LSST DESC
Dark Energy Science Collaboration

ALCCA
Avestruz Lab for Computational Cosmology and Astrophysics

Probabilistic Machine Learning Group (UM)

Bayesian Light Source Separator (BLISS): Probabilistic detection, deblending and measurement of astronomical light sources

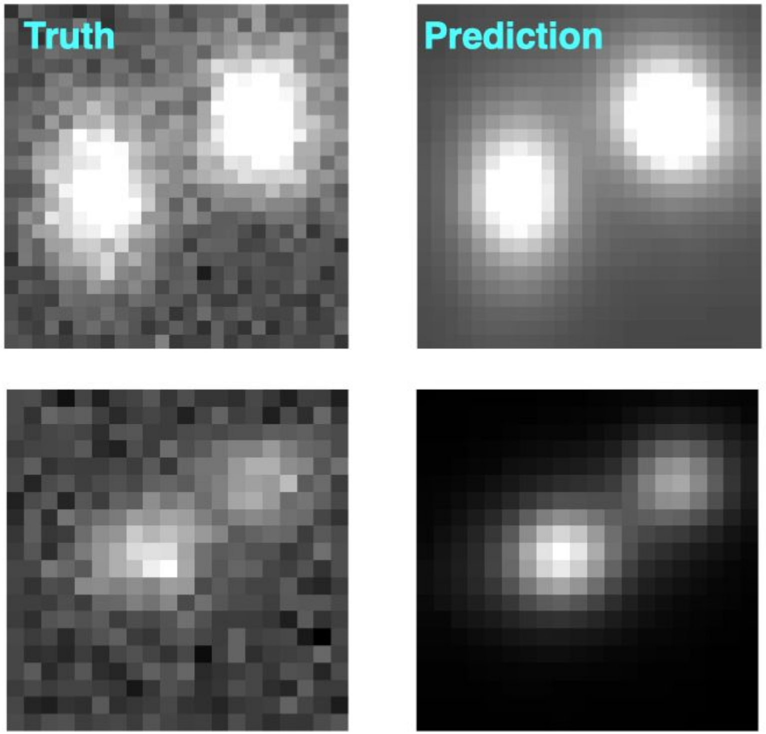
Ismael Mendoza
Department of Physics, University of Michigan

C. Avestruz
Department of Physics, University of Michigan

J. Regier, D. Hansen, R. Liu, Z. Zhao, Z. Pang
Department of Statistics, University of Michigan

Outline

- The Blending Problem
- Overview of BLISS Framework
- Results: BLISS applied on simulated data
- Future Directions and Conclusions



Motivation: The Blending Problem in Stage-IV surveys

- Stage-IV Cosmological surveys (such as LSST) will reach higher depths than ever before.

- Higher depth ⇒ higher number density of galaxies ⇒ **~ 60% will visually overlap (blend)!**

- For **precision cosmology**, need accurate source detection and measurement. (e.g. flux and photo-z's, shapes and weak lensing)

- **⇒ Need accurate, robust and fast deblenders.**

Simulated 5x5 arcsec LSST patch

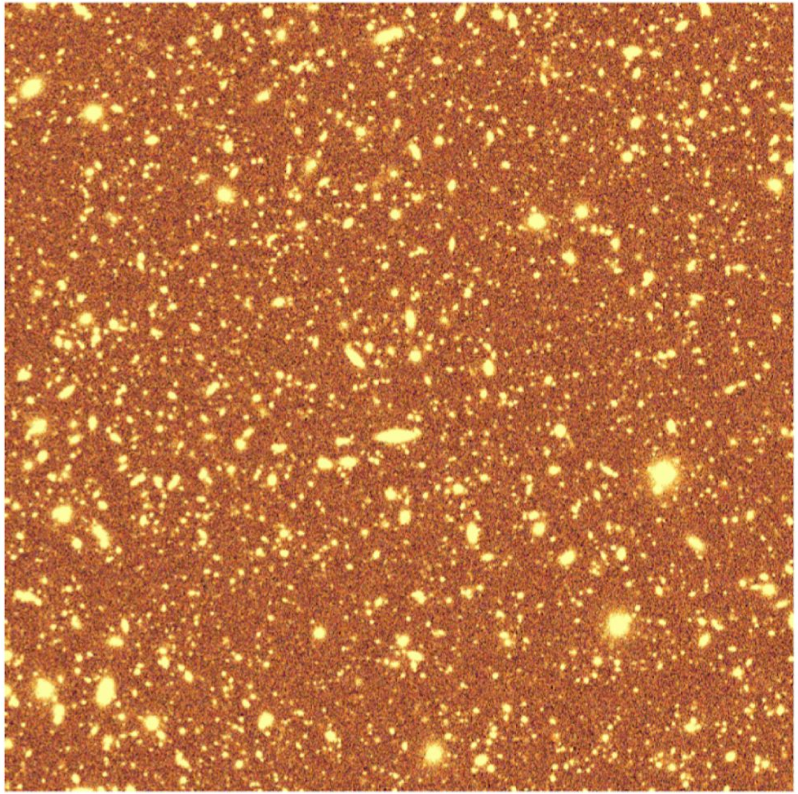
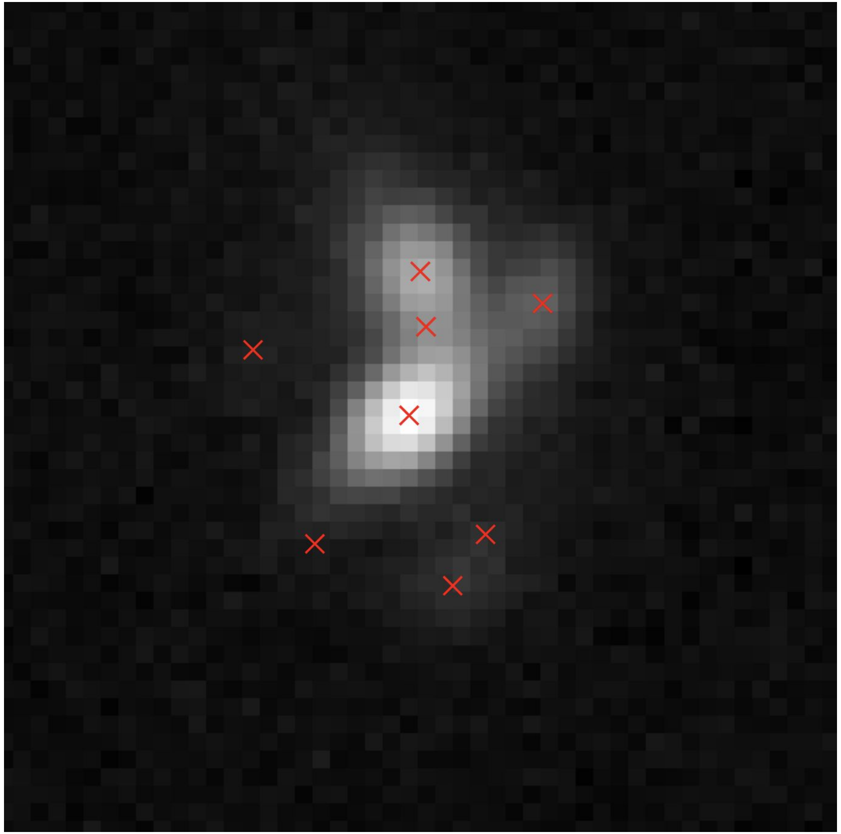


Image Credit: Kamath 2020 Thesis

Probabilistic cataloging and uncertainty

- Number of sources could be highly uncertain for any given galaxy blend.
- There is also high flux uncertainty as different galaxies could share different portions of the flux in the image.
- **⇒ Need a way to estimate and propagate uncertainty to downstream analysis**

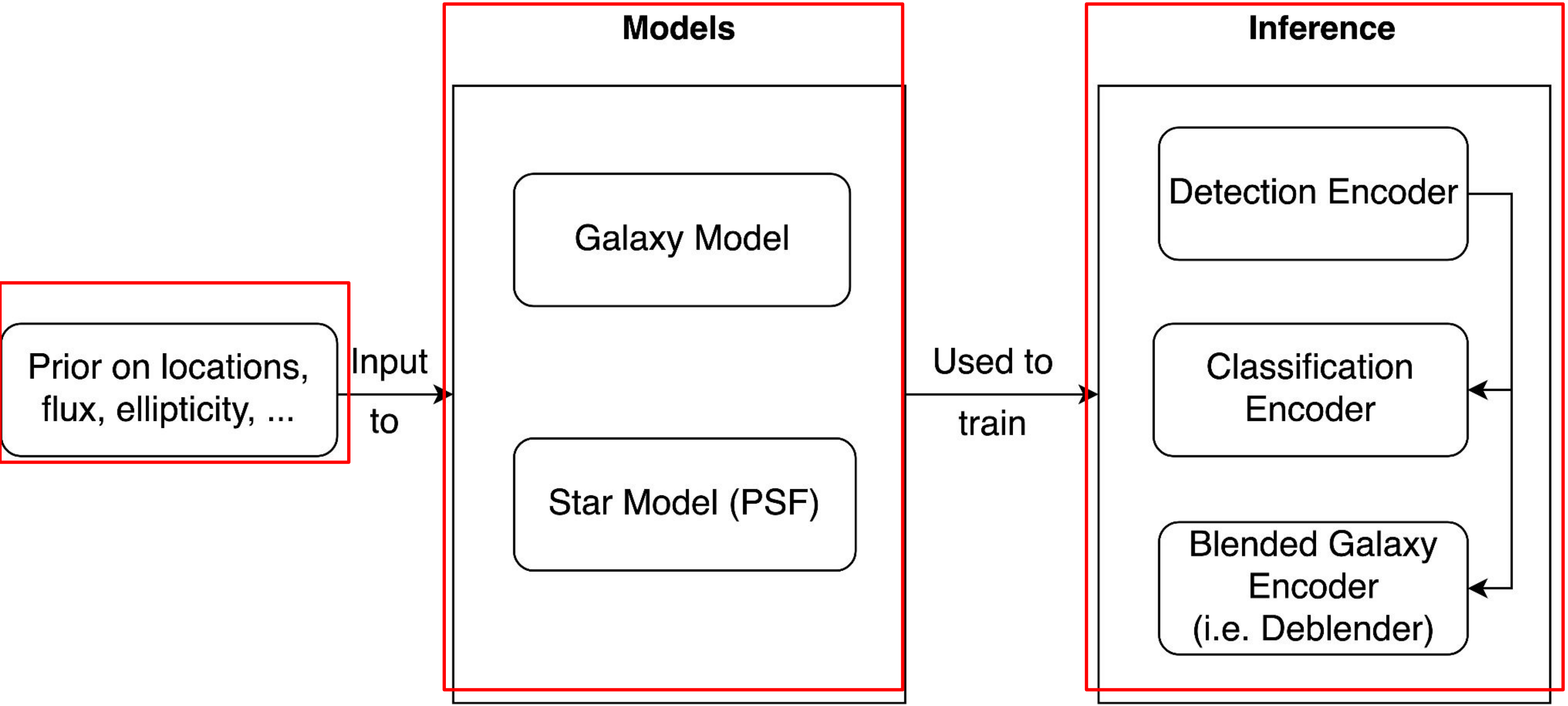


Overview of BLISS Framework

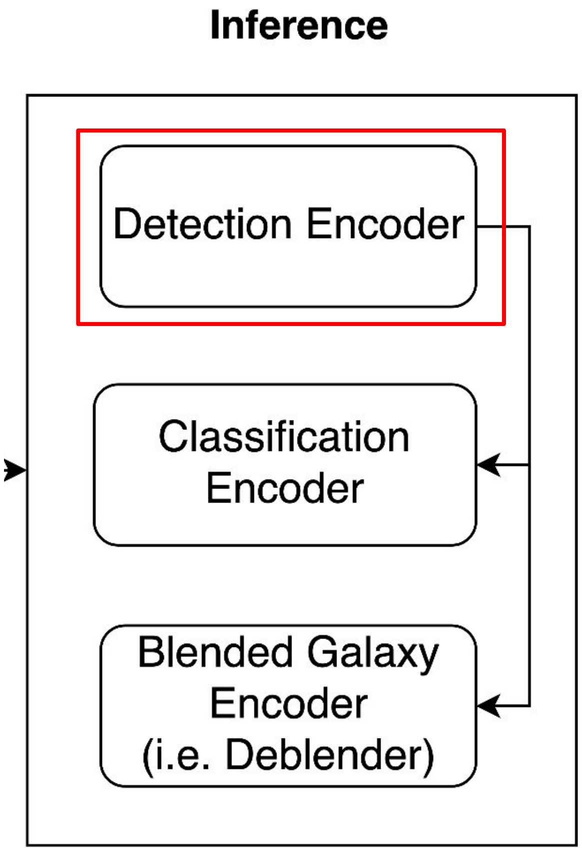
What is the Bayesian Light Source Separator (BLISS) ?

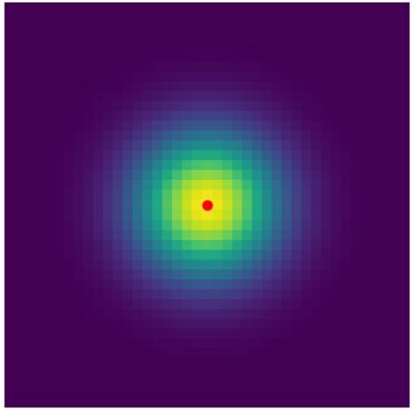
A Bayesian framework for detection, deblending, and measurement of galaxies and stars.

