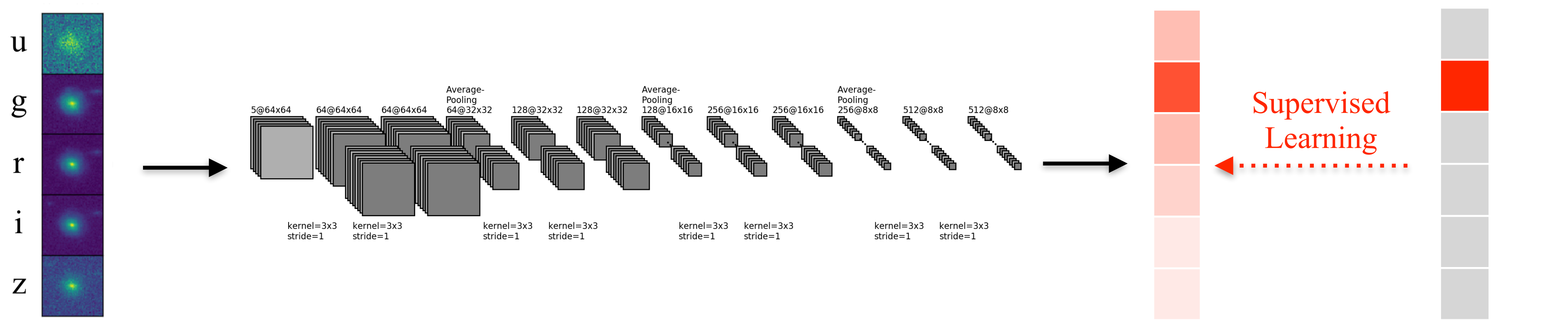
Estimation of photo-z probability density functions via deep learning with statistical basis

Presenter: Qiufan Lin (林秋帆)

Photo-z estimation as a computer vision problem supervised by spec-z

- ✓ State-of-the-art, best accuracy
- ✗ Direct PDF prediction lacks statistical basis, and may suffer from biases
- ✗ The network lacks interpretability

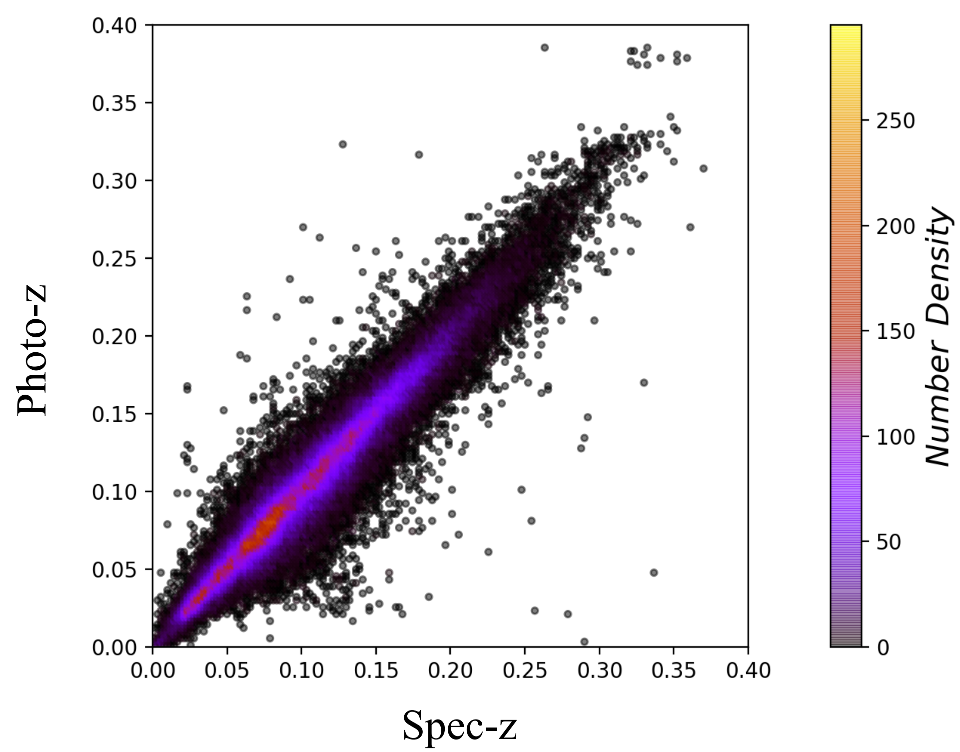


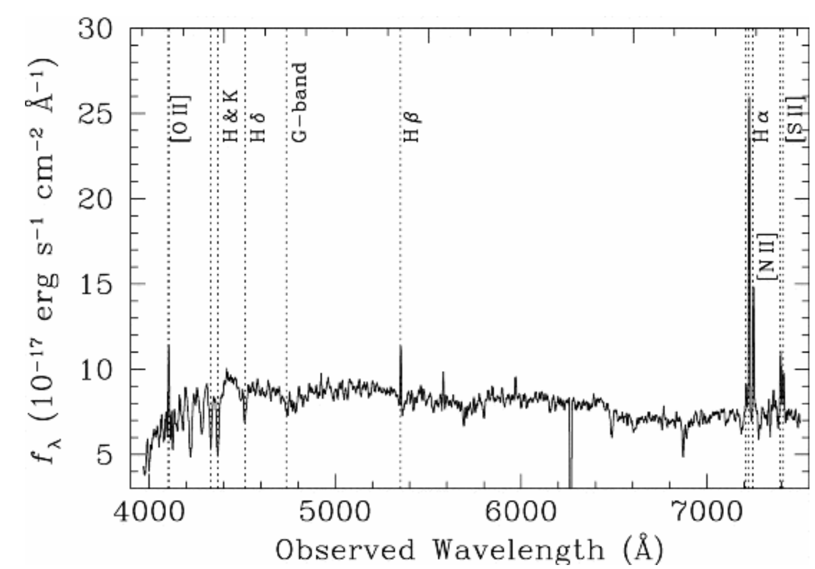
Input Image

Network

“PDF” Output

Label (spec-z)





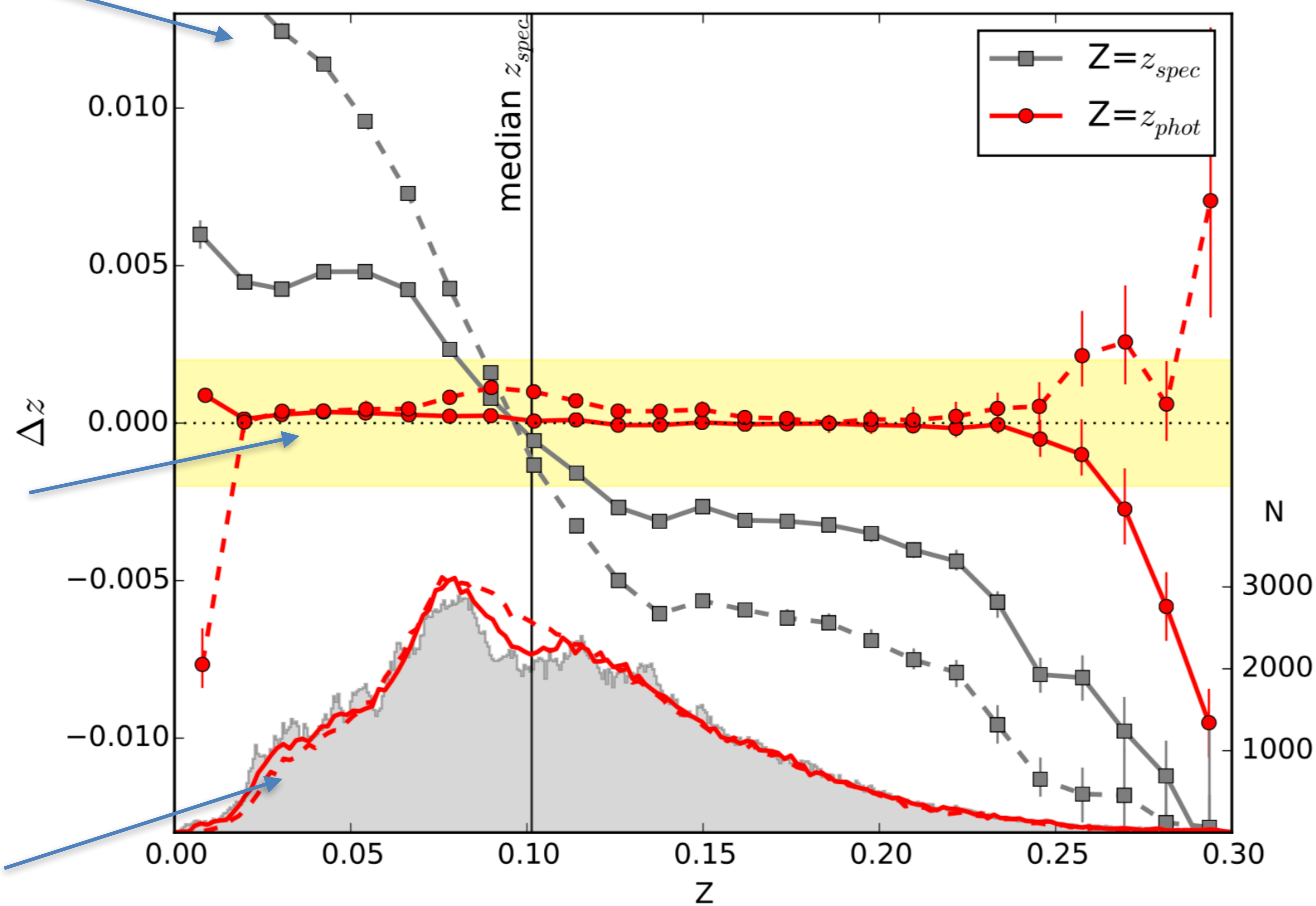
Spectroscopy

One form of bias:
mean redshift residuals as a function of spec-z or photo-z

$$\Delta z = \frac{z_{photo} - z_{spec}}{1 + z_{spec}}$$

w.r.t spec-z
(biased)

w.r.t photo-z
(not guaranteed
to be unbiased)



