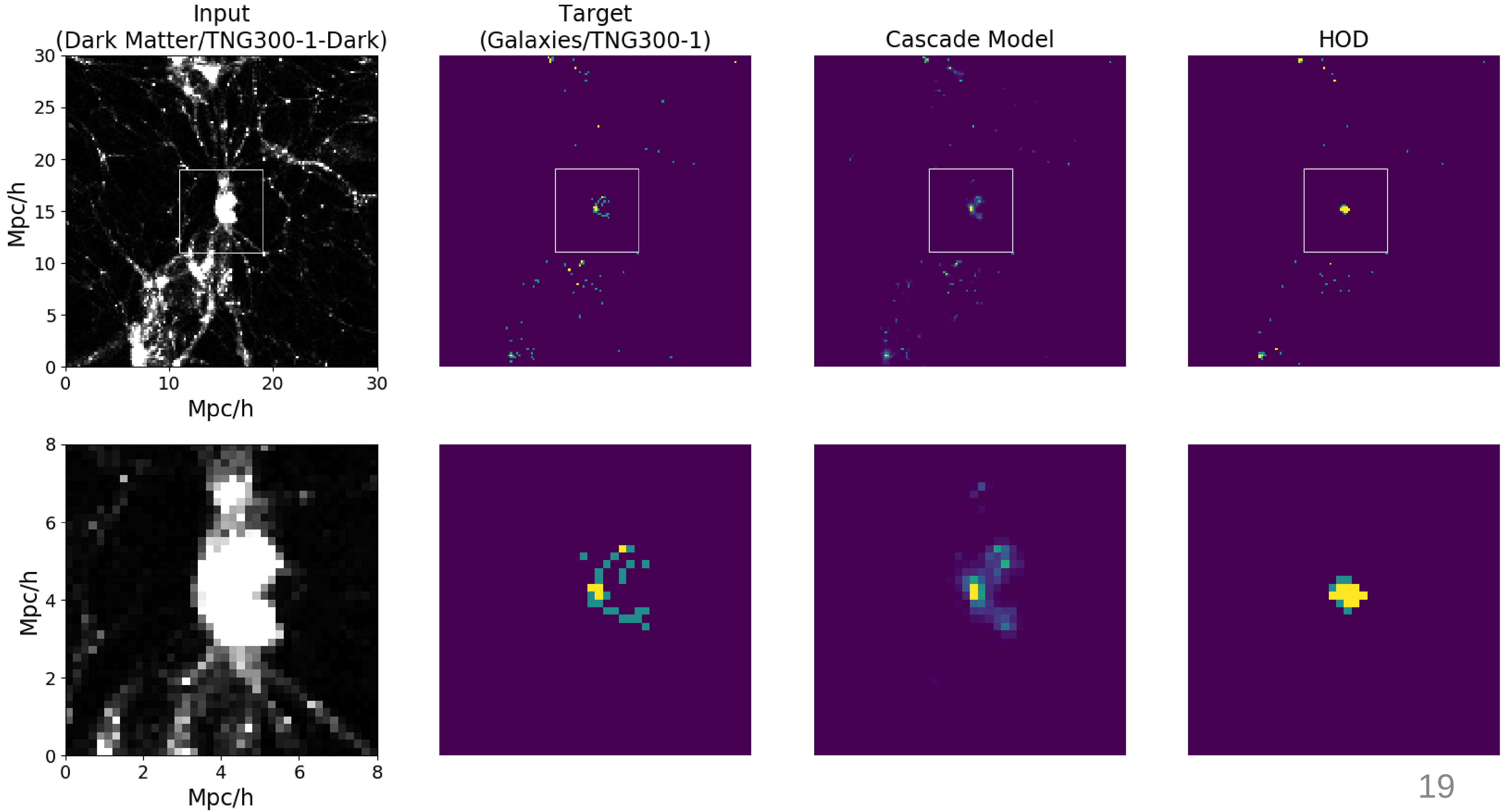
- Commonly used in the cosmological community to link dark matter halos and galaxies

- Identifies halo-mass-related-only parameters to determine the number of galaxies that a dark matter halo holds, then randomly place them within a radius inside the halo