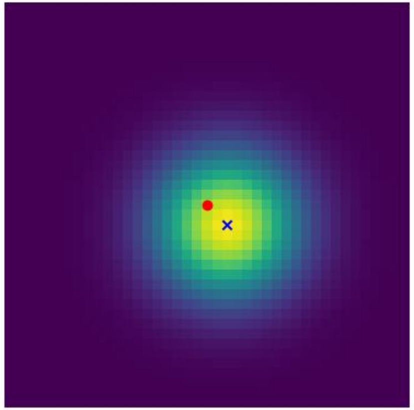
BLISS at a glance



Not requiring centroids as additional input

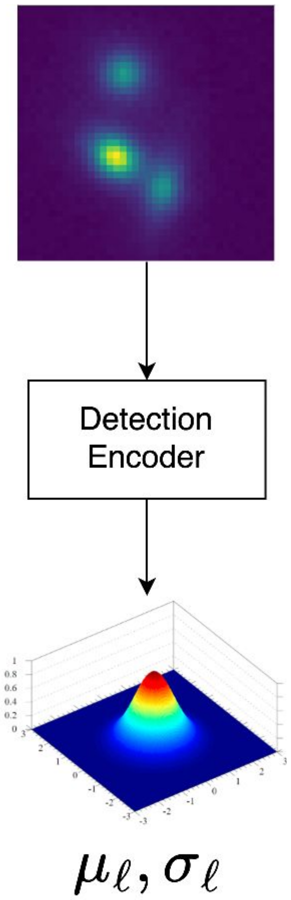






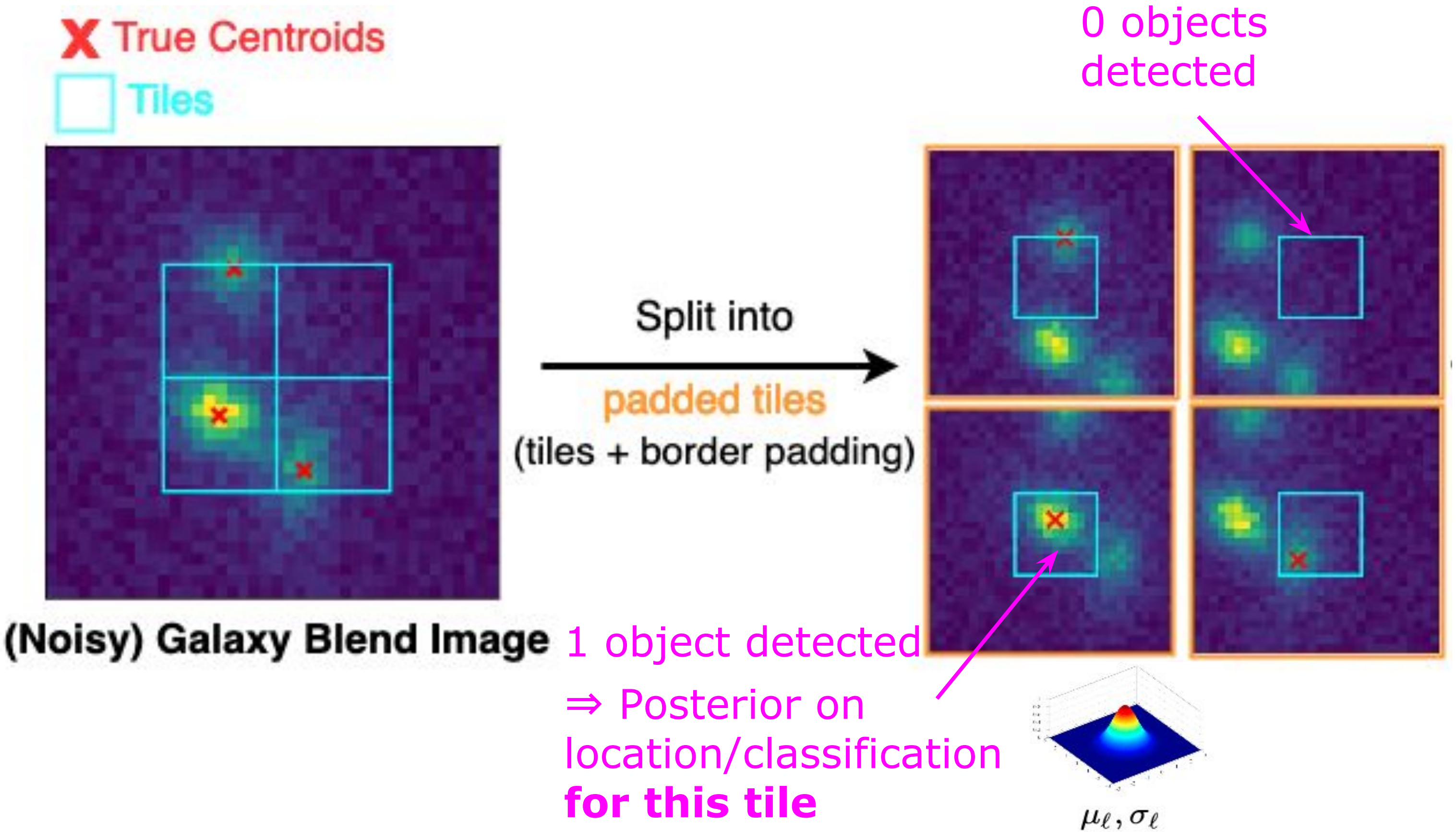
Neural networks parametrize distribution on measured quantities

- Number of sources
- Locations
- Classification boolean
  (star vs galaxy)
- Star fluxes
- Galaxy properties (shape + flux)

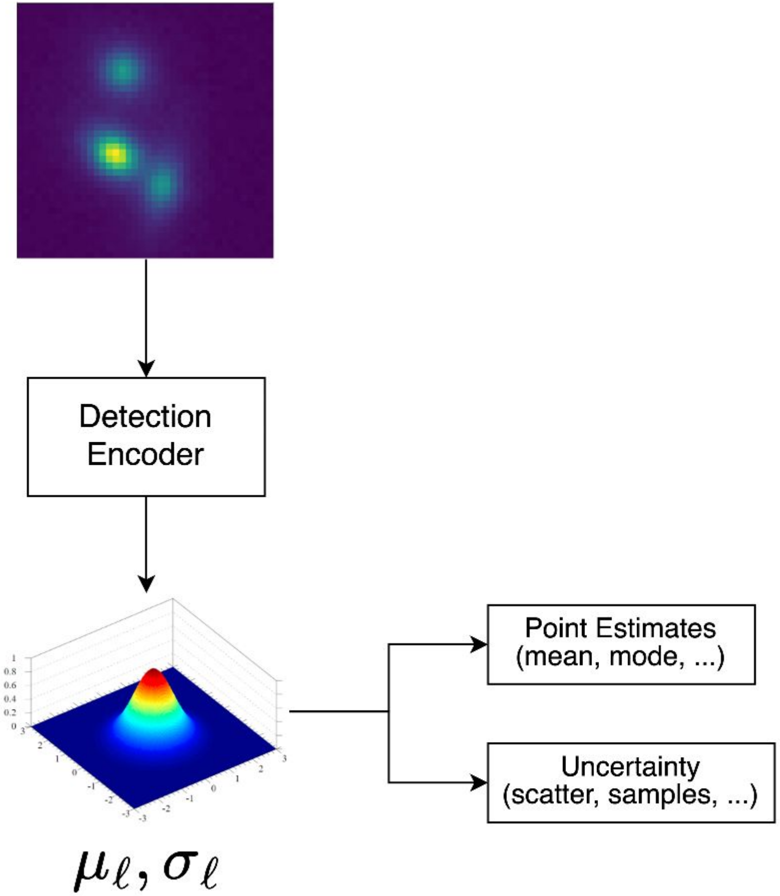


BLISS predictions are done “per tile”

- Images of blended galaxies are split into **tiles**, predictions of properties are done per **tile**.



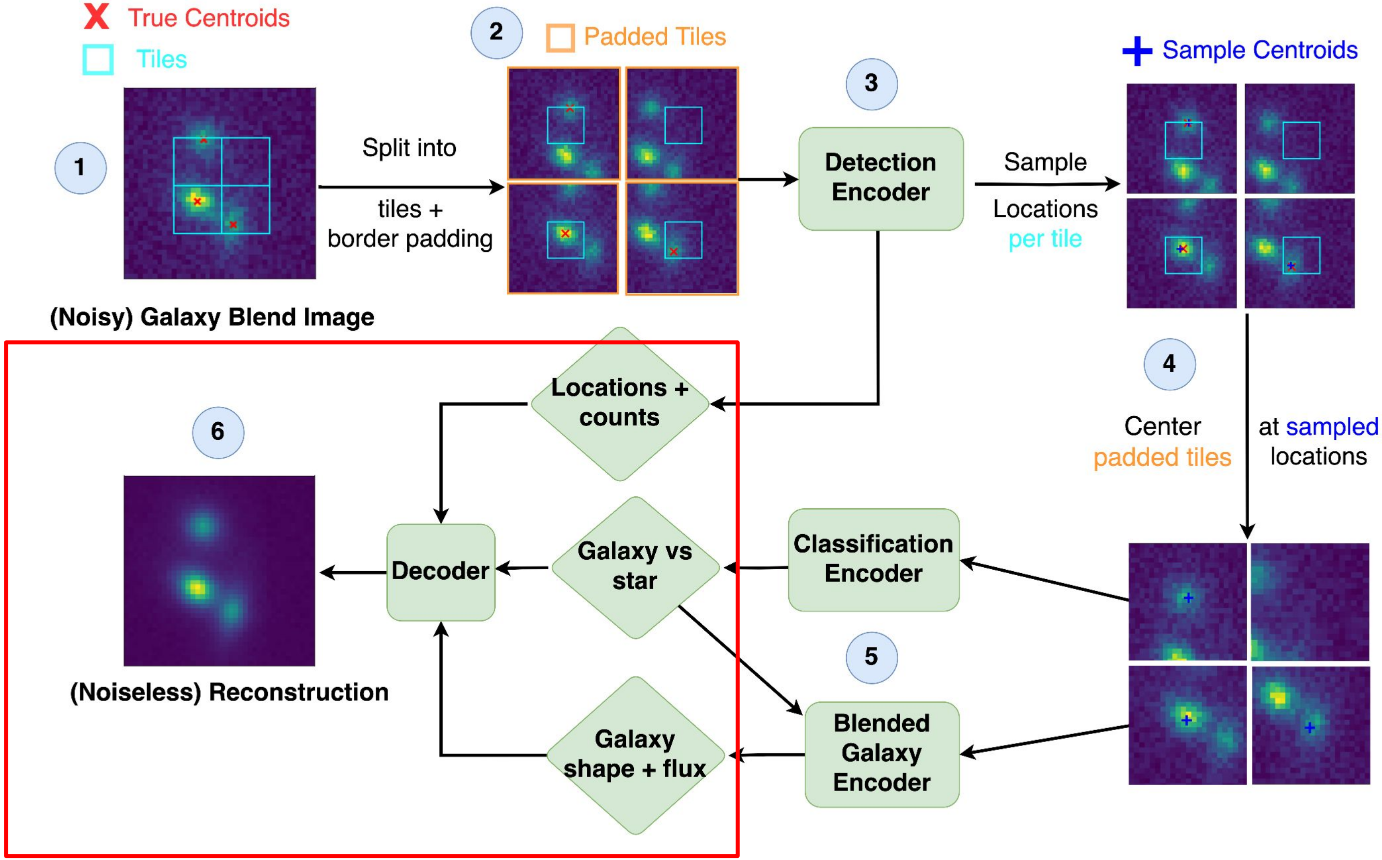
Output of algorithm on images provides a **probabilistic interpretation** of measured quantities and corresponding uncertainties.



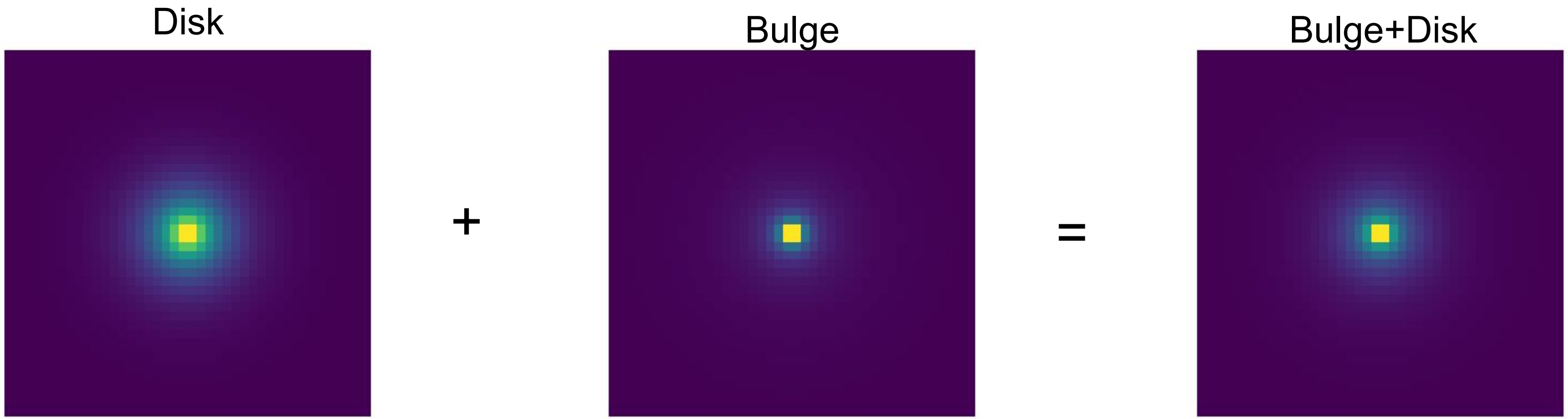
A Deep Learning backbone using pytorch

PyTorch

Outline



True Centroids
Tiles
1
(Noisy) Galaxy Blend Image
Split into
tiles +
border padding
2
Padded Tiles
3
Detection Encoder
Sample
Locations
per tile
Sample Centroids
4
Center
padded tiles
at sampled
locations
Classification Encoder
5
Blended Galaxy Encoder
Locations + counts
Galaxy vs star
Galaxy shape + flux
Decoder
6
(Noiseless) Reconstruction