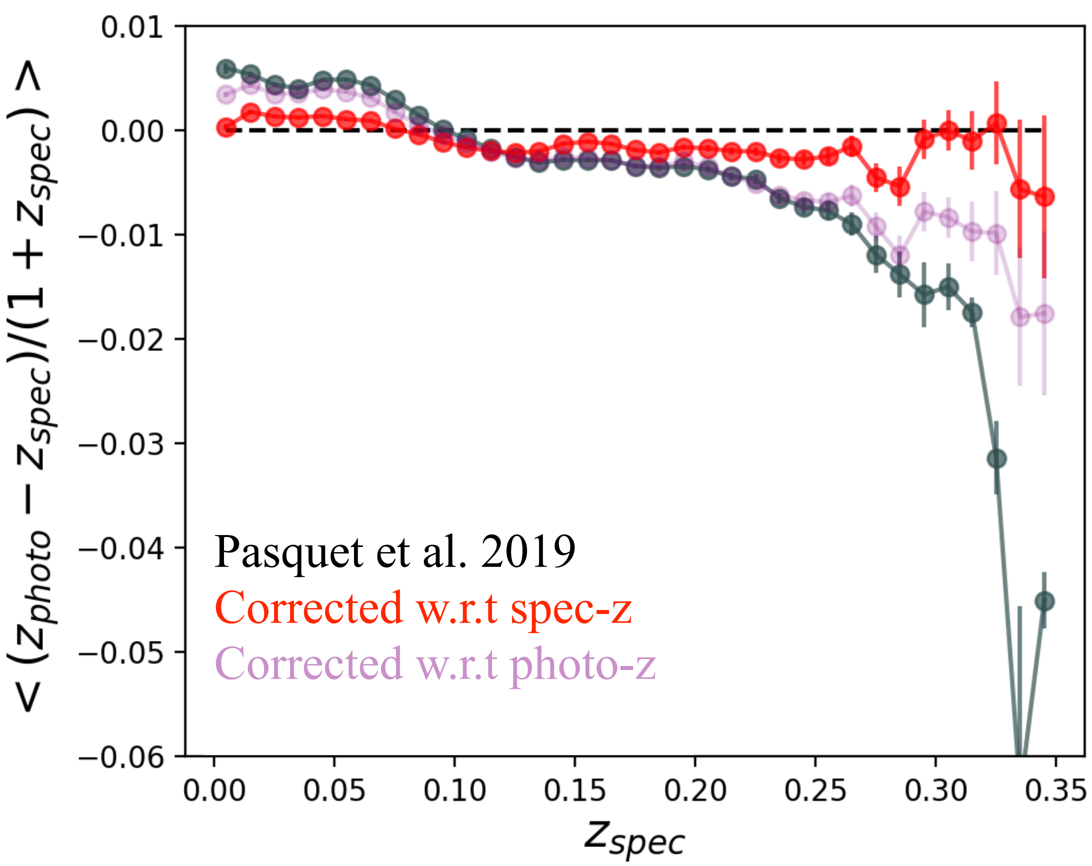
Unbalanced
Distribution

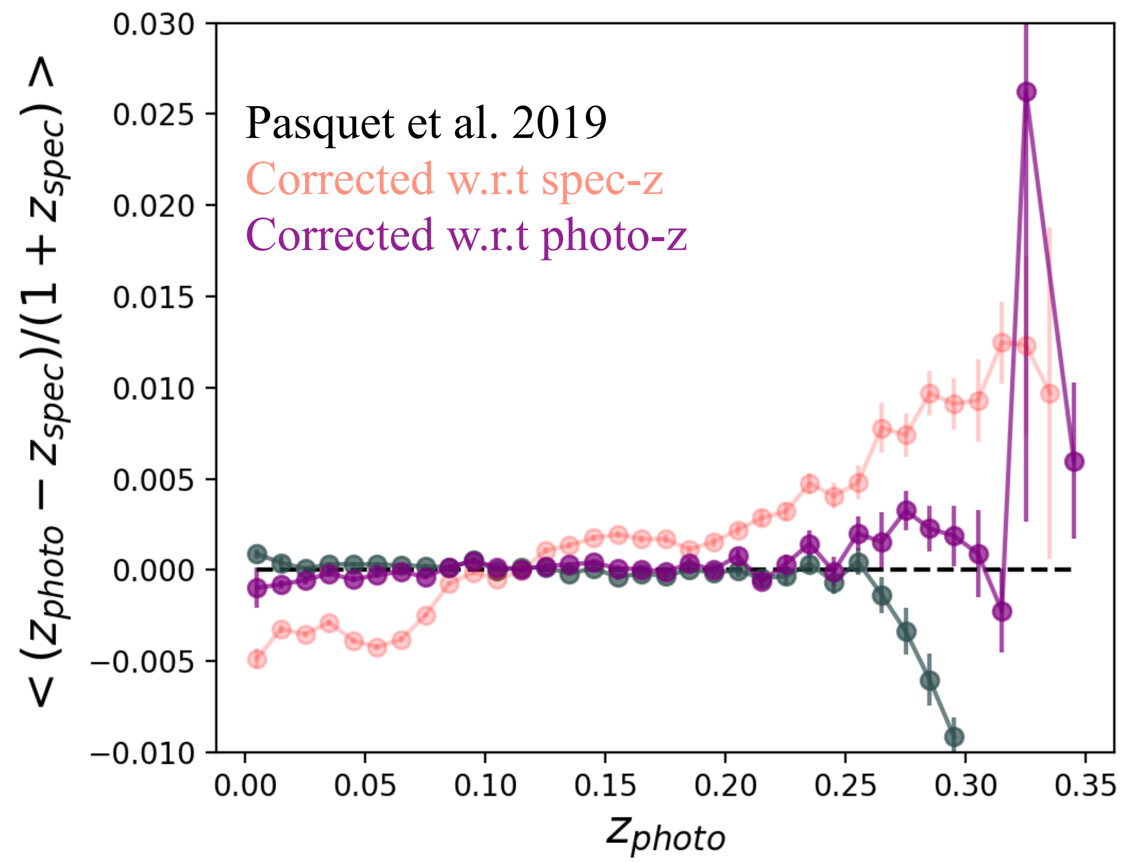
SDSS z < 0.4
(Pasquet et al. 2019)

Bias correction via splitting representation and estimation (Lin et al. 2022)