Result – Visualizations



Result – Visualizations



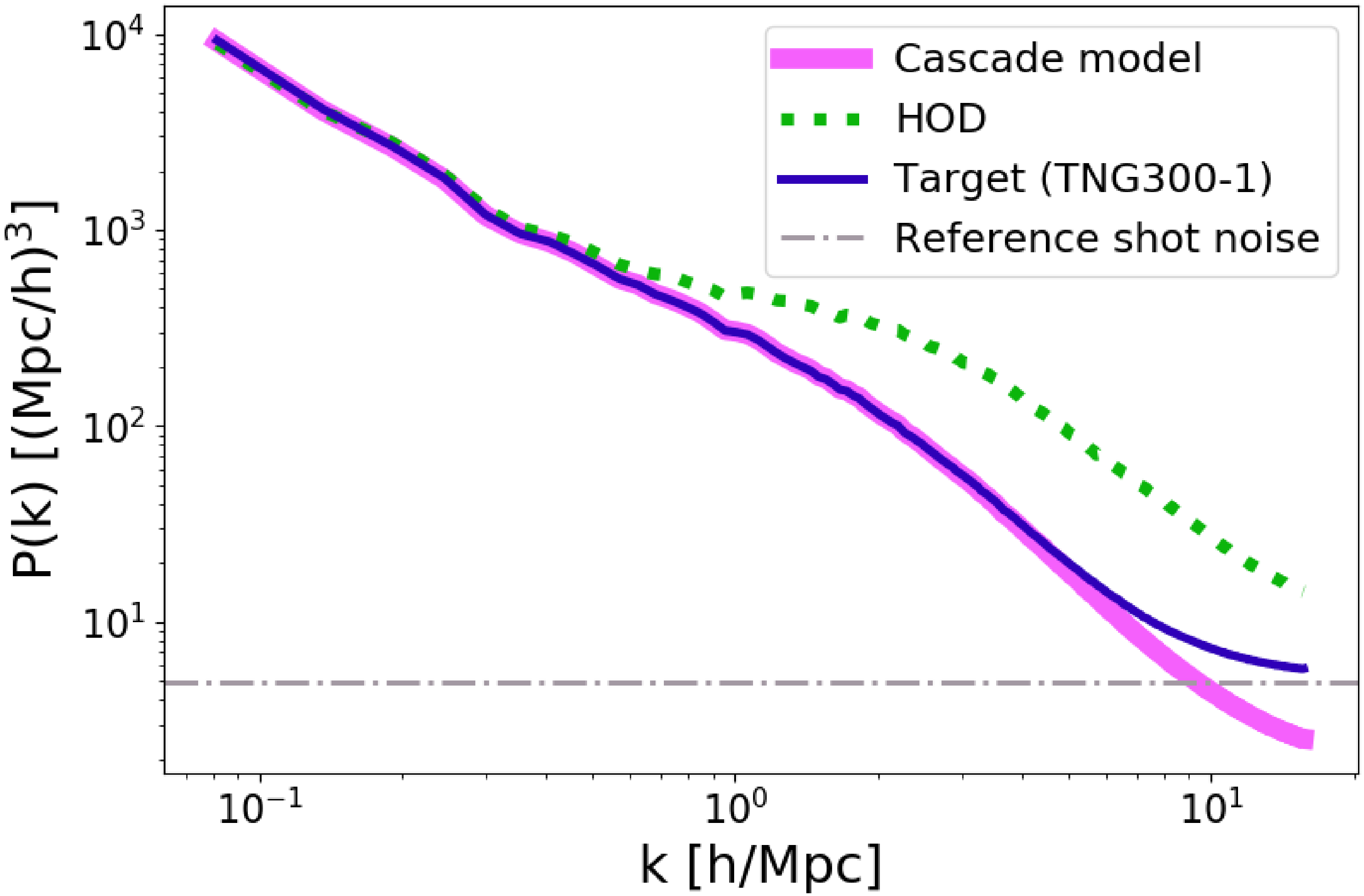
Result – Power spectra

- Fourier transform of the correlation function:

$$\xi(|\boldsymbol{r}|) = \langle \delta_A(\boldsymbol{r}')\delta_B(\boldsymbol{r}' + \boldsymbol{r})\rangle$$

$$P(|\boldsymbol{k}|) = \int \mathrm{d}^3\boldsymbol{r}\ \xi(r) e^{i\boldsymbol{k}\cdot\boldsymbol{r}}$$

- Accounts for the deviation of the galaxy distribution from a random field at different scale k



Conclusion and future work

- **Our cascade model can efficiently predict the galaxy number density field given only the dark matter density field**
  - Outperforms benchmark method in statistics of interest

- Extensions / on-going work
  - Predict additional properties of galaxies: mass, galaxy formation rate, metallicity… etc
  - Explore more environmental factors in the dark matter field: velocity… etc

Thank you,
CUHK Physics & Shirley!