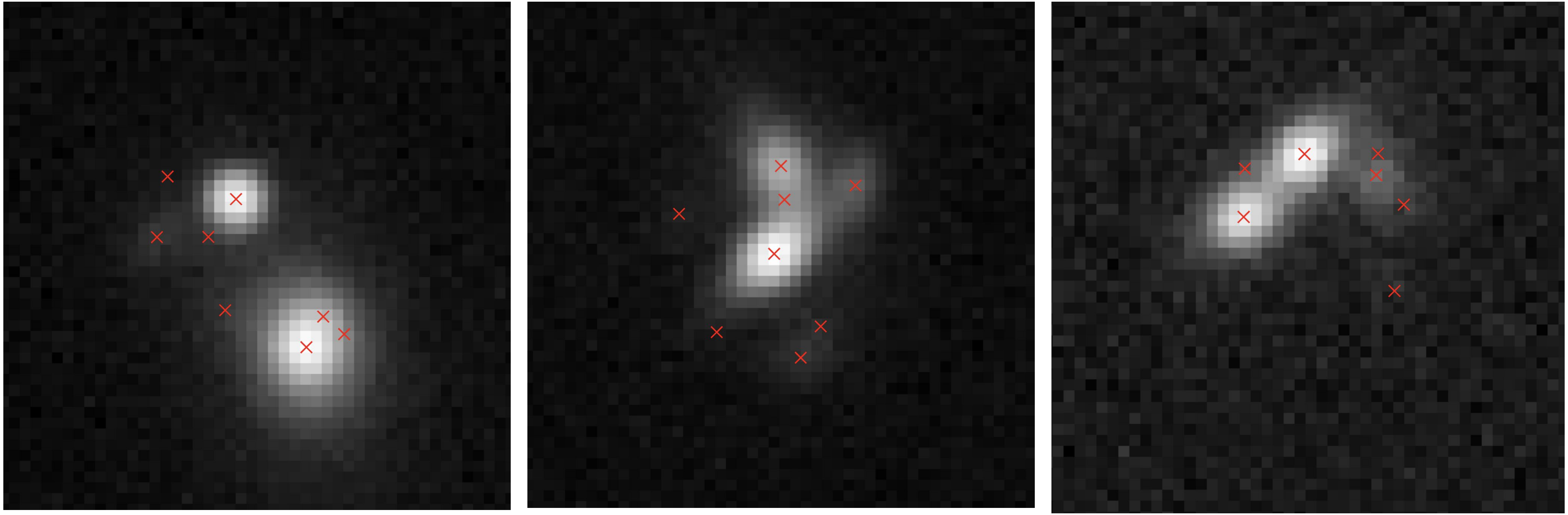


Results: BLISS applied on simulated data

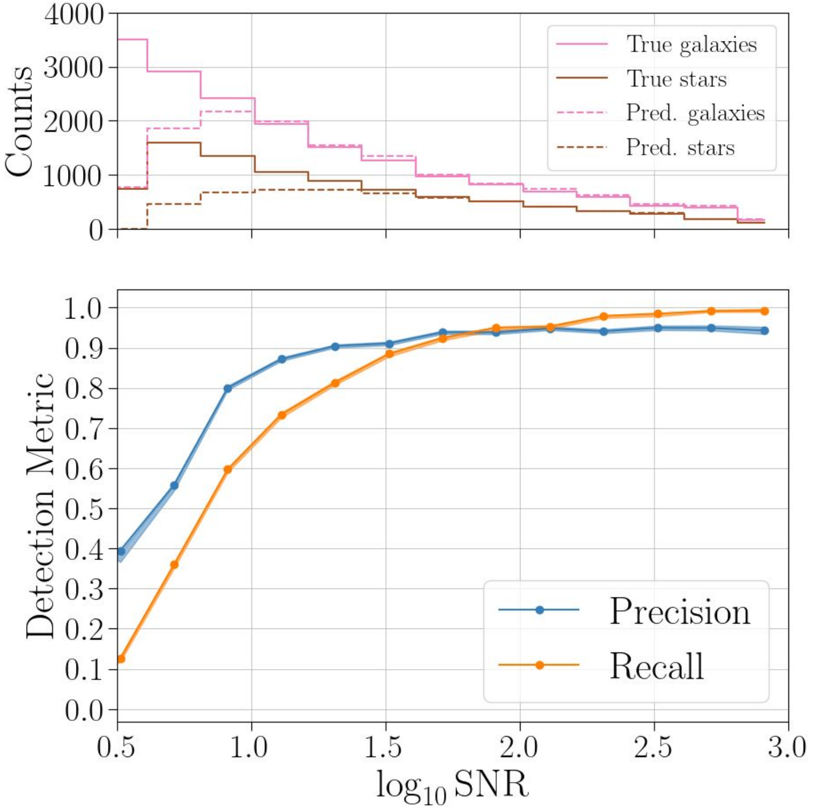
- Produced via a galaxy simulator 'Galsim'. Used to create galaxy blends for both training and testing of BLISS.

- Galaxies are **"Bulge+Disk"** parametric galaxies.

Example of blends with true centroids



BLISS point-estimate **detection** metrics on 10k **blend dataset**



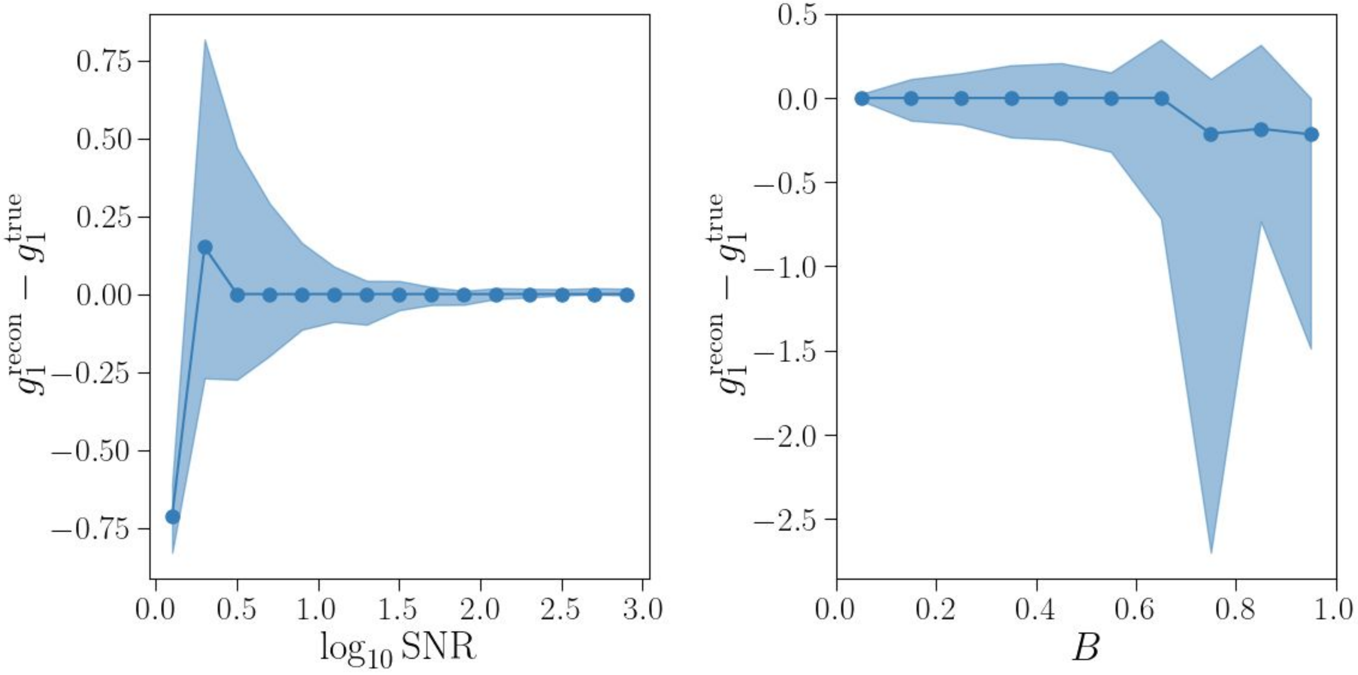
Precision =
\# Matched predicted sources /
\# Predicted sources

Recall =
\#Matched predicted sources /
\#True sources

Matches are detected (>50%** detection probability) galaxy/star centroids less than 1 pixel way from true centroid.

Takeaway: BLISS can detect majority of sources > 7 SNR.
BLISS maintains higher precision than recall for low SNR sources.

BLISS measurement residuals on **ellipticity** on 10k **blend dataset.**



Takeaway: BLISS residuals on ellipticities are consistent with 0 for matched objects (but low SNR and high B regions are noisy).

Future directions and Conclusions

Future Directions

- **Flexible galaxy probabilistic model:** Using a VAE, normalizing flows, or diffusion models.

- **Real data application** on SDSS, DES, and (future) LSST data.

- **Propagate uncertainties** to cosmological downstream analysis.

Thank you!

Conclusions

- BLISS splits astronomical images into **tiles**. Each tile is an independently input to a neural network that outputs a distribution of **light source parameters** for that tile.

- BLISS can output number of sources, centroids, classification, star flux, and galaxy flux and shapes.

- BLISS can **accurately detect, classify, and deblend** simulated galsim blends for relatively low SNRs and high blendedness.

- Case study suggests **probably estimates are reasonable** e.g. uncertainty in detection increases for high blendedness regime.

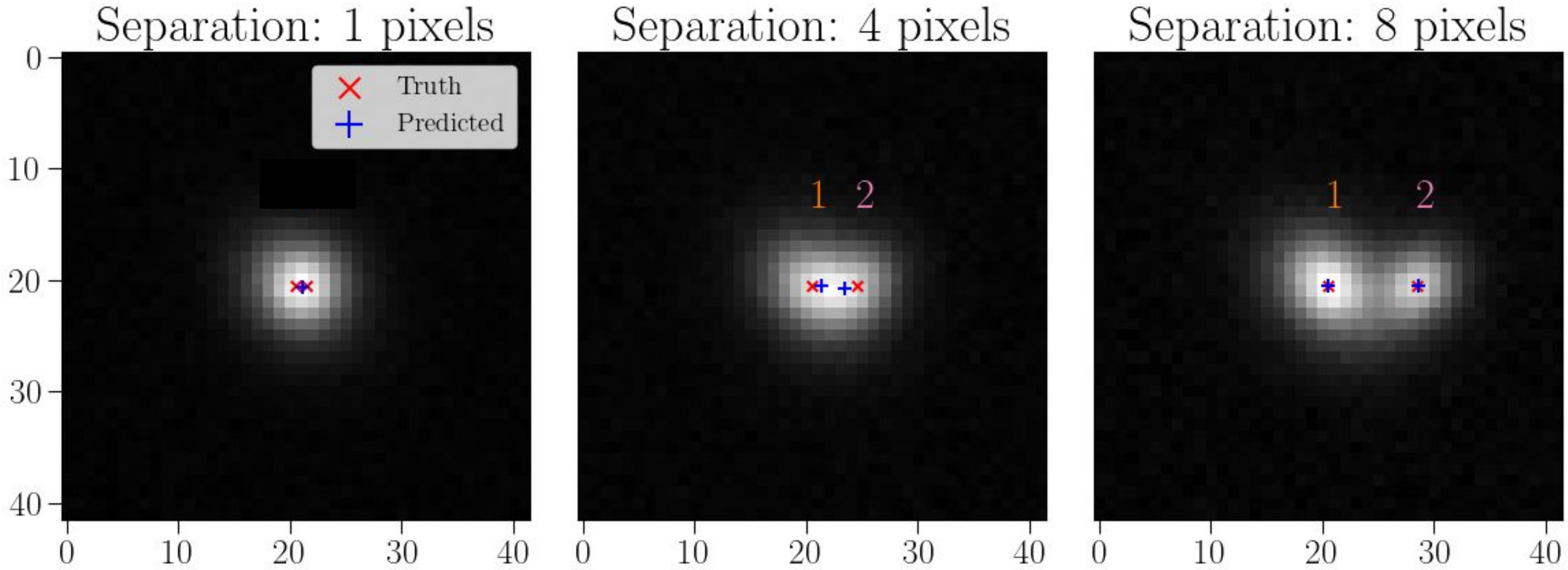
For more info on BLISS…

- Feel free to email imendoza@umich.edu for any follow-up questions on BLISS.
- You can also ask in in our channel #desc-bl-bliss
- For latest developments, follow our github: https://github.com/prob-ml/bliss

Extra slides

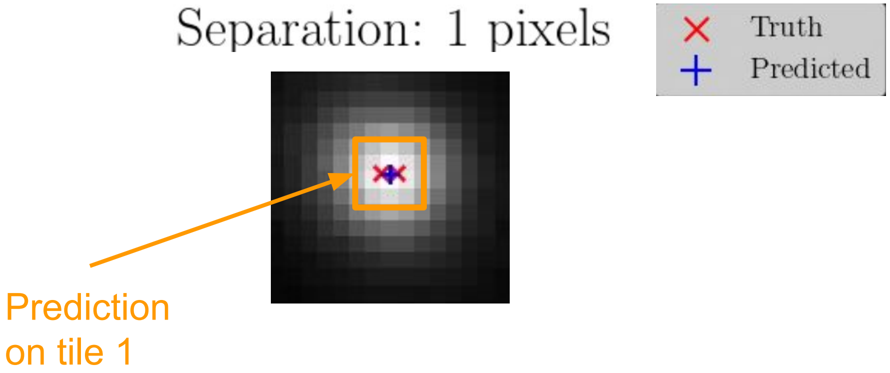
BLISS case study on separation

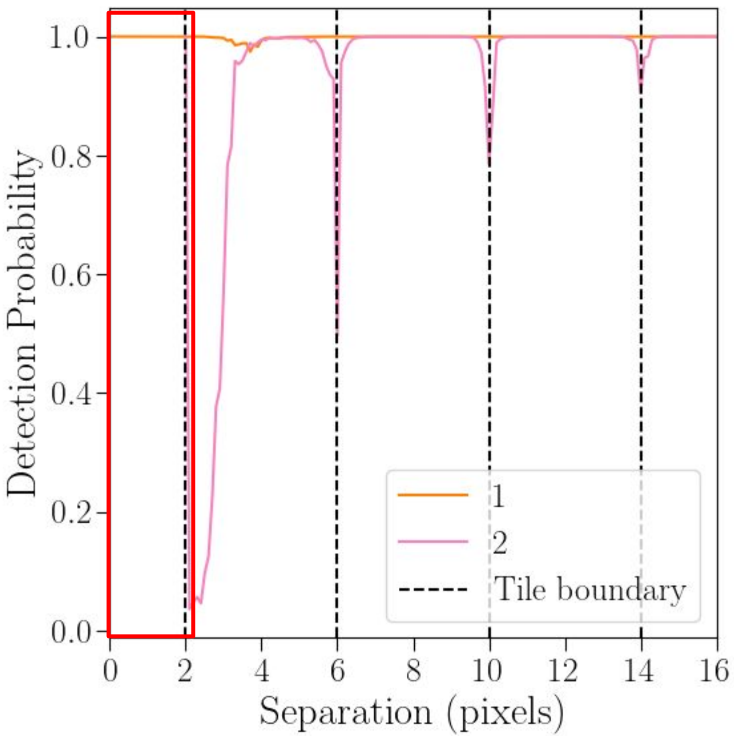
- Galaxy 1 has double the flux of Galaxy 2 and is 50% bigger.