Representation Learning (all data) **Estimation (a near-balanced subset)**

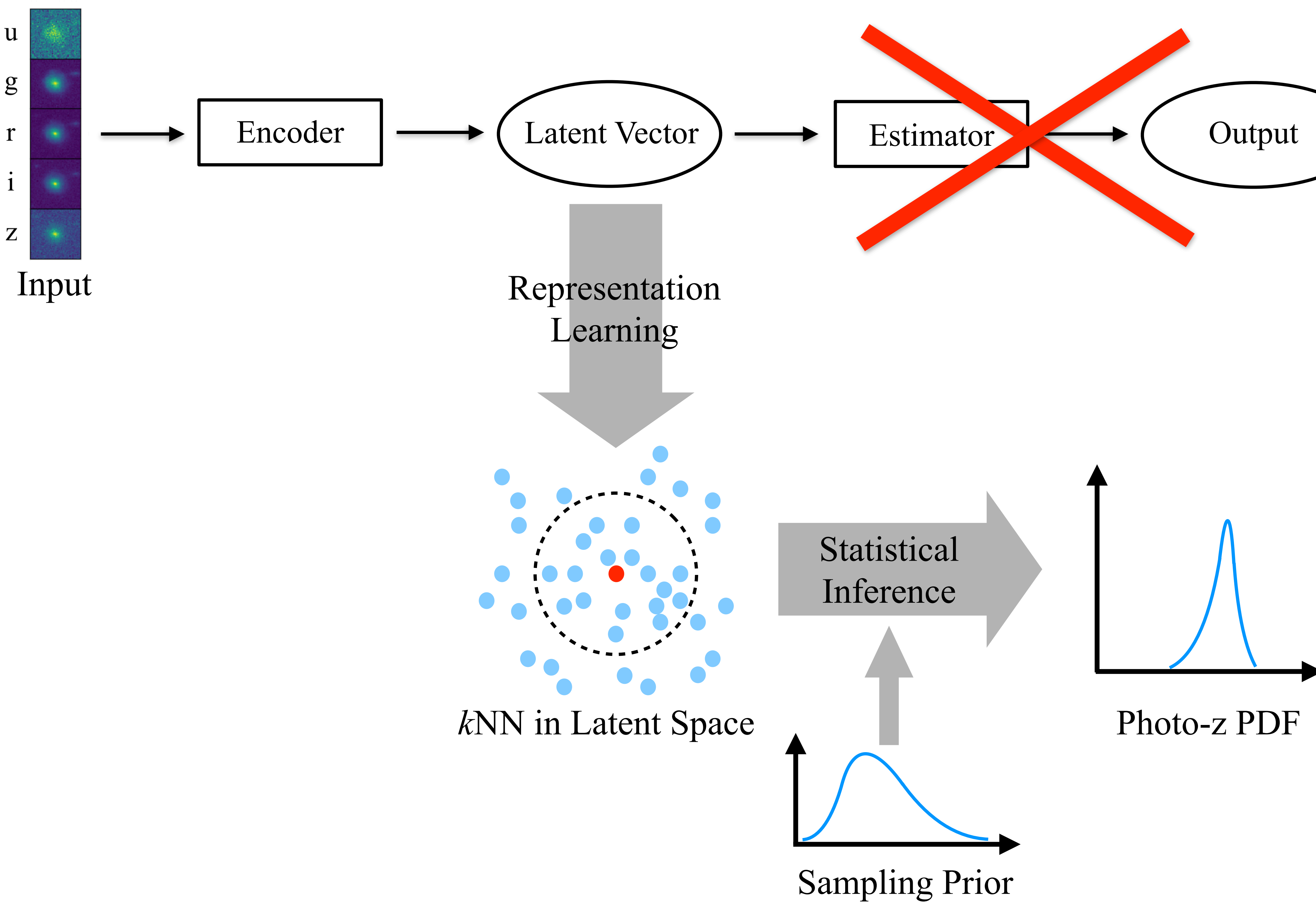
- Treat spectroscopic & photometric spaces separately:





Current work: empower deep learning with statistical basis

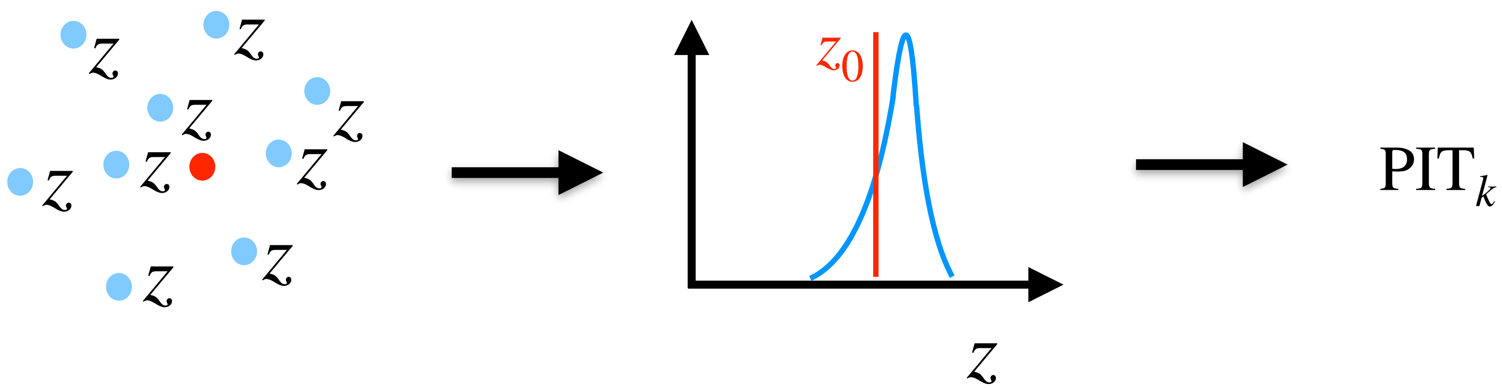
- Representation learning + statistical inference (or *k*NN)



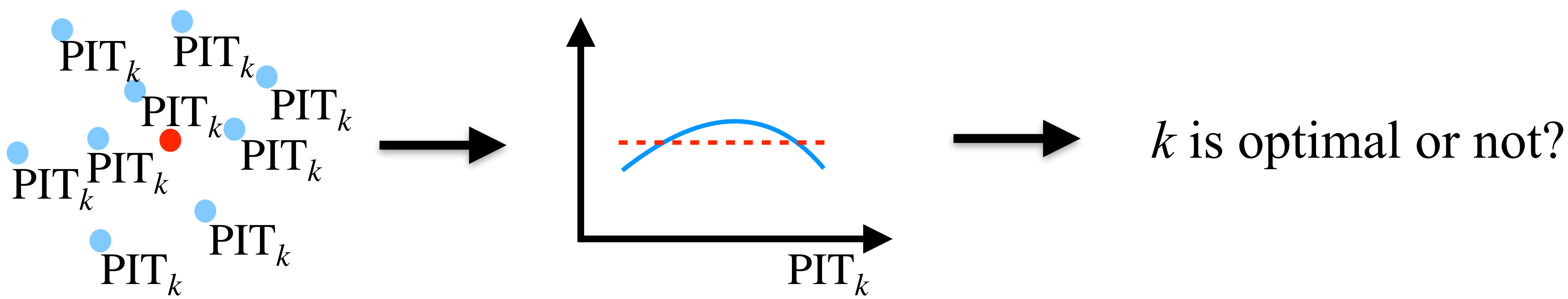
Determine the optimal *k* via local Probability Integral Transform (PIT) diagnostics

- Select k from [5, 10, 15, … , 2000]
- For each labeled galaxy:

$$\mathrm{PIT}(z_{spec}) = \int_0^{z_{spec}} p(z)dz$$



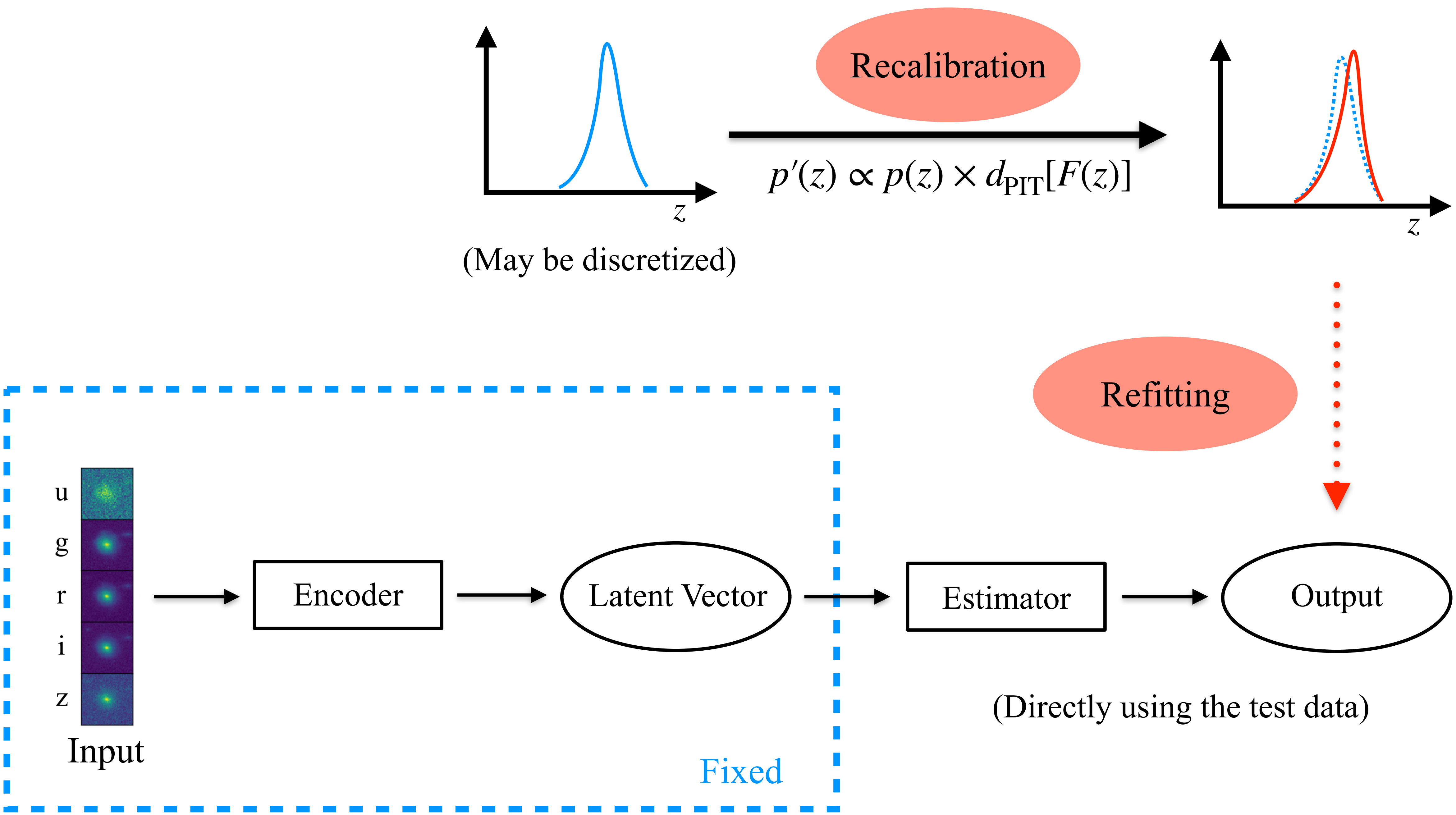
- For each query (unlabeled) galaxy:



- k is optimal when the PIT_k distribution is closest to a **uniform** distribution

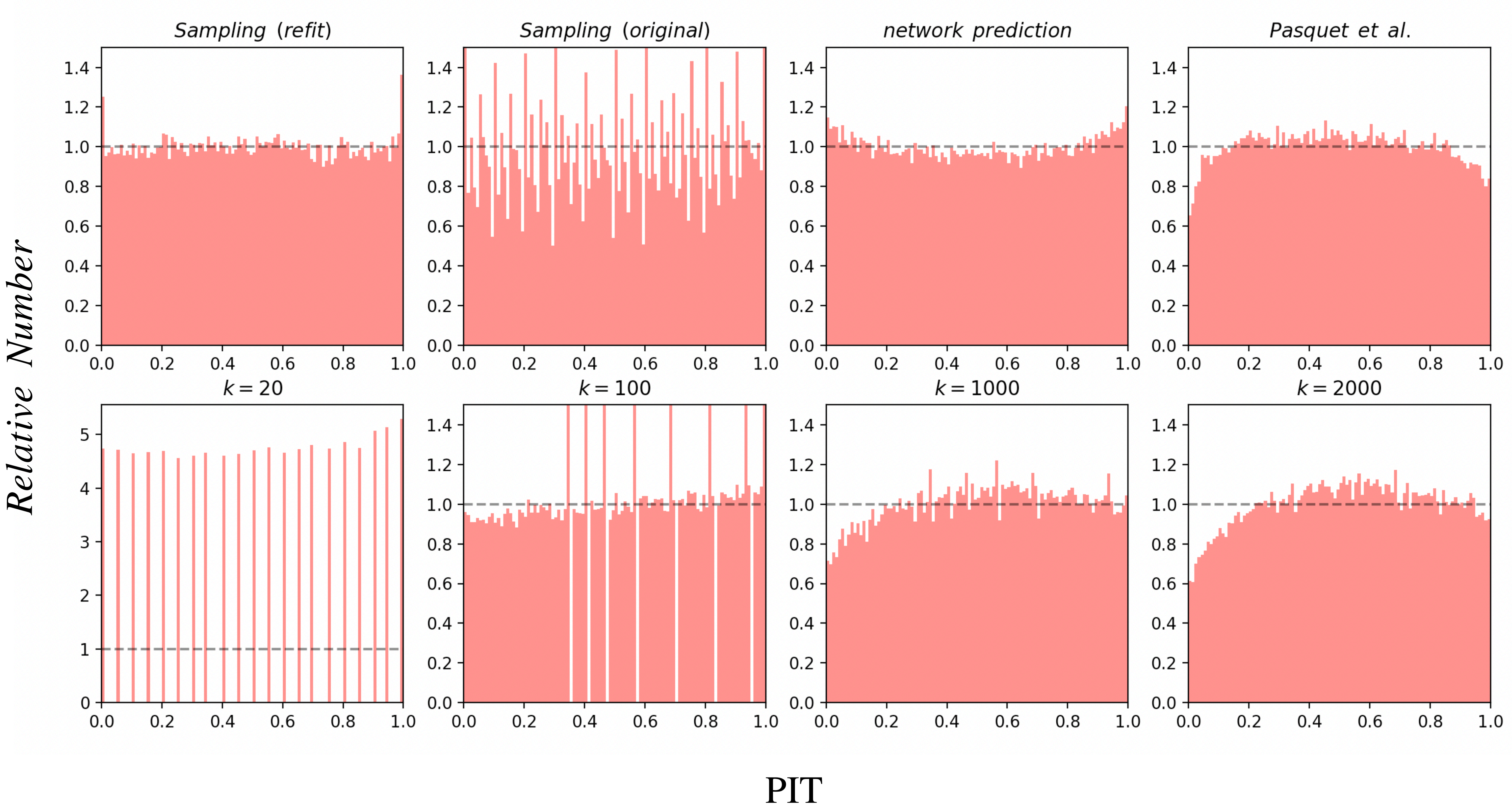
Recalibration + Refitting

- Discretized —> Recalibration
- Non-uniform —> Refitting



Results: PITs

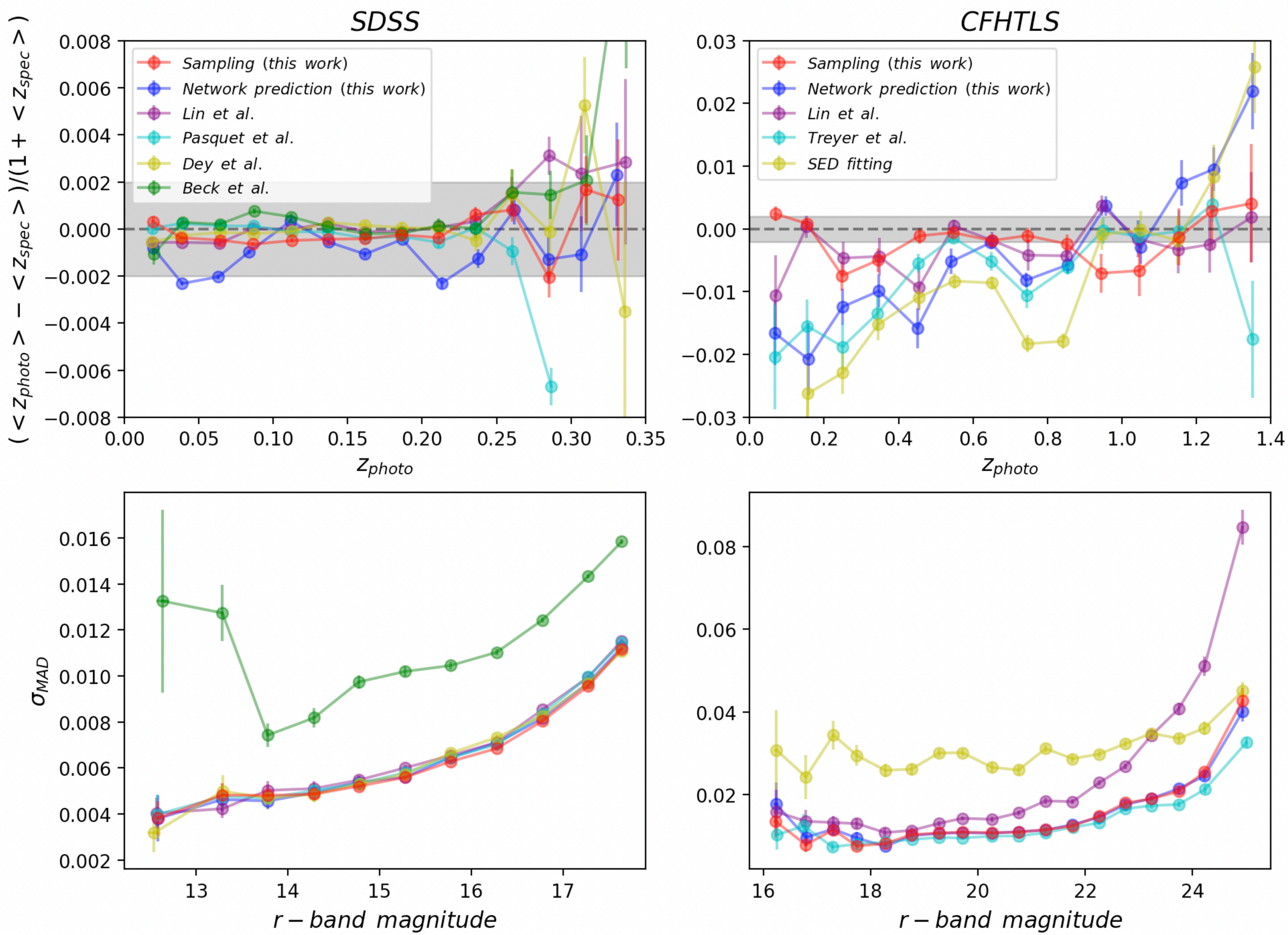
- Good PDF calibration achieved by sampling/inference (shown for the SDSS data)