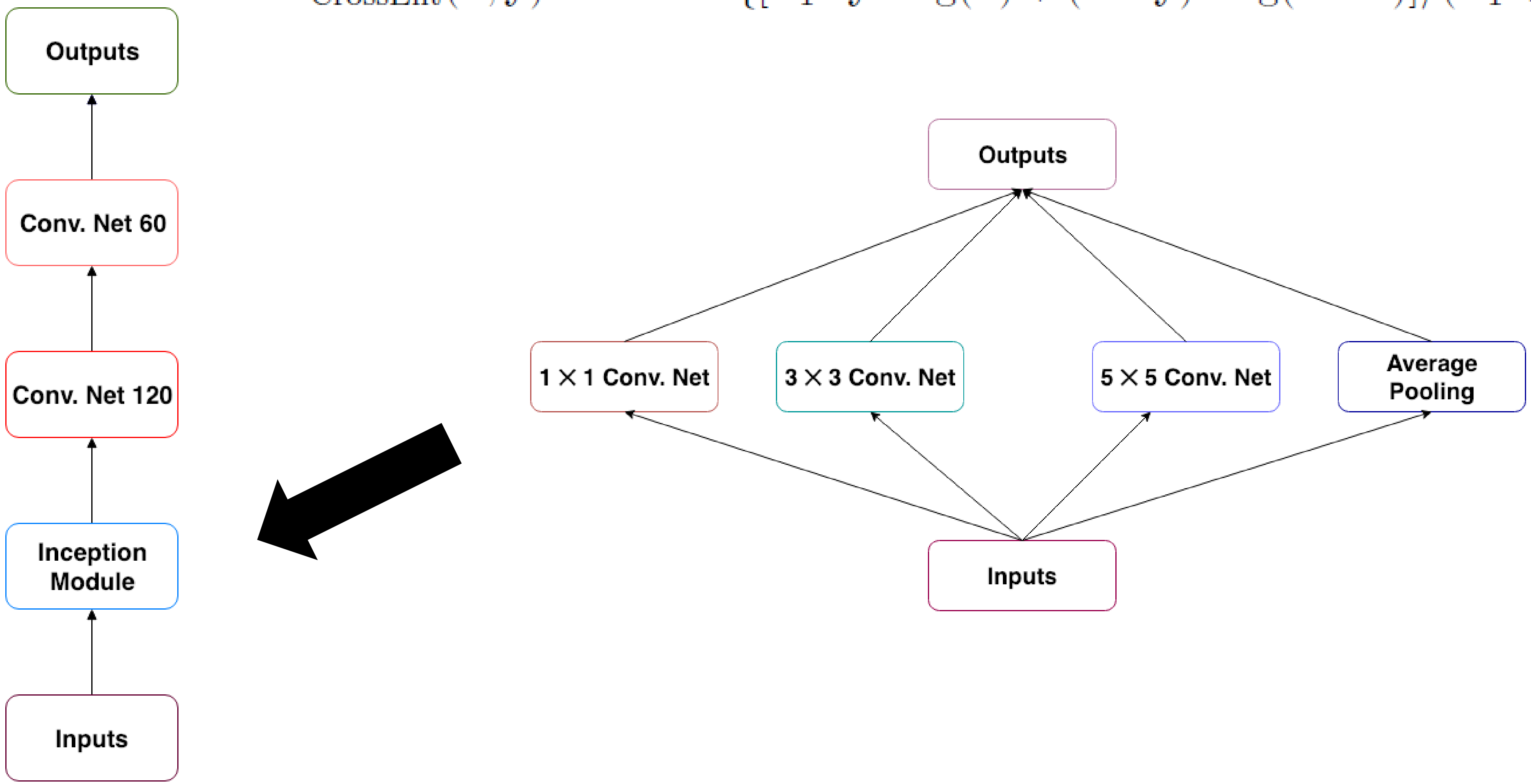
THE CHINESE UNIVERSITY OF HONG KONG

Department of Physics

Summer Undergraduate Research Exchange Program (SURE)