Let’s try to understand BLISS prediction’s step by step at each separation

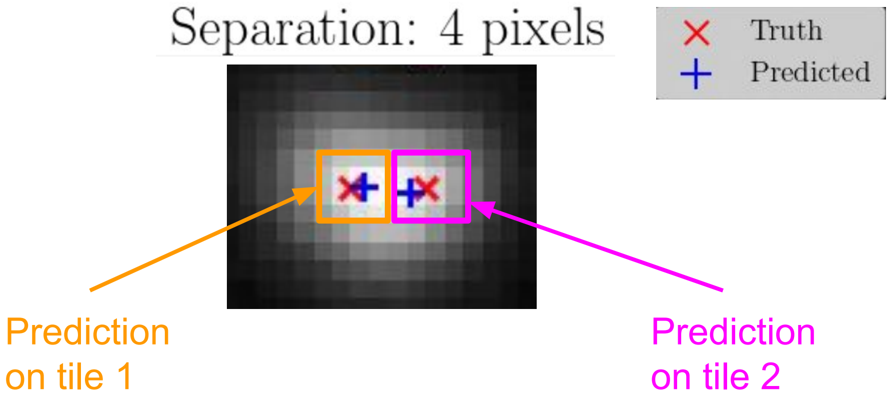
Detection probability as a function of distance





- Initially, the centroid of both galaxies are in the same (central) tile.
- BLISS assumes there is at most 1 object per tile. So only one prediction is made.
- BLISS is highly confident of a galaxy being present in this tile.

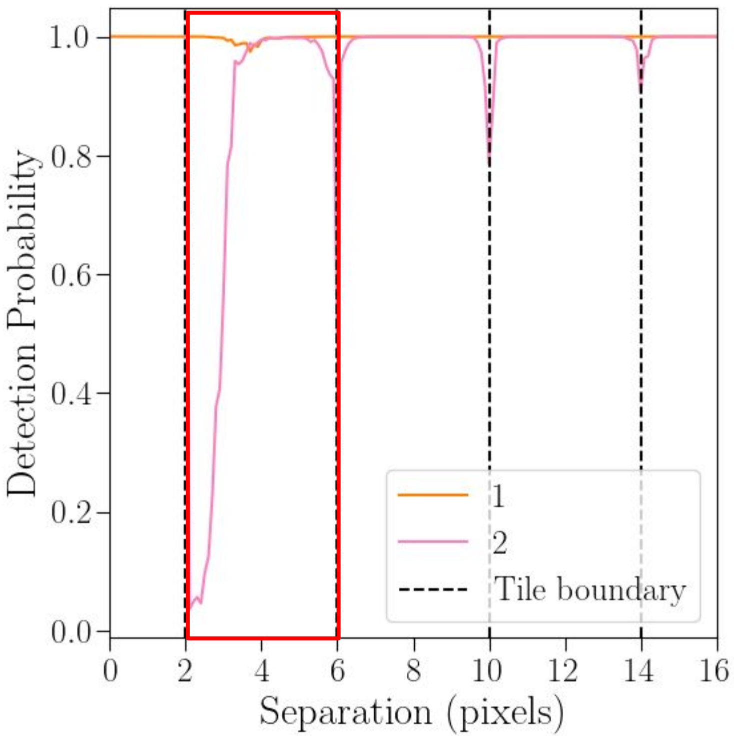
Detection probability as a function of distance



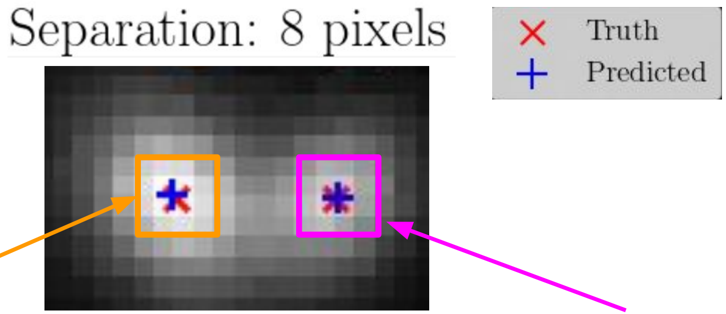
- Once the second galaxy enters the adjacent tile, BLISS outputs prediction at both tiles.

- Initially, both galaxies are highly blended and BLISS does not detect a second galaxy.

- Probability becomes > 50% at separation > 3 pixels.



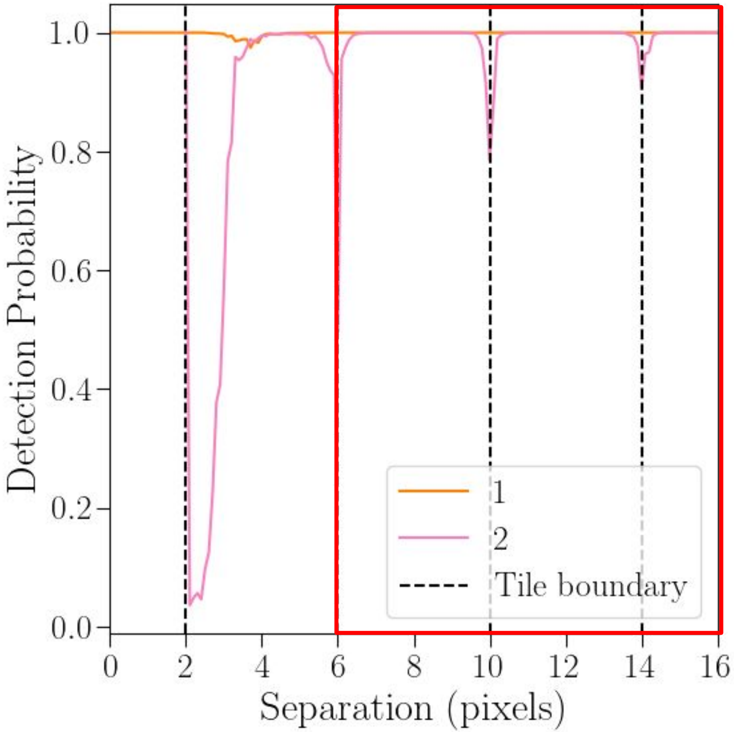
Detection probability as a function of distance



Prediction on tile 1

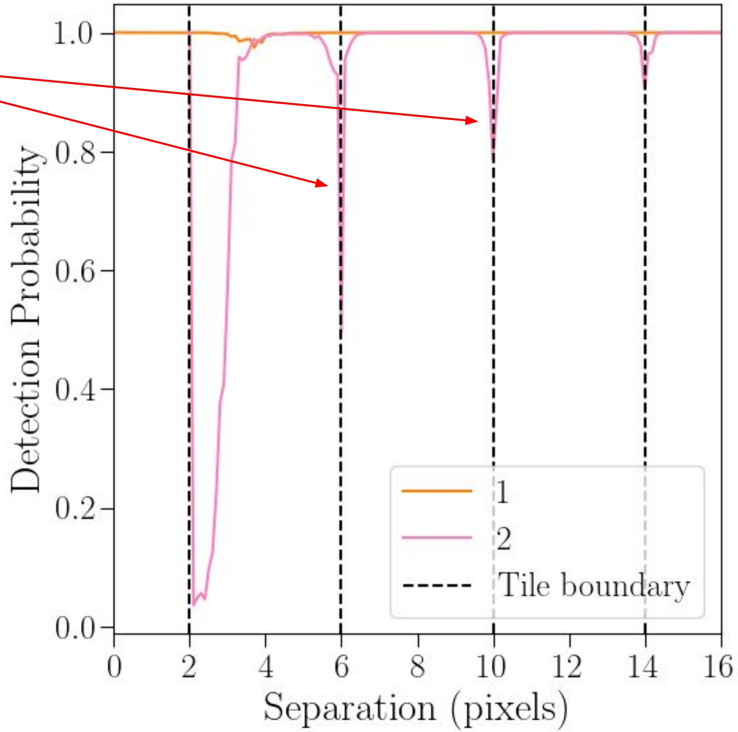
Prediction on tile 2

- As galaxies get farther apart, detection of each individual light sources becomes easier and both detection probabilities converge to 1.

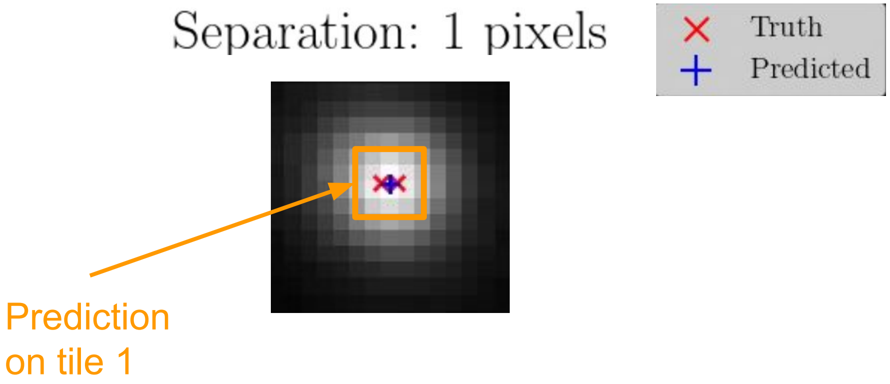


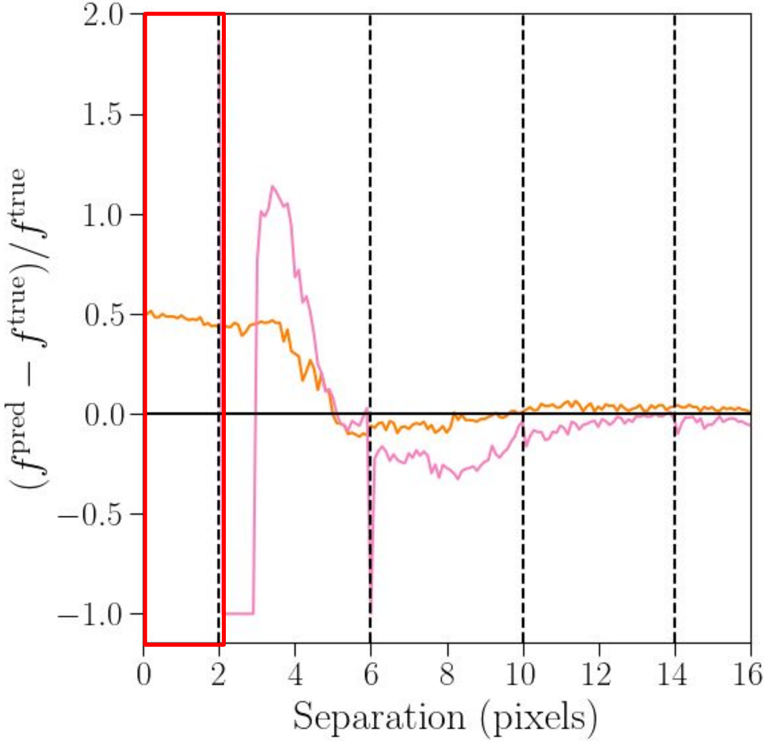
Detection probabilities at tile boundaries

- Detection probability has **sharp valleys** when centroid of galaxy lands close to tile boundaries.
- This is an artifact of our loss **not** being “symmetric” w.r.t tile boundaries.
- Further investigation in future work.



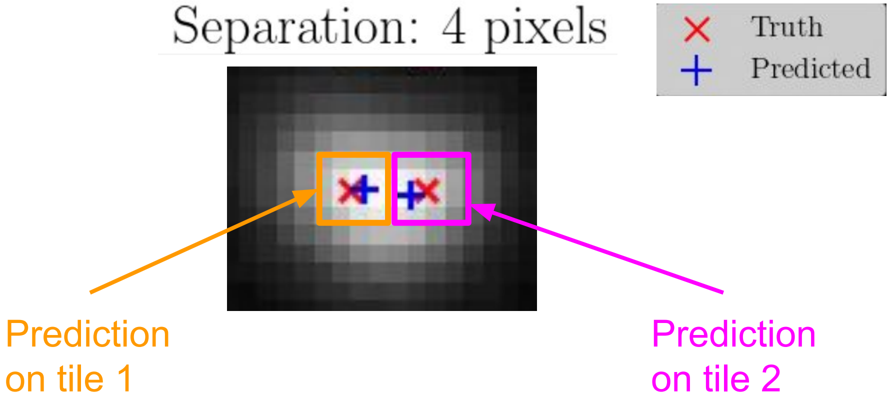
Flux reconstruction residual as a function of distance

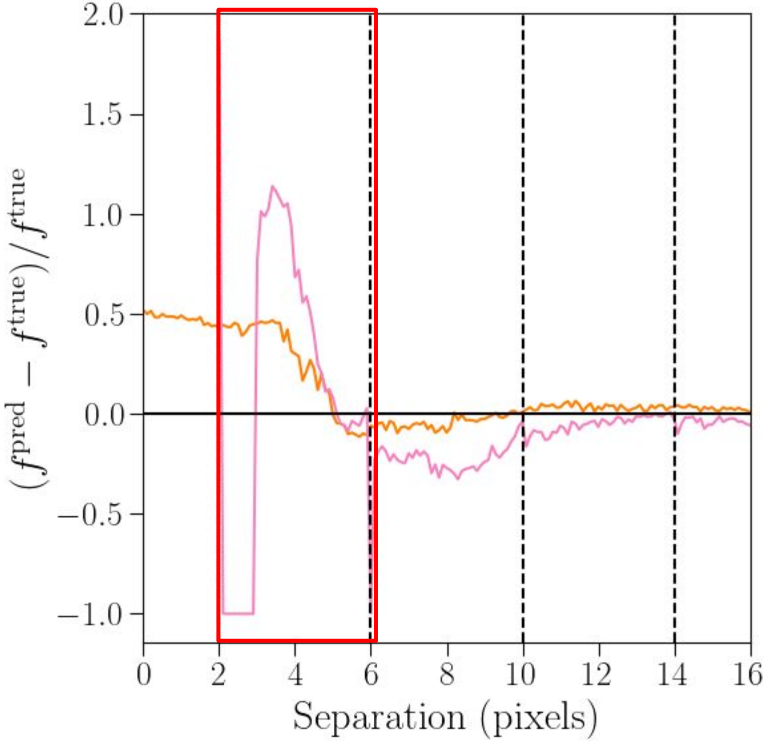




- Initially, the centroid of both galaxies are in the same (central) tile.
- Flux reconstructed for both galaxies at the central tile is 50% bigger than Galaxy 1's flux. As expected.