Results: point estimates

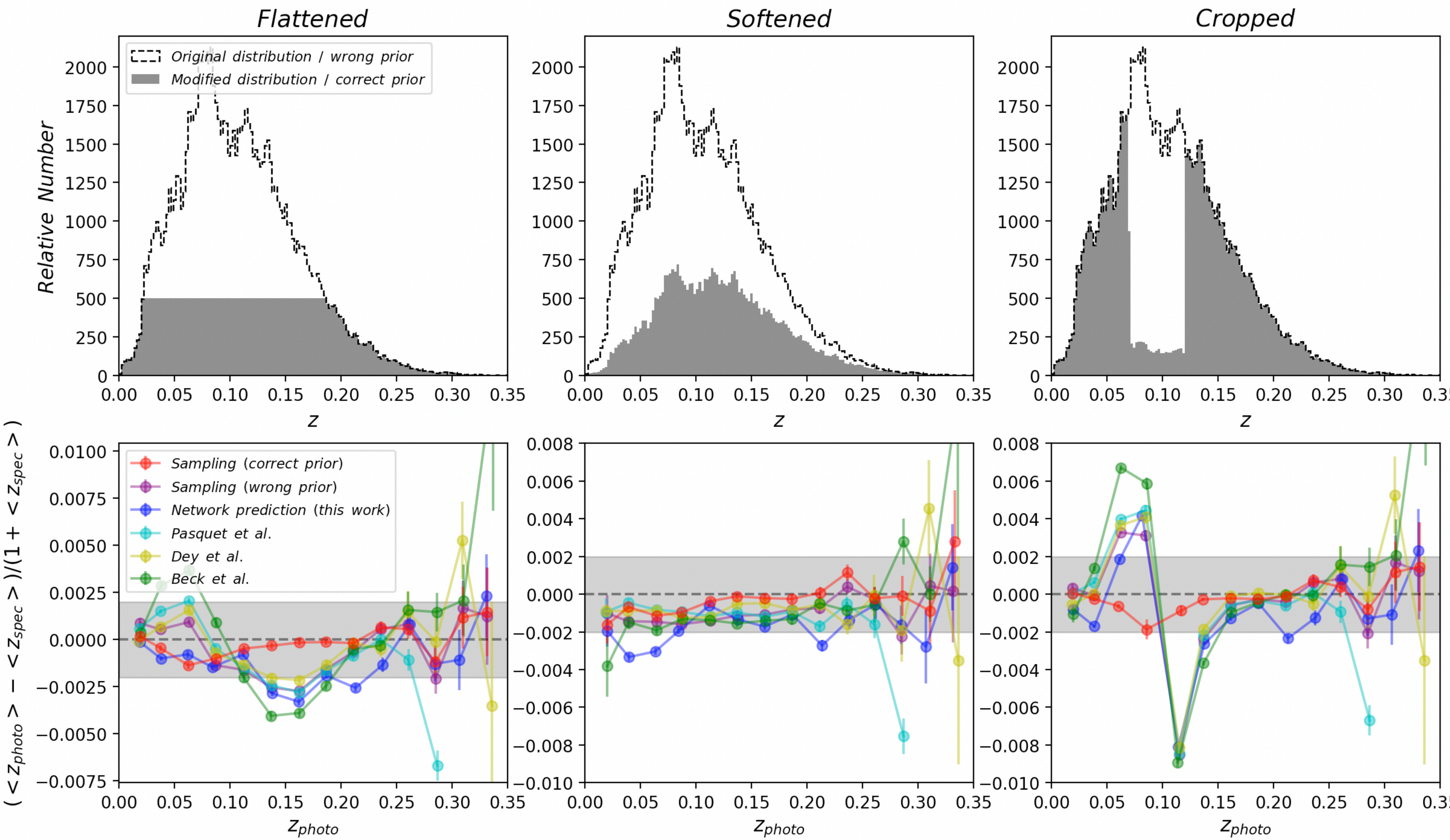
- Mean redshift bias correction achieved by sampling/inference (similar to Lin et al.)
- No loss in accuracy (contrary to Lin et al.)

$$z_{photo} = \int_0^{z_{max}} z \times p(z)dz$$



Results: the impact of distribution mismatch

- Robustness under distribution mismatch with correct sampling prior

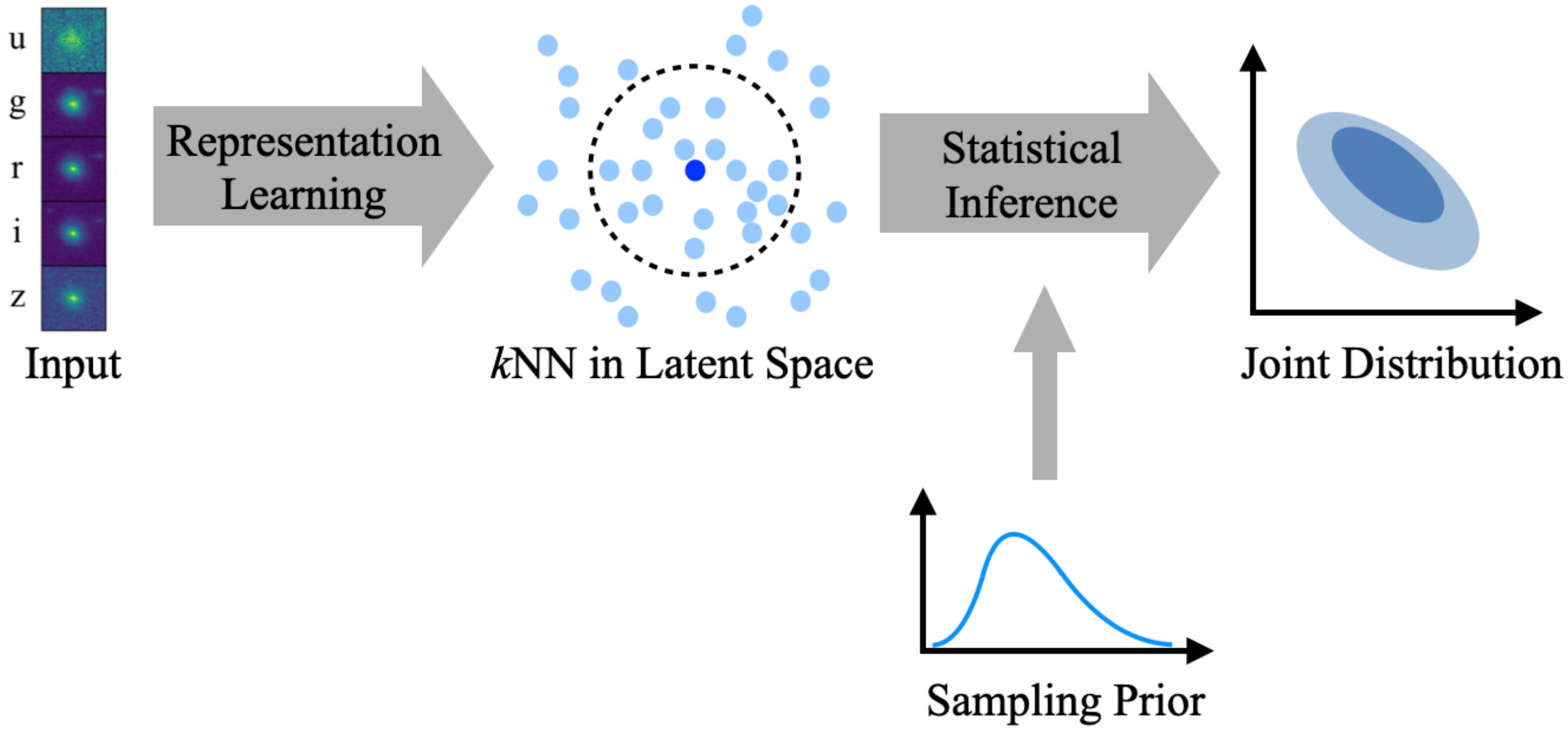


Summary

- Key idea: combine deep learning and statistical basis
- Representation learning, statistical inference, recalibration & refitting
- Better results over benchmark methods:
  - Well-calibrated PDFs
  - Good control of photo-z-dependent residuals without compromising accuracy
  - Robustness under distribution mismatch