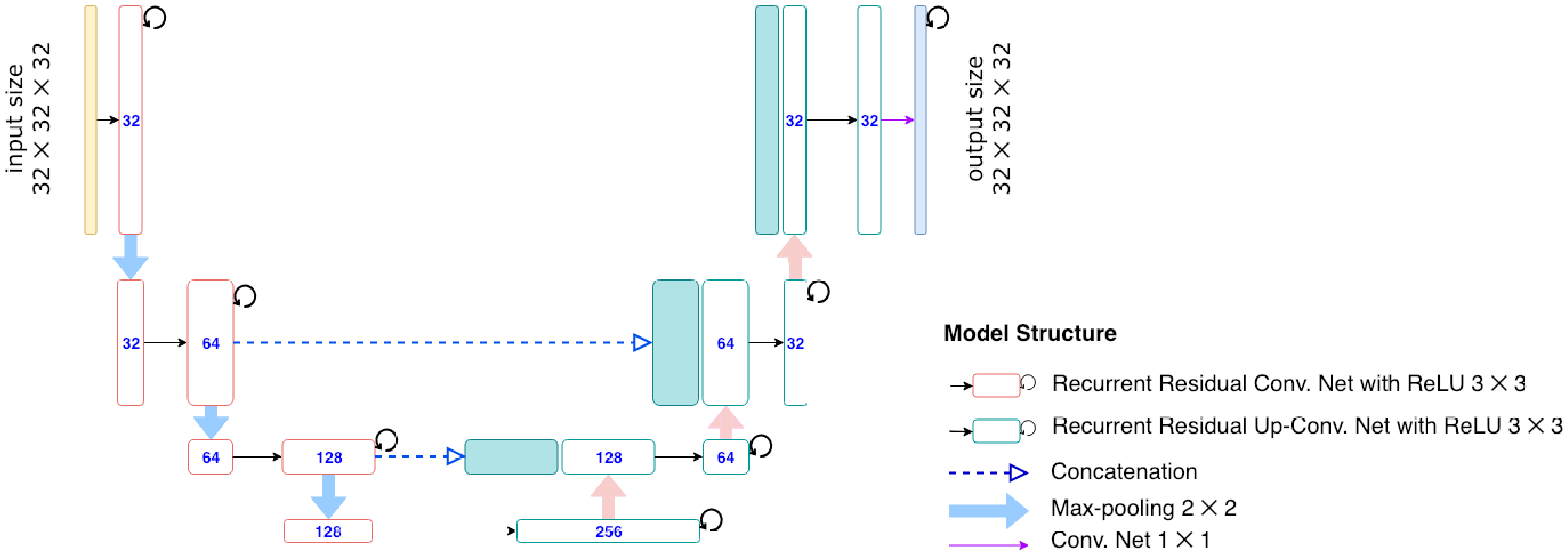
First phase: Inception Network

$$\mathbb{L}_{\text{CrossEnt}}(\mathbf{x},\mathbf{y}) = -mean\{[w_1 \cdot \mathbf{y} \cdot \log(\mathbf{x}) + (1-\mathbf{y}) \cdot \log(1-\mathbf{x})]/(w_1+1)\}$$



Second phase: Recurrent Residual U-net

$$\mathbb{L}_{\mathrm{MSE}}(\mathbf{n_g}, \mathbf{n_t}) = mean\{W(\mathbf{n_t})(\mathbf{n_g} - \mathbf{n_t})^2\}$$