Flux reconstruction residual as a function of distance

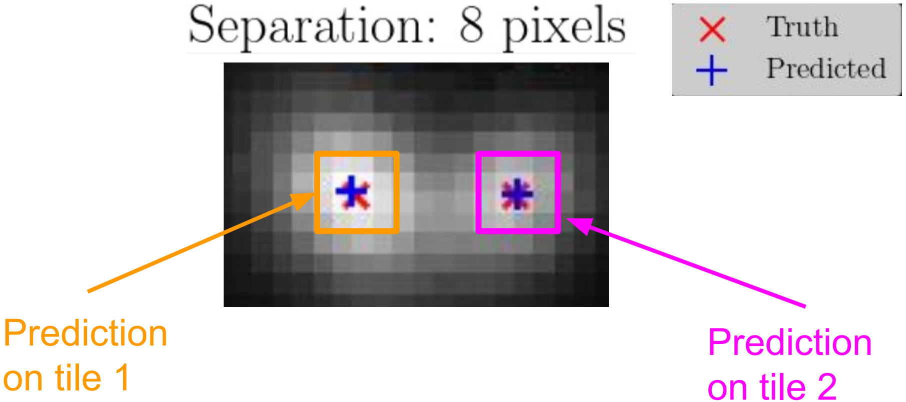


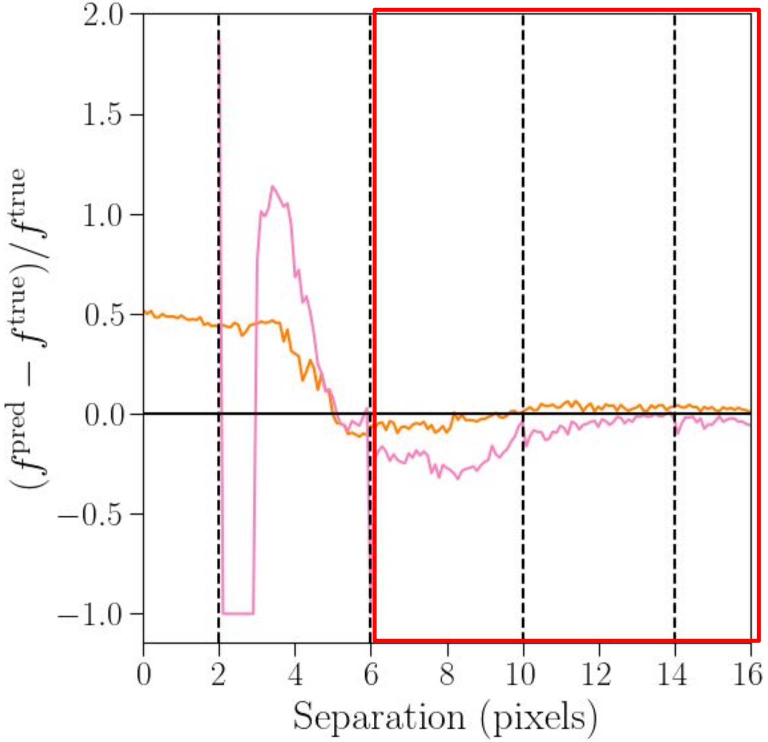


- Flux residual predicted for Galaxy 1 (left) slowly drops to 0 as the Galaxy 2 (right) separates away.

- Galaxy 2's flux prediction is initially 0 as BLISS thinks there is no source in right tile. As Galaxy 2 further enters the right tile, flux is over-predicted due to high blending.

Flux reconstruction residual as a function of distance

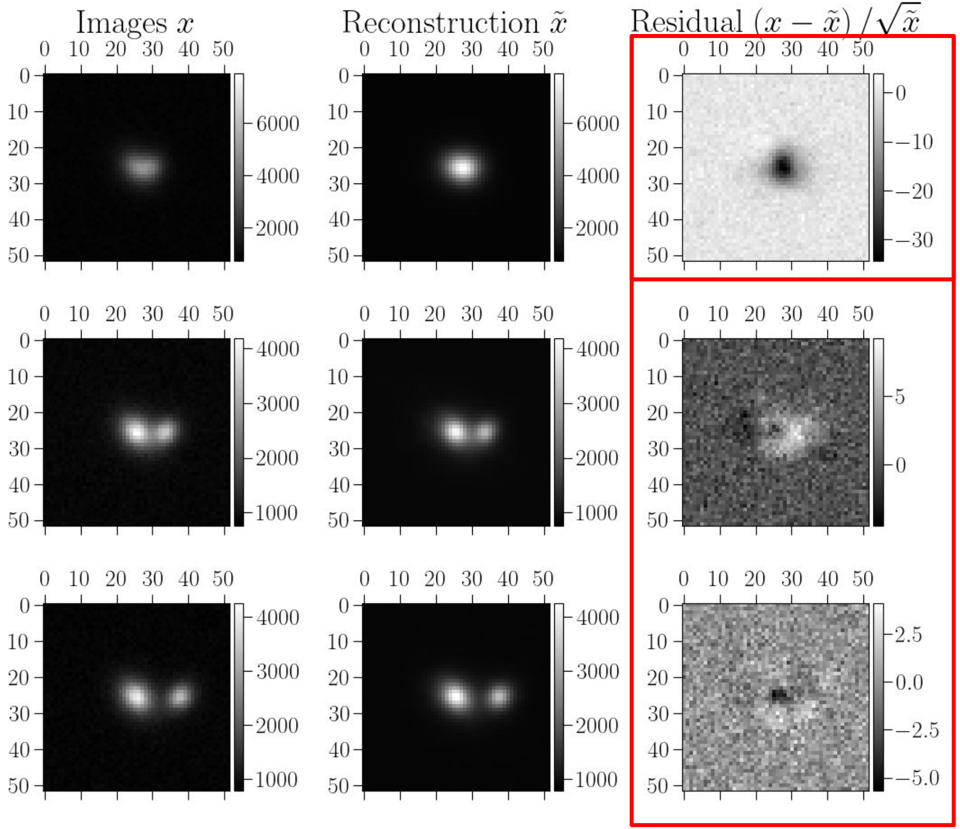




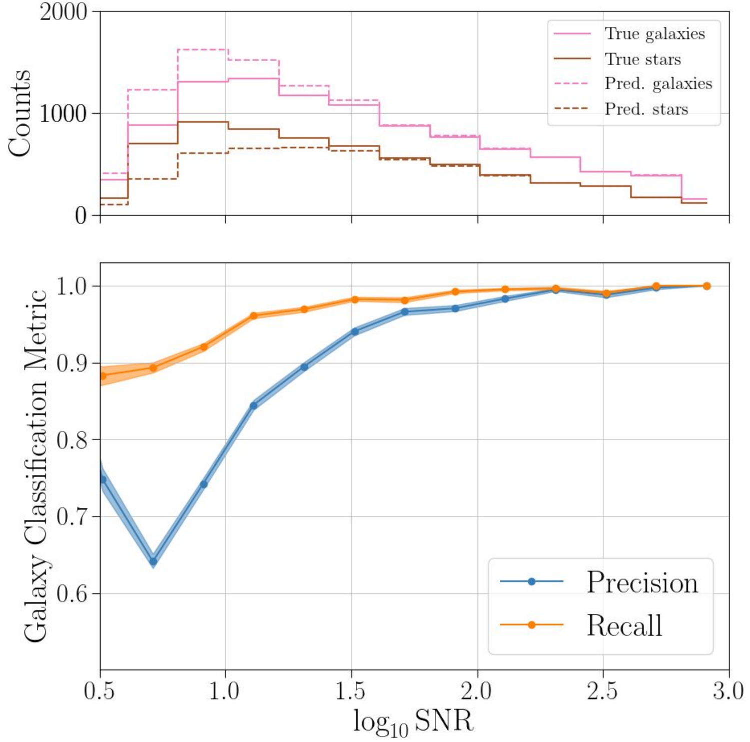
- As galaxies get farther apart, flux residuals tend to 0.
- Overall under-estimation of flux for moderate blending (6-10 pixel separation)

BLISS reconstructions of blends at different degrees of separation

- BLISS has difficulty reconstructing blends < 4 pixels apart.
- Doesn’t preserve total flux for extreme blends.
- As separation grows, residual improves, as expected.



BLISS point-estimate **classification** metrics on 10k **blend dataset**



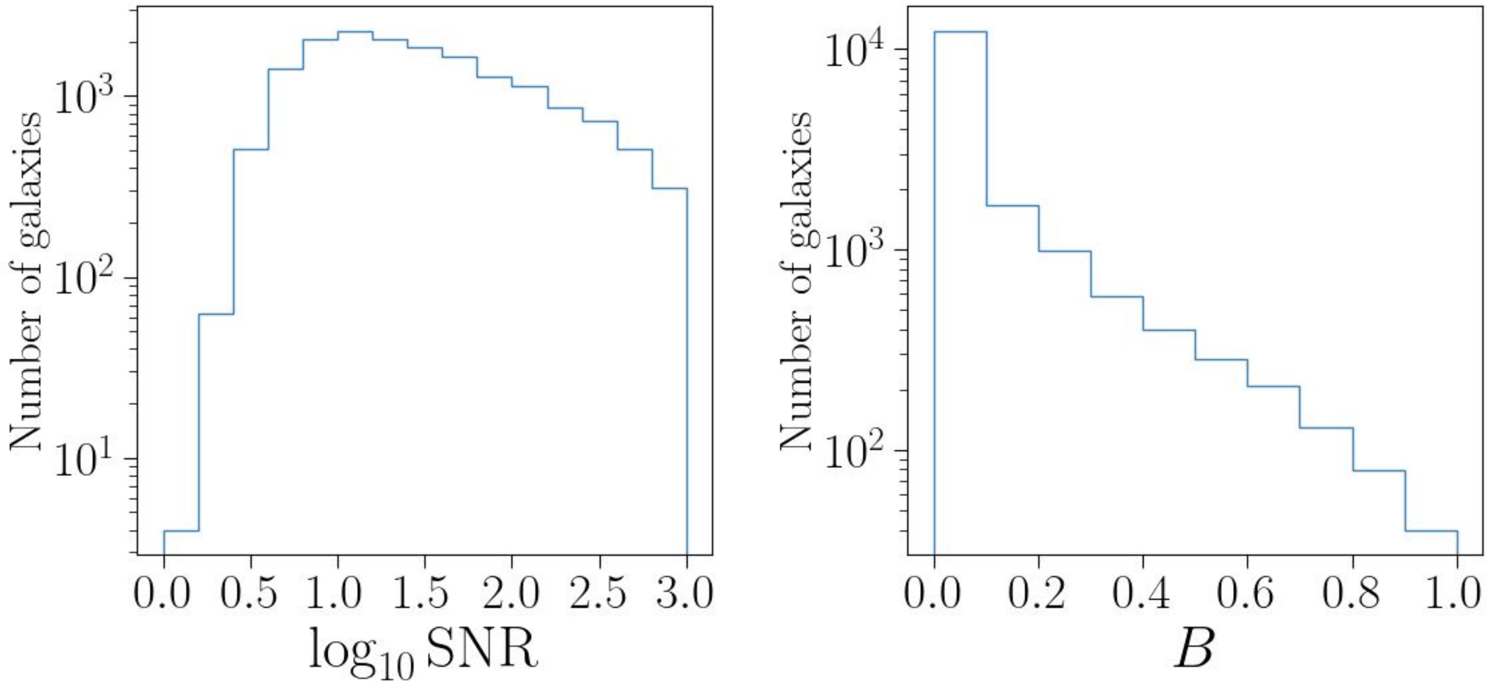
Precision =
\# Matched true galaxy classified as galaxy /
\# Matched sources classified as galaxies

Recall =
\# Matched true galaxies classified as galaxy /
\# True matched galaxies

Matches are detected (>50%** detection probability) galaxy/star centroids less than 1 pixel way from true centroid.

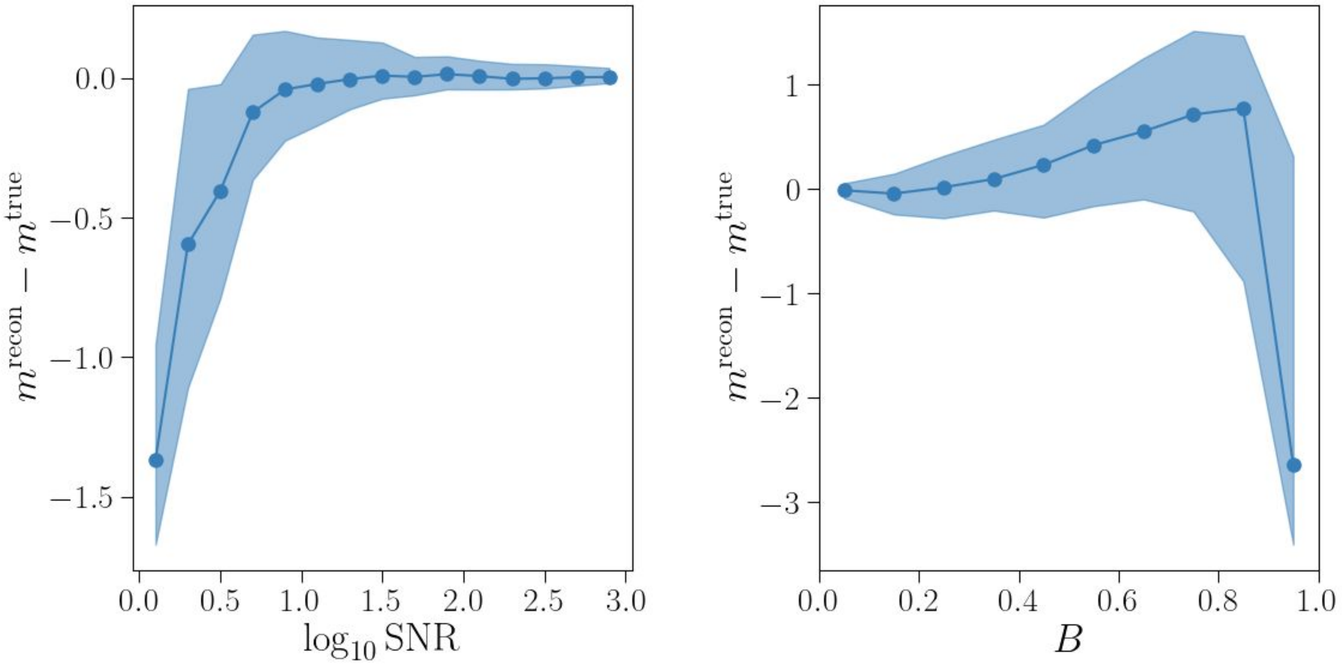
Takeaway: BLISS can disambiguate stars and galaxies with high accuracy for sources > 7 SNR.