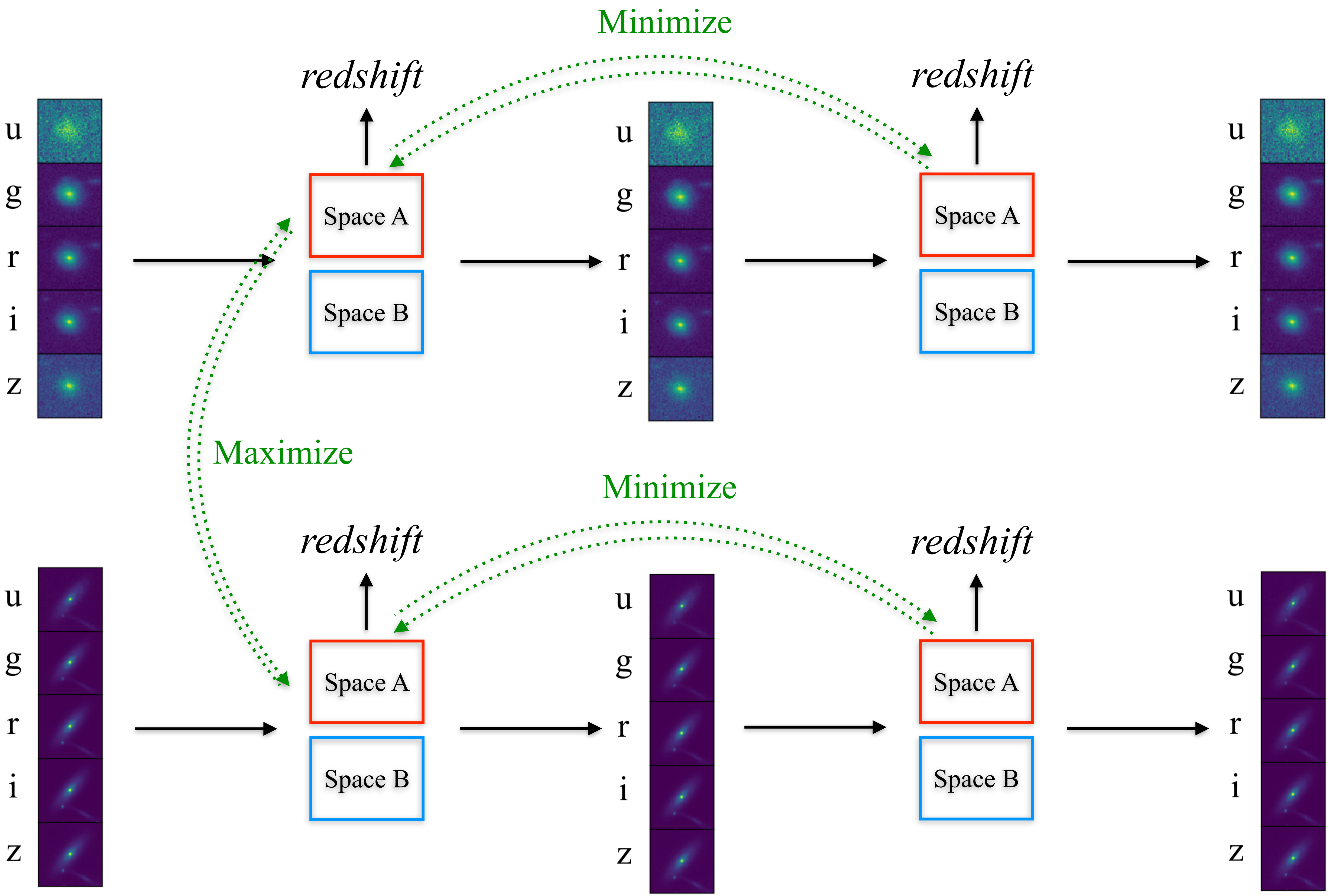
For interpretability: analyze *redshift-variable* correlations

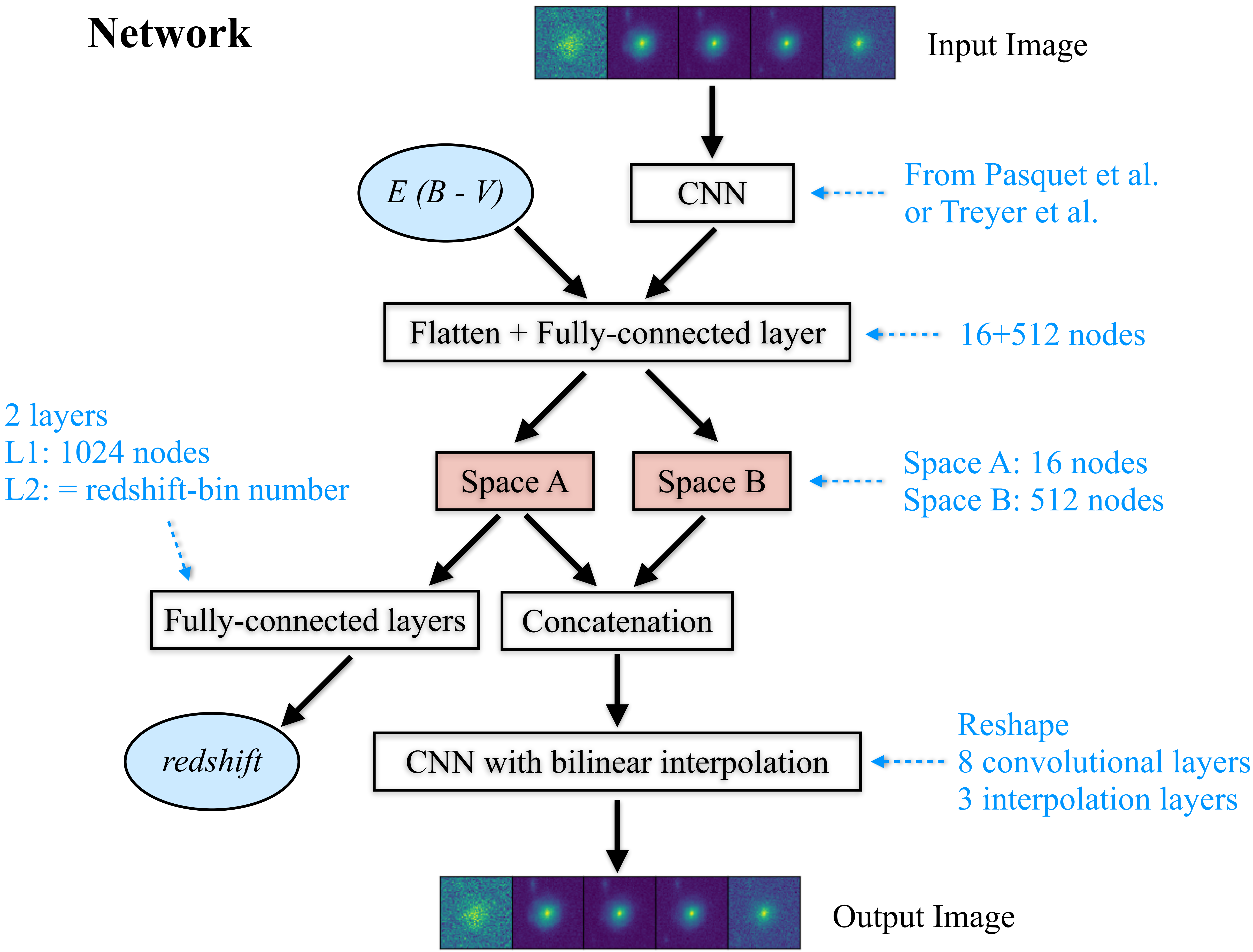
- Information/variables to be exploited for reducing redshift residuals (e.g., galaxy structures, environmental properties, etc.)
- Relations between variables