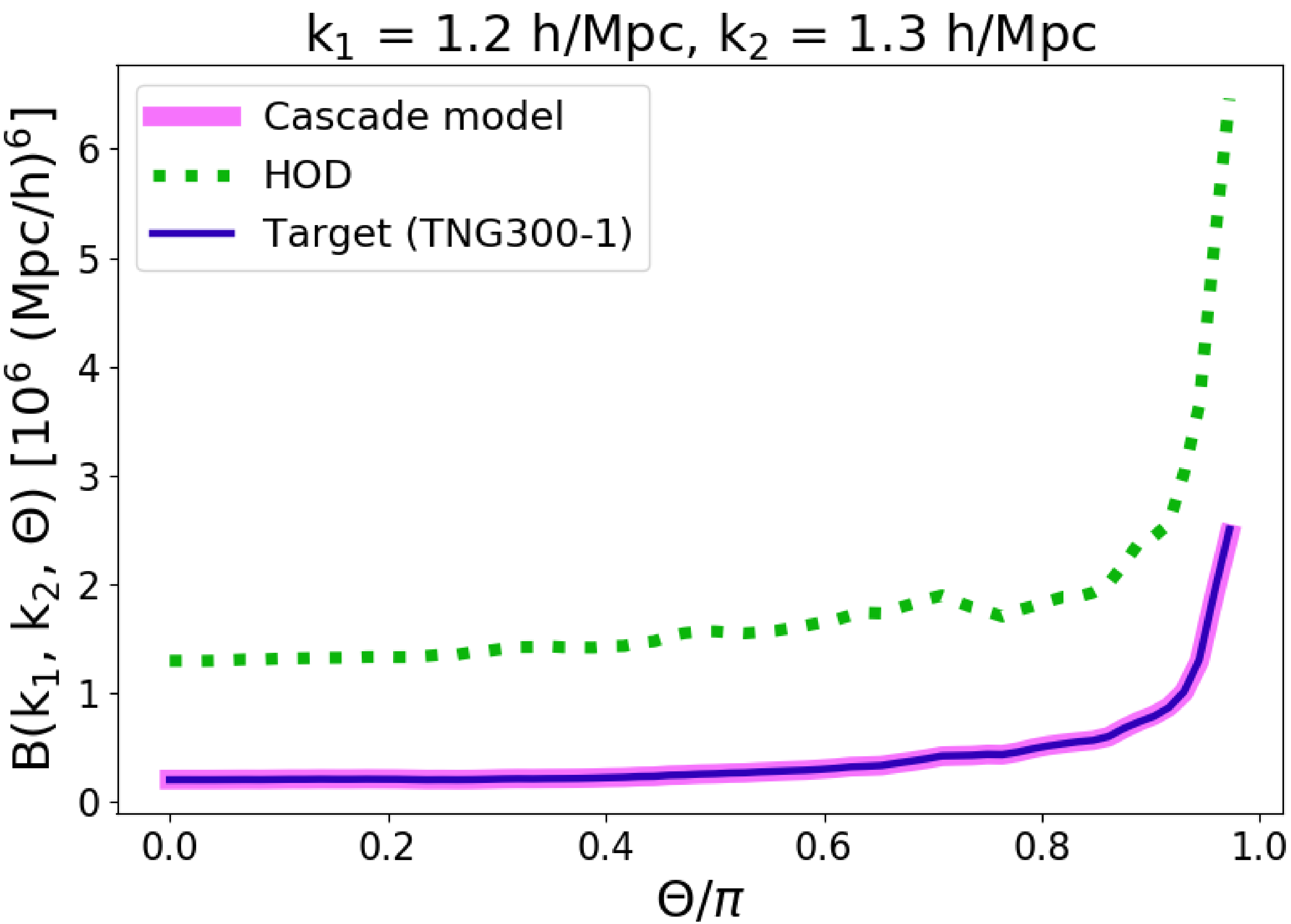
Halo Occupation Distribution (HOD) Model

1) Read all the positions and radii of the halos

2) If it has a central galaxy (if $M > M_{min}$), put it in the center of the halo

3) Compute the mean number of satellites in the halo as $(M/M1)^{alpha}$

4) Given that number, draw a random number with a Poissonian distribution with that mean. That will be the number of satellites of that particular halo. Place them randomly within the radius of the halo



k1 = 1.2 h/Mpc, k2 = 1.3 h/Mpc
Cascade model
HOD
Target (TNG300-1)
B(k1, k2, Θ) [10^6 (Mpc/h)^6]
0
1
2
3
4
5
6
0.0
0.2
0.4
0.6
0.8
1.0
Θ/π