10k Blends Histograms of matched objects



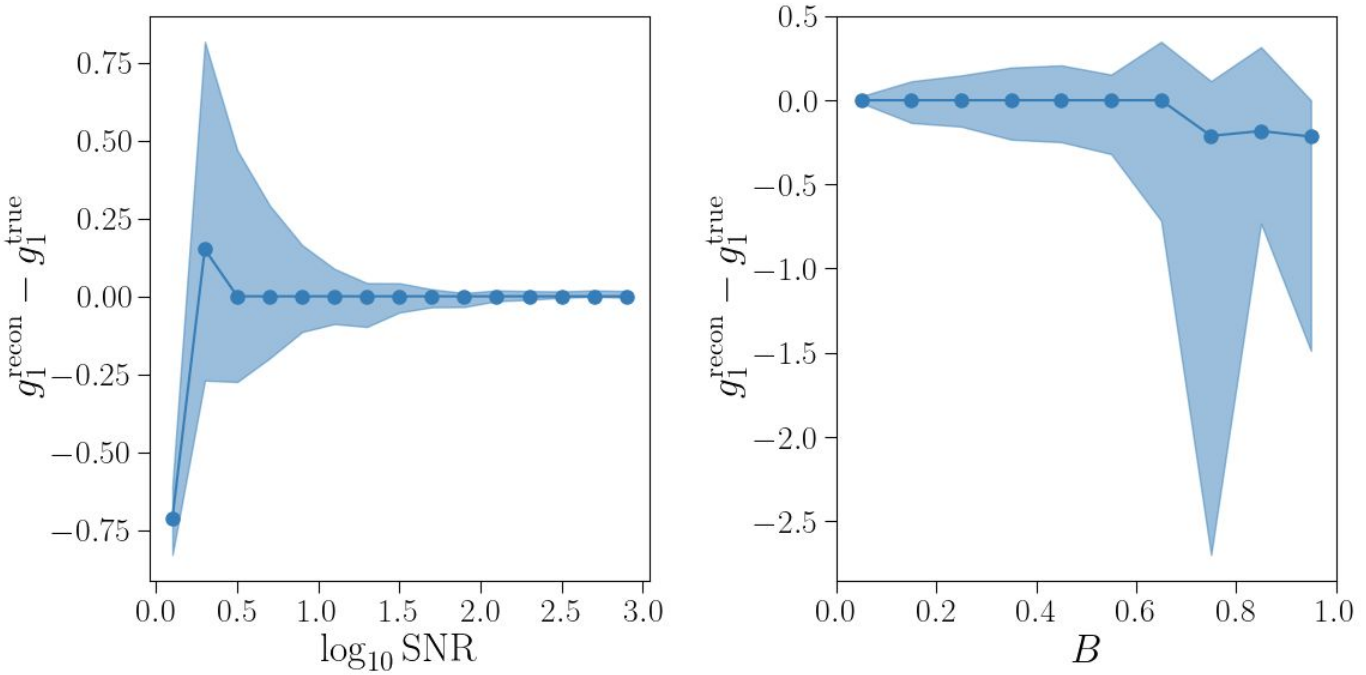
Residuals on measured properties will be noisy with highly blended and low SNR objects. It's hard to detect and thus match them.

10k Blends measurement residuals on matched true galaxies:
Flux residuals



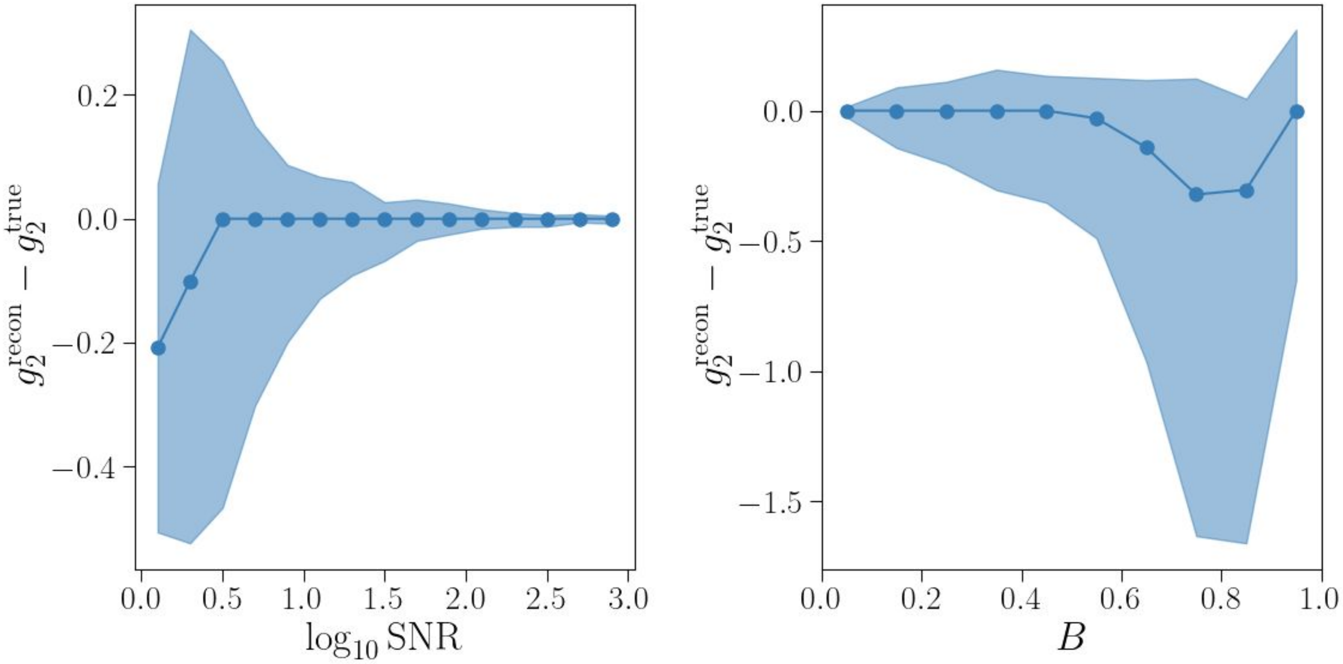
BLISS measured flux residuals on galaxy+star blends are consistent with 0 for matched objects over 7 SNR and 0.8 Blendedness

10k Blends measurement residuals on matched true galaxies:
Ellipticity residuals



BLISS residuals on ellipticities are consistent with 0 for matched objects (but low SNR and high B regions are noisy).

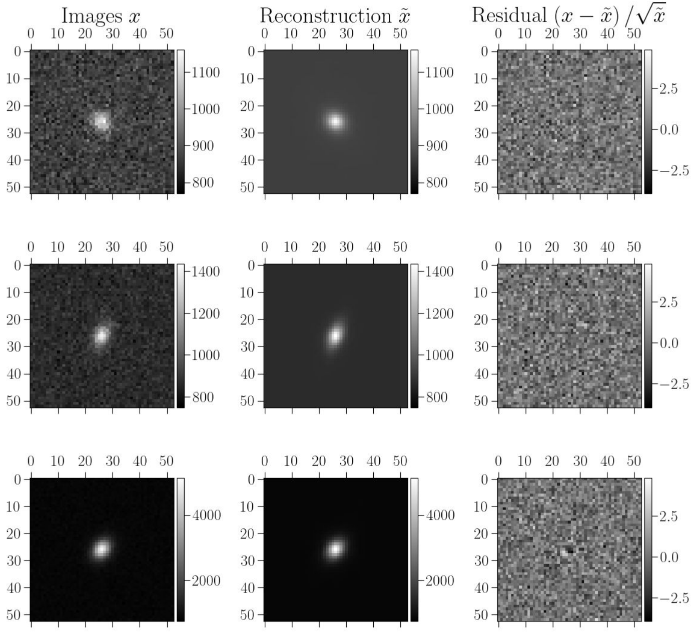
10k Blends measurement residuals on matched true galaxies:
Ellipticity residuals



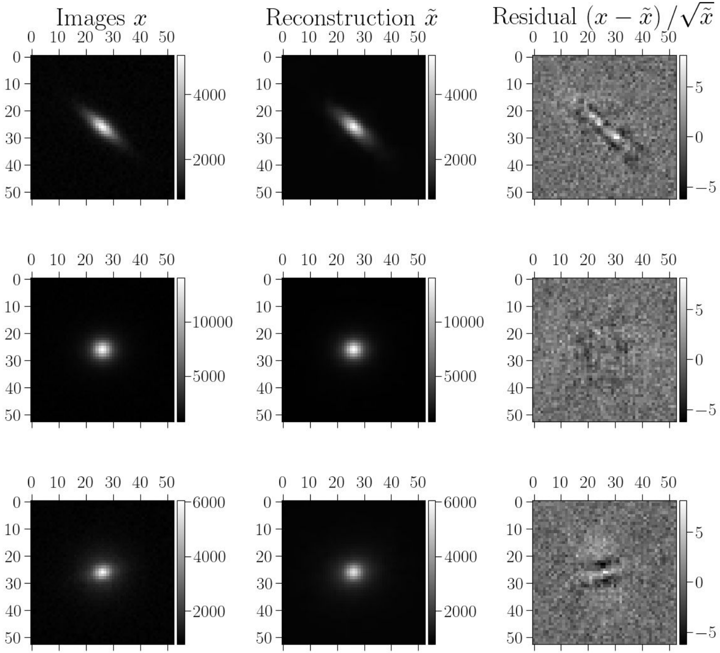
BLISS residuals on ellipticities are consistent with 0 for matched objects (but low SNR and high B regions are noisy).

BLISS can capture isolated galaxies

Random examples

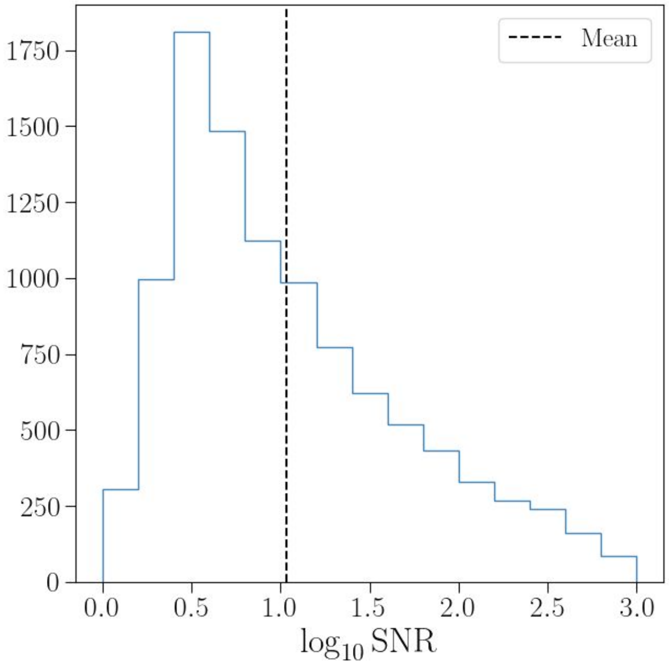


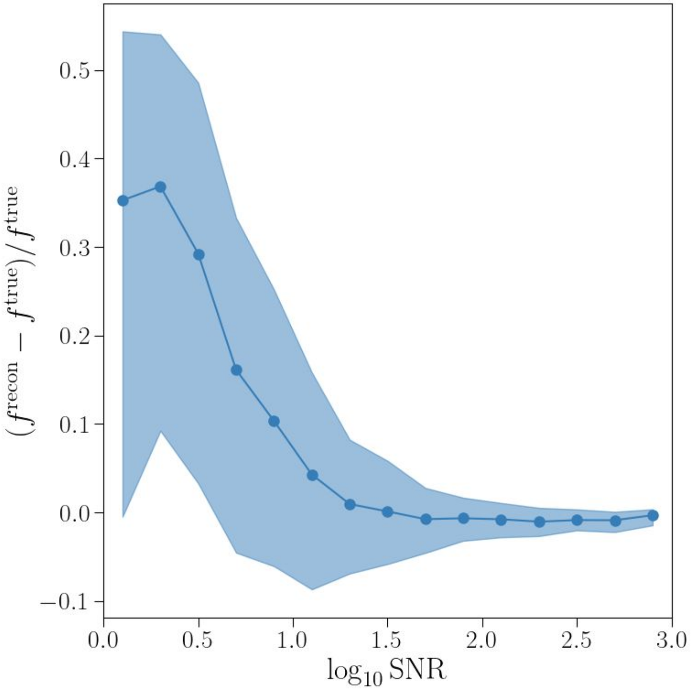
"Worst" residuals



Autoencoder galaxy model residuals are consistent with Gaussian noise for majority of examples.

BLISS can capture isolated galaxy **fluxes**

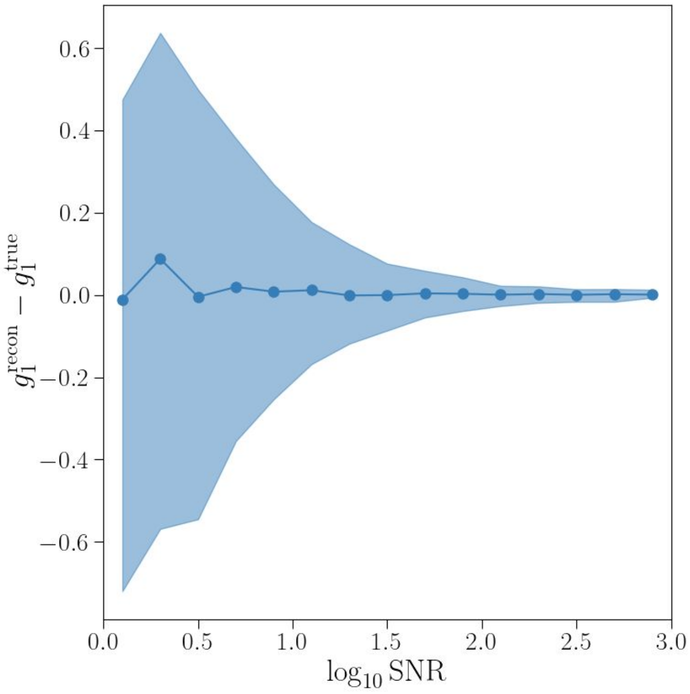


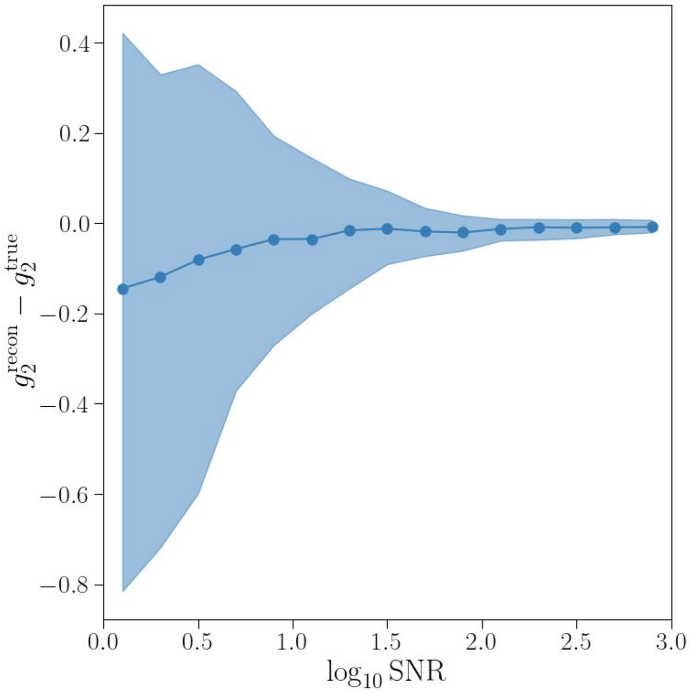


Autoencoder galaxy model reasonably captures galaxy fluxes for SNR > 7.