Back-up slides

Representation learning

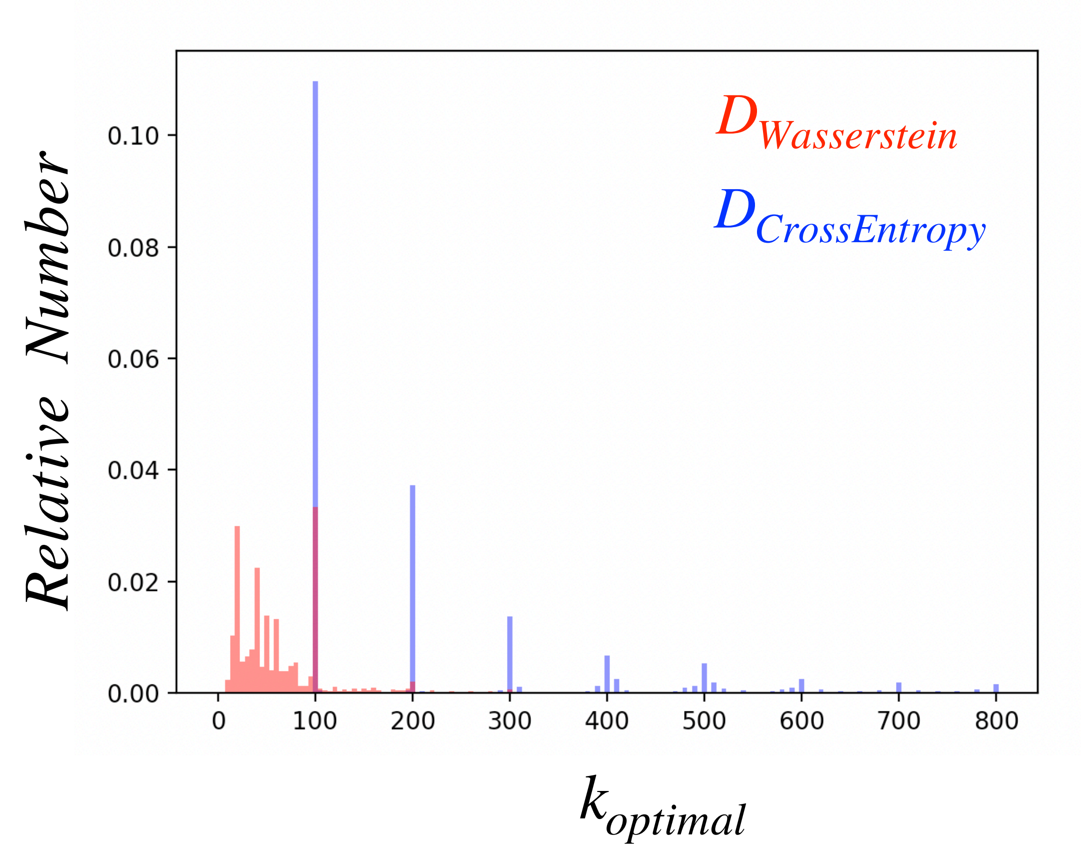


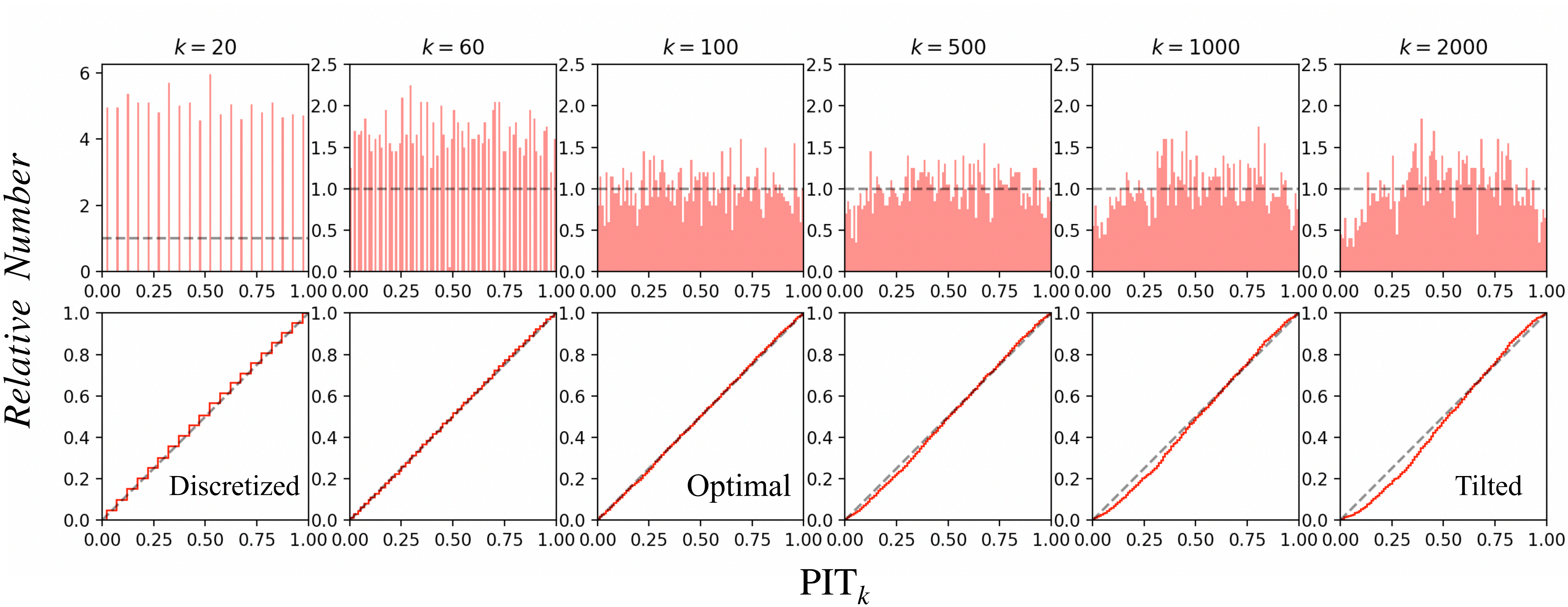

Network
Input Image
E (B - V)
CNN
From Pasquet et al.
or Treyer et al.
Flatten + Fully-connected layer
16+512 nodes
2 layers
L1: 1024 nodes
L2: = redshift-bin number
Space A
Space B
Space A: 16 nodes
Space B: 512 nodes
Fully-connected layers
Concatenation
redshift
CNN with bilinear interpolation
Reshape
8 convolutional layers
3 interpolation layers
Output Image

Optimal metric for PIT diagnostics: Wasserstein distance

$$D_{Wasserstein}[P_{\mathrm{PIT}_k}, P_{uniform}] = \int_0^1 |F_{\mathrm{PIT}_k}(p) - F_{uniform}(p)|\, dp$$

$$D_{CrossEntropy}[P_{\mathrm{PIT}_k}, P_{uniform}] = \int_0^1 [P_{uniform}(p) \log P_{\mathrm{PIT}_k}(p) + (1 - P_{uniform}(p)) \log(1 - P_{\mathrm{PIT}_k}(p))]\, dp$$