Bias on flux residuals of lower SNR objects reflects choice of realistic flux prior.

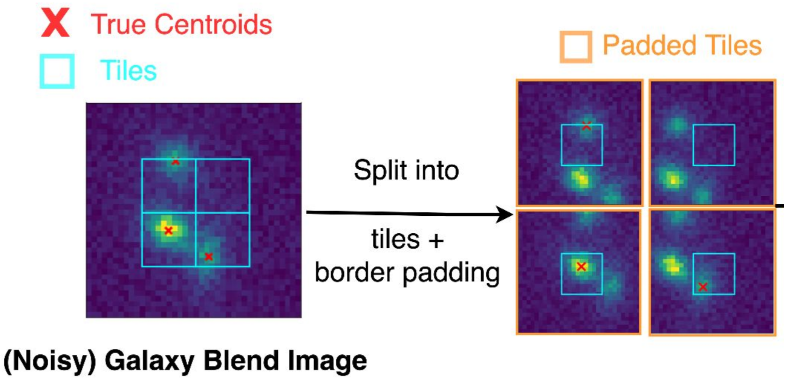
BLISS can capture isolated galaxy **shapes**



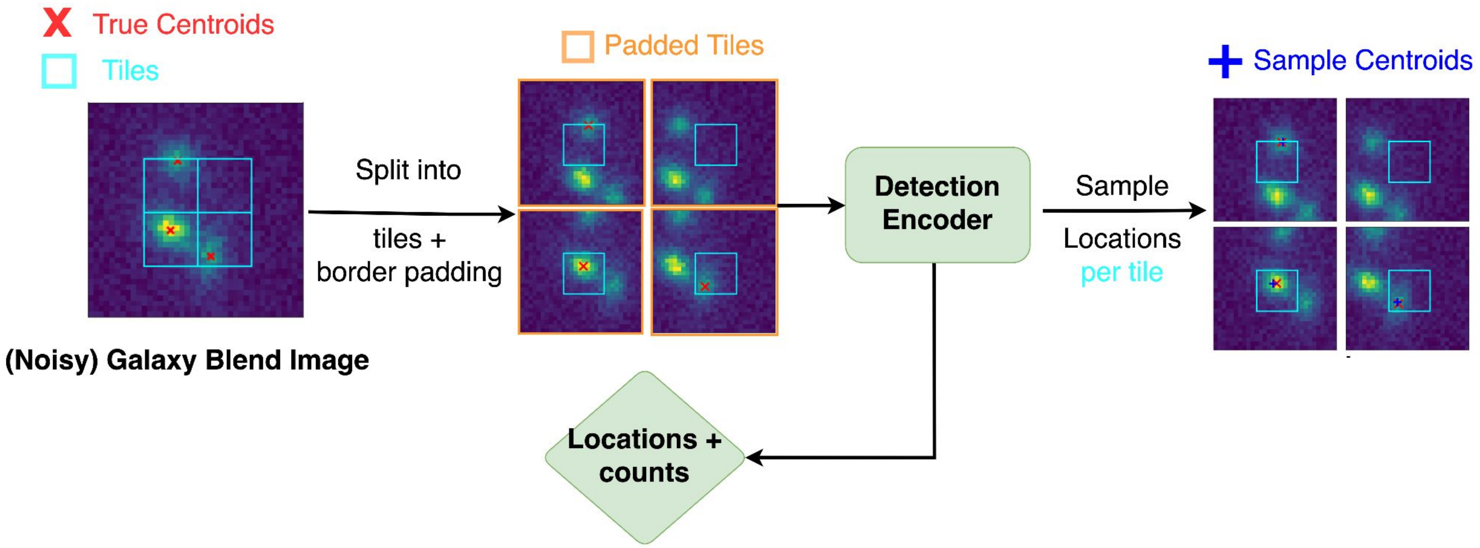


Autoencoder model reasonably captures galaxy shapes.

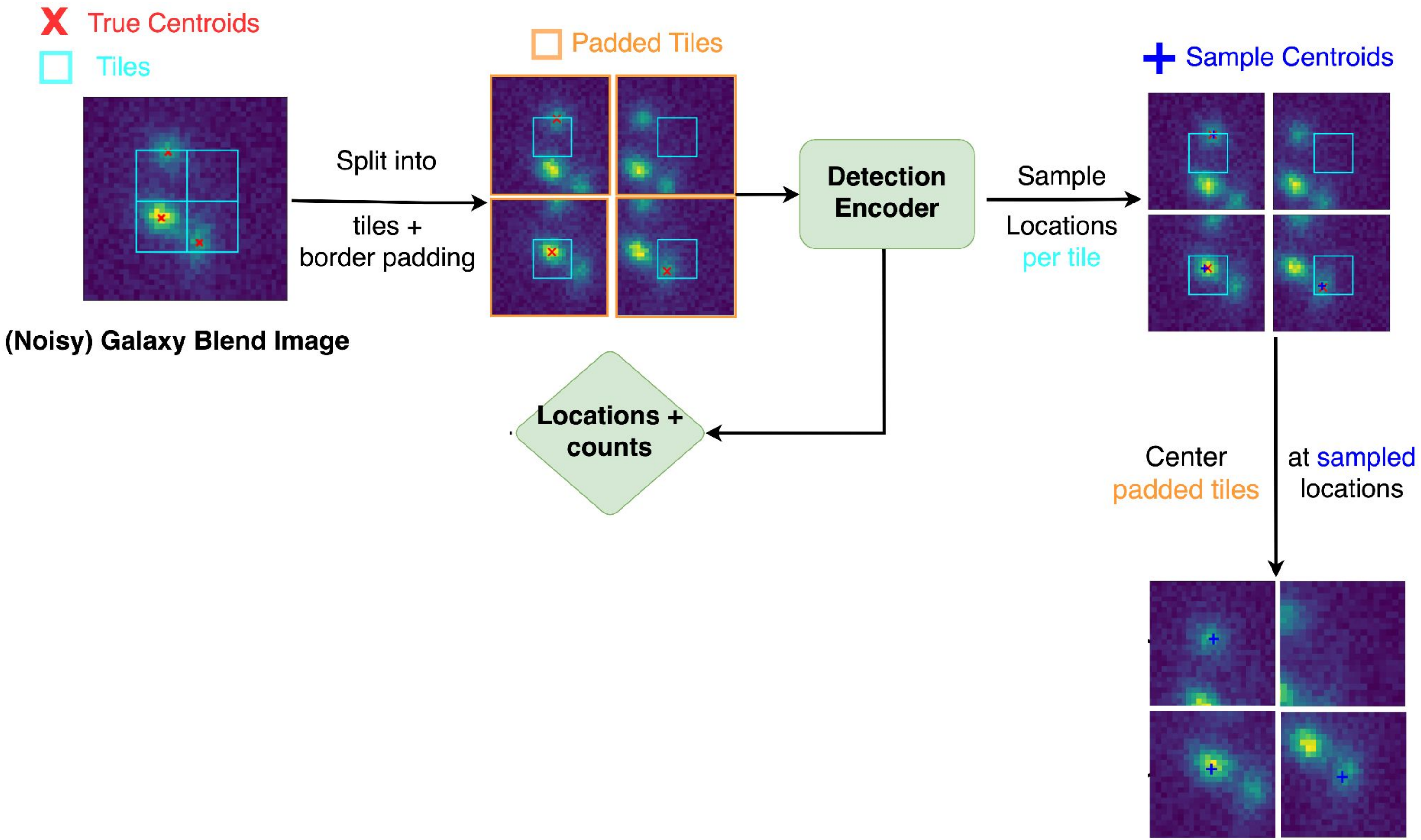
Outline Walkthrough: Prediction

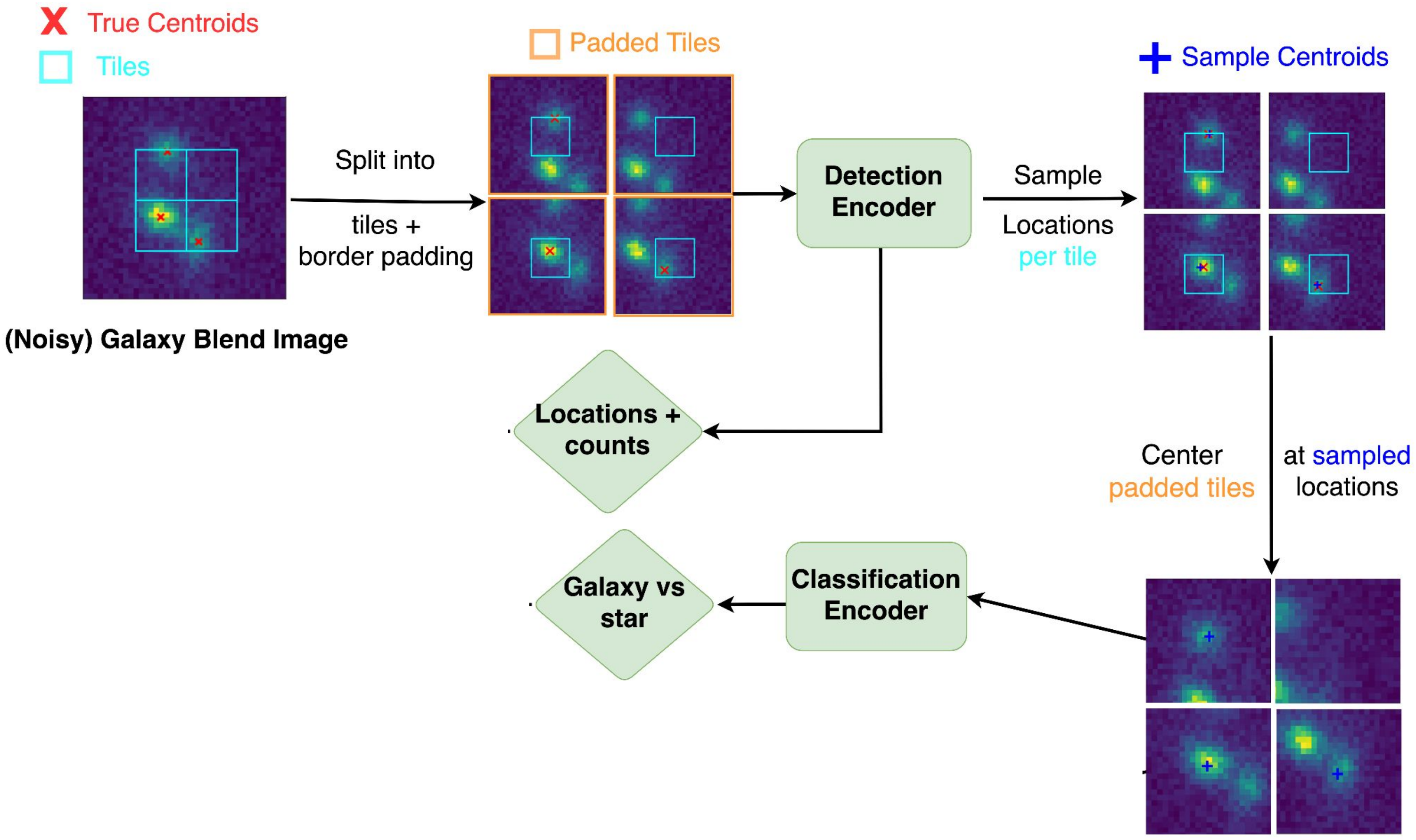


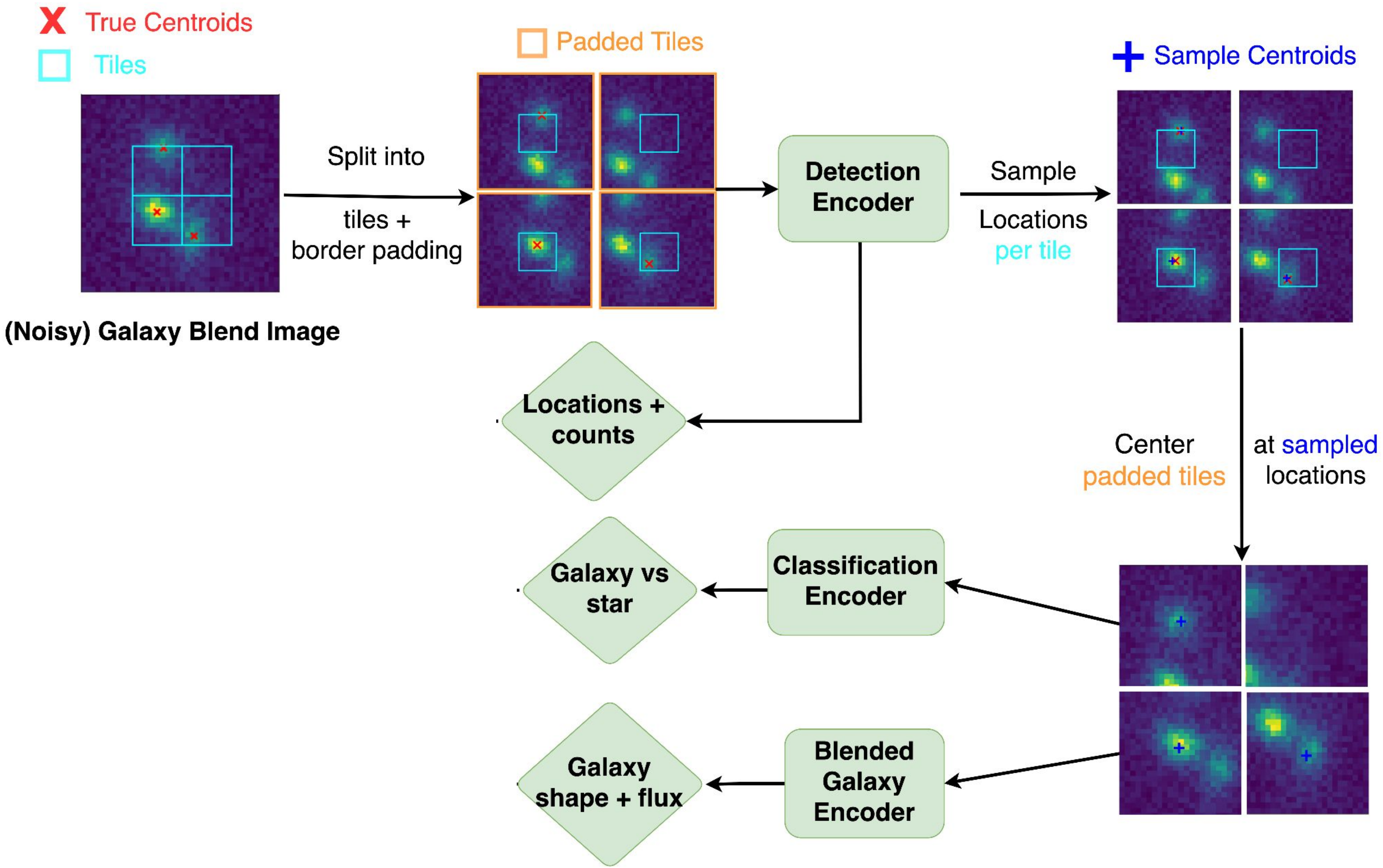
***Note:** Tiles are much smaller in actuality.

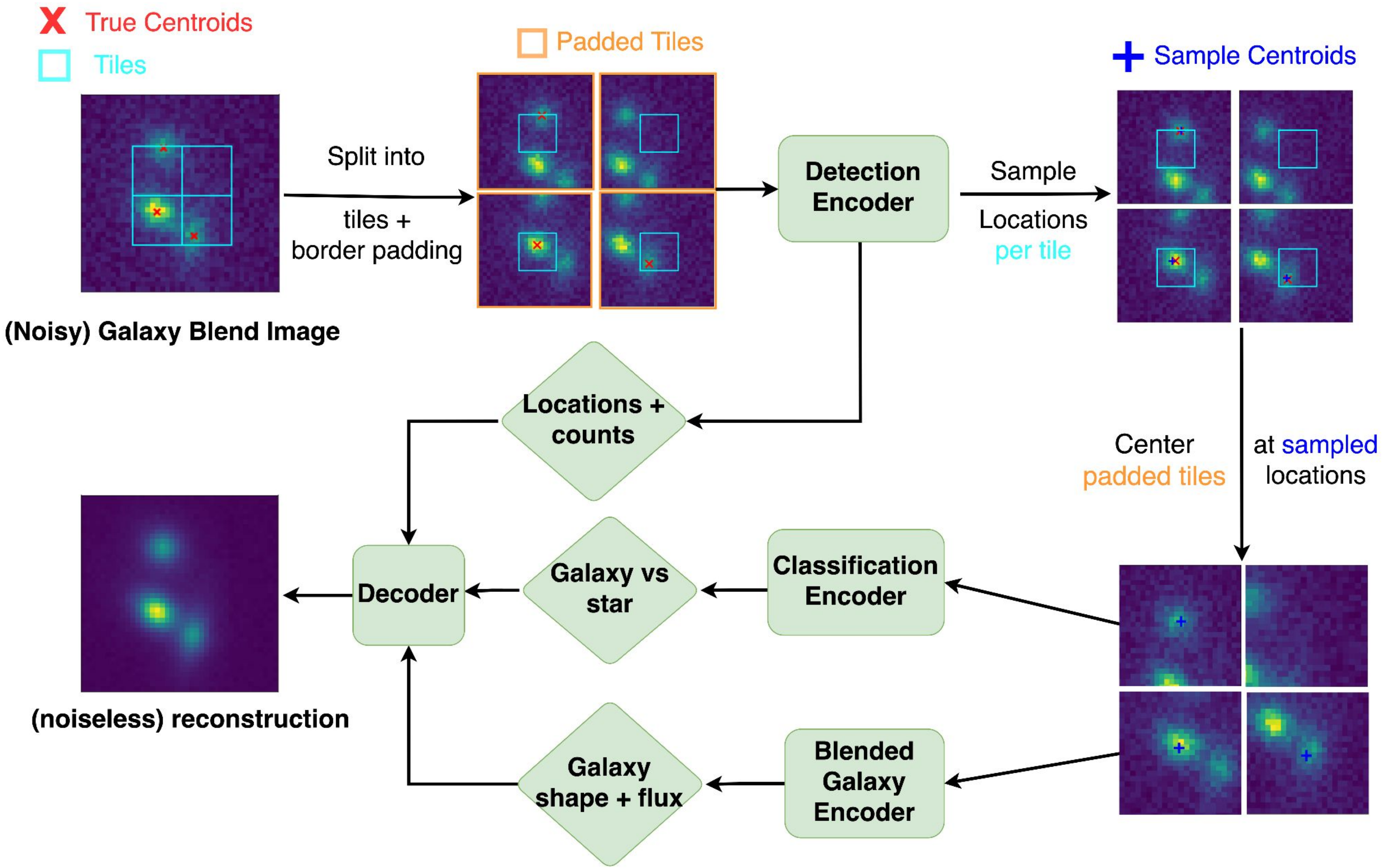


True Centroids
Tiles
(Noisy) Galaxy Blend Image
Split into
tiles +
border padding
Padded Tiles
Detection Encoder
Locations + counts
Sample
Locations
per tile
Sample Centroids


True Centroids
Tiles
(Noisy) Galaxy Blend Image
Split into
tiles +
border padding
Padded Tiles
Detection
Encoder
Locations +
counts
Sample
Locations
per tile
Sample Centroids
Center
padded tiles
at sampled
locations


True Centroids
Tiles
Padded Tiles
Sample Centroids
Split into
tiles +
border padding
(Noisy) Galaxy Blend Image
Detection
Encoder
Sample
Locations
per tile
Locations +
counts
Center
padded tiles
at sampled
locations
Classification
Encoder
Galaxy vs
star


True Centroids
Tiles
(Noisy) Galaxy Blend Image
Split into
tiles +
border padding
Padded Tiles
Detection Encoder
Sample
Locations
per tile
Sample Centroids
Locations + counts
Center
padded tiles
at sampled
locations
Classification Encoder
Galaxy vs star
Blended Galaxy Encoder
Galaxy shape + flux



True Centroids
Tiles
Padded Tiles
Sample Centroids
Split into
tiles +
border padding
Detection Encoder
Sample
Locations
per tile
(Noisy) Galaxy Blend Image
Locations + counts
Center
padded tiles
at sampled
locations
Decoder
Galaxy vs star
Classification Encoder
(noiseless) reconstruction
Galaxy shape + flux
Blended Galaxy Encoder

Statistical Methods

FAVI for detection and classification

In this context of variational inference, we choose to approximate the posterior in eq. (11) with a *variational distribution* that factorizes over tiles $Q_\phi(\mathcal{Z})$:

$$Q_\phi(\mathcal{Z}) = \prod_{t=1}^{T} Q_\phi(\mathcal{Z}_t), \tag{14}$$

We explicitly define the variational distribution $Q_\phi(\mathcal{Z}_t) = Q_\phi(n_t, b_t, f_t, \boldsymbol{\ell}_t)$ on each tile t as:

$$n_t \sim \text{Bernoulli}\left(\omega_{n,t}\right) \tag{15}$$

$$b_t | n_t = 1 \sim \text{Bernoulli}\left(\omega_{g,t}\right) \tag{16}$$

$$\boldsymbol{\ell}_t | n_t = 1 \sim \text{LogitNormal}\left(\mu_{\ell,t}, \text{diag}\left(\nu_{\ell,t}\right)\right) \tag{17}$$

$$f_t | n_t = 1, b_t = 1 \sim \text{LogNormal}\left(\mu_{f,t}, \sigma^2_{f,t}\right) \tag{18}$$

with the so-called *variational parameters*

$$\theta_t = \{\omega_{n,t}, \omega_{g,t}, \boldsymbol{\mu}_{\ell,t}, \text{diag}(\boldsymbol{\nu}_{\ell,t}), \mu_{f,t}, \sigma^2_{f,t}\}. \tag{19}$$

FAVI for detection and classification

We choose to target the following loss function introduced in Ambrogioni et al. (2019) to find the set of optimal weights ϕ for our neural networks characterizing the variational distributions:

$$L(\phi) = \mathbb{E}_{\mathcal{X}\sim P(\mathcal{X})} D_{\mathrm{KL}}\big(P\,(\mathcal{Z}|\mathcal{X})\;\|\,Q_{\phi}\,(\mathcal{Z})\big). \tag{20}$$

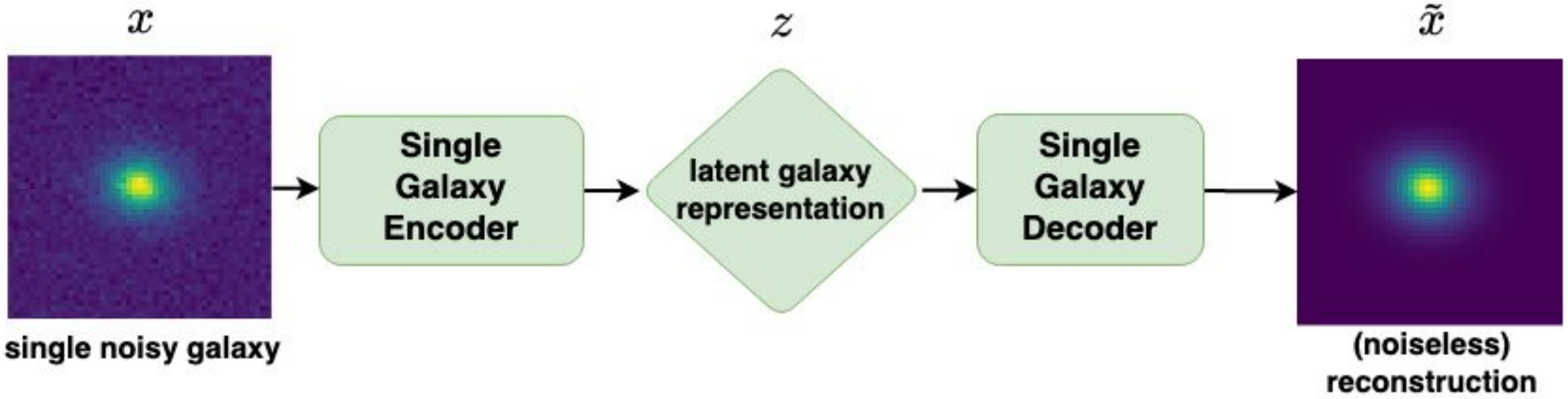
$D_{\mathrm{KL}}(P\,\|\,Q)$ is the *KL divergence* between distributions P and Q defined as:

$$D_{\mathrm{KL}}(P\,\|\,Q) \equiv \int P\,(z) \log\left(\frac{Q\,(z)}{P\,(z)}\right) dz \tag{21}$$

Using a specific factorization for the variational distribution we can rewrite the loss function from eq. (20) as:

$$\begin{aligned} L(\phi) = \mathbb{E}_{\mathcal{X},\mathcal{Z}\sim P(\mathcal{X},\mathcal{Z})} \sum_{t=1}^{T} &\log Q_{\phi}(n_t) + \log Q_{\phi}(\boldsymbol{\ell}_t|n_t) \\ &+ \log Q_{\phi}(f_t|n_t,\boldsymbol{\ell}_t,b_t) + \log Q_{\phi}(b_t|n_t,\boldsymbol{\ell}_t) \end{aligned} \tag{22}$$

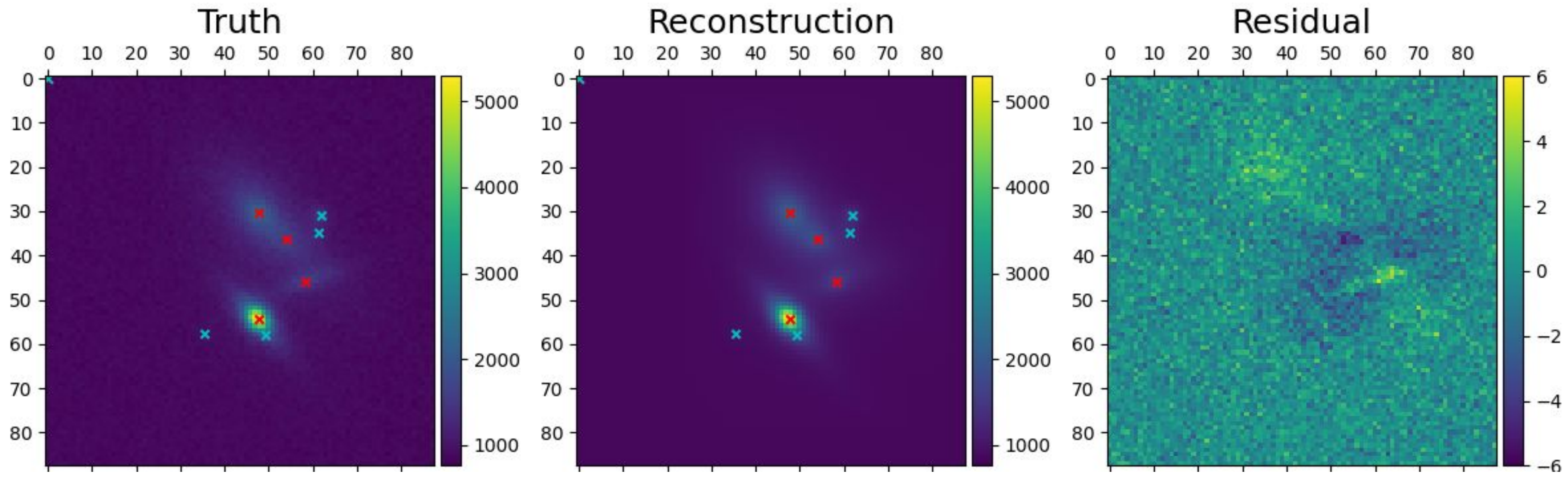
Autoencoder for galaxy modeling and deblending



$$L(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = -\sum_{p \in \text{pixels}} \log \mathcal{N}\left(x_p \,;\, \tilde{x}_p, \tilde{x}_p\right)$$

Simulated galaxy blend reconstruction