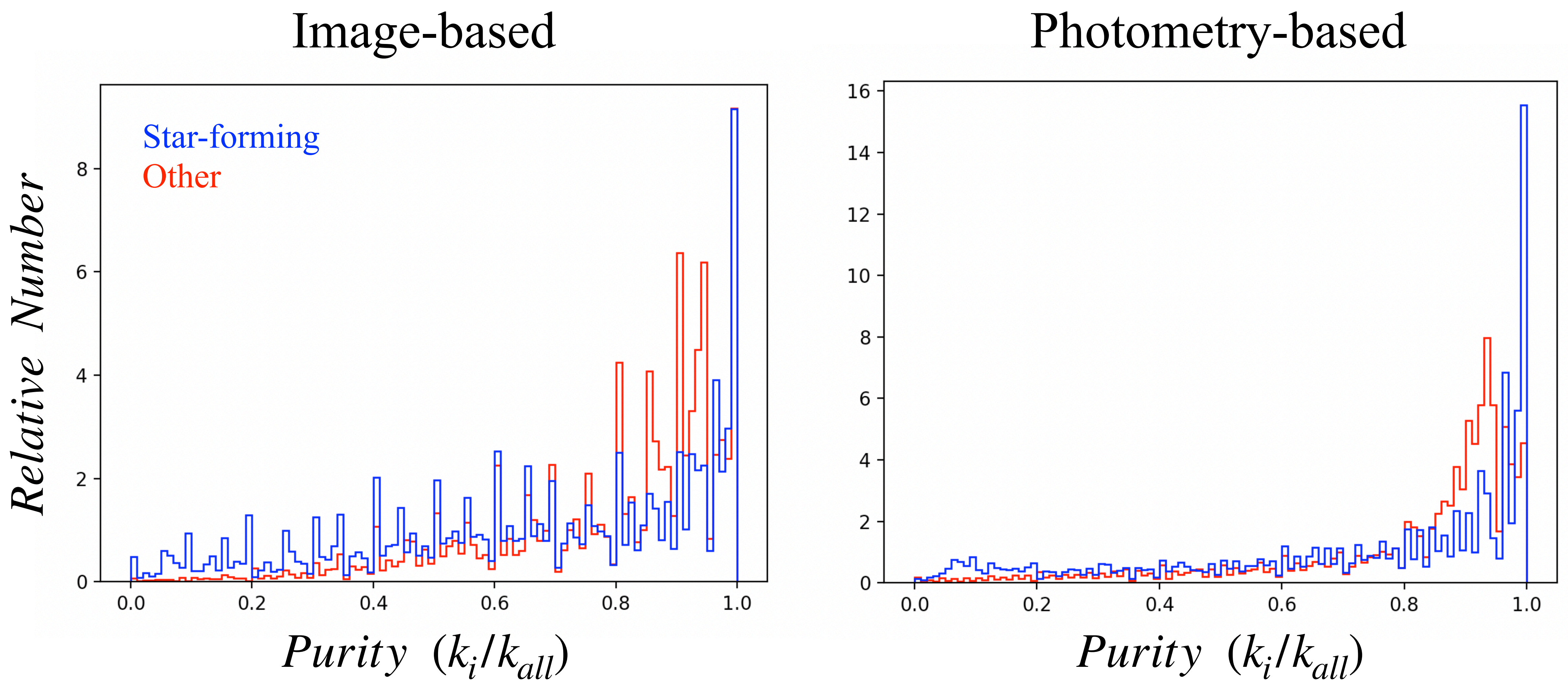
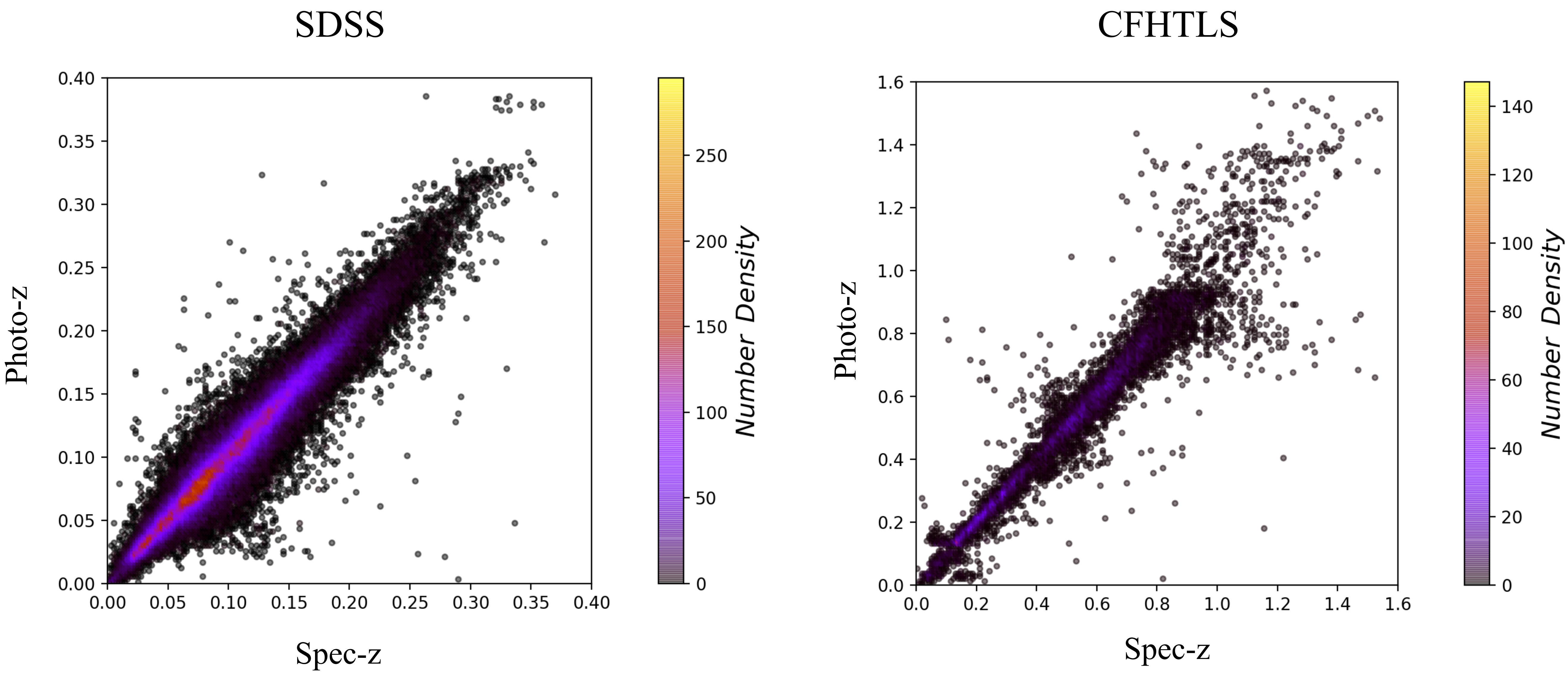
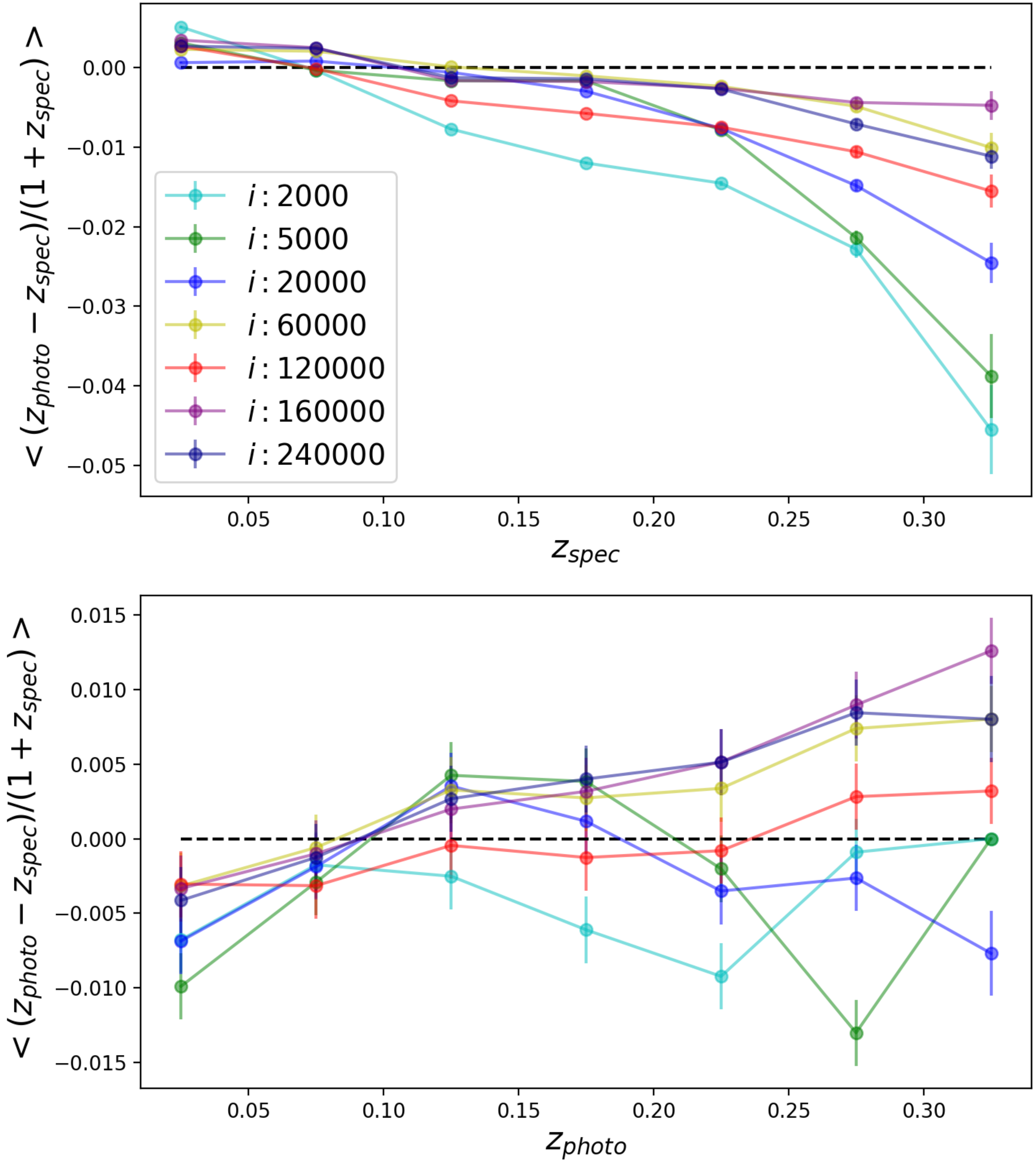


Image-based
Photometry-based
Star-forming
Other
Relative Number
Purity (k_i/k_{all})
Purity (k_i/k_{all})



SDSS
Photo-z
Spec-z
Number Density
CFHTLS
Photo-z
Spec-z
Number Density



$< (z_{photo} - z_{spec})/(1 + z_{spec}) >$
0.00
−0.01
−0.02
−0.03
−0.04
−0.05
$i : 2000$
$i : 5000$
$i : 20000$
$i : 60000$
$i : 120000$
$i : 160000$
$i : 240000$
0.05
0.10
0.15
0.20
0.25
0.30
z_{spec}
$< (z_{photo} - z_{spec})/(1 + z_{spec}) >$
0.015
0.010
0.005
0.000
−0.005
−0.010
−0.015
0.05
0.10
0.15
0.20
0.25
0.30
z_{photo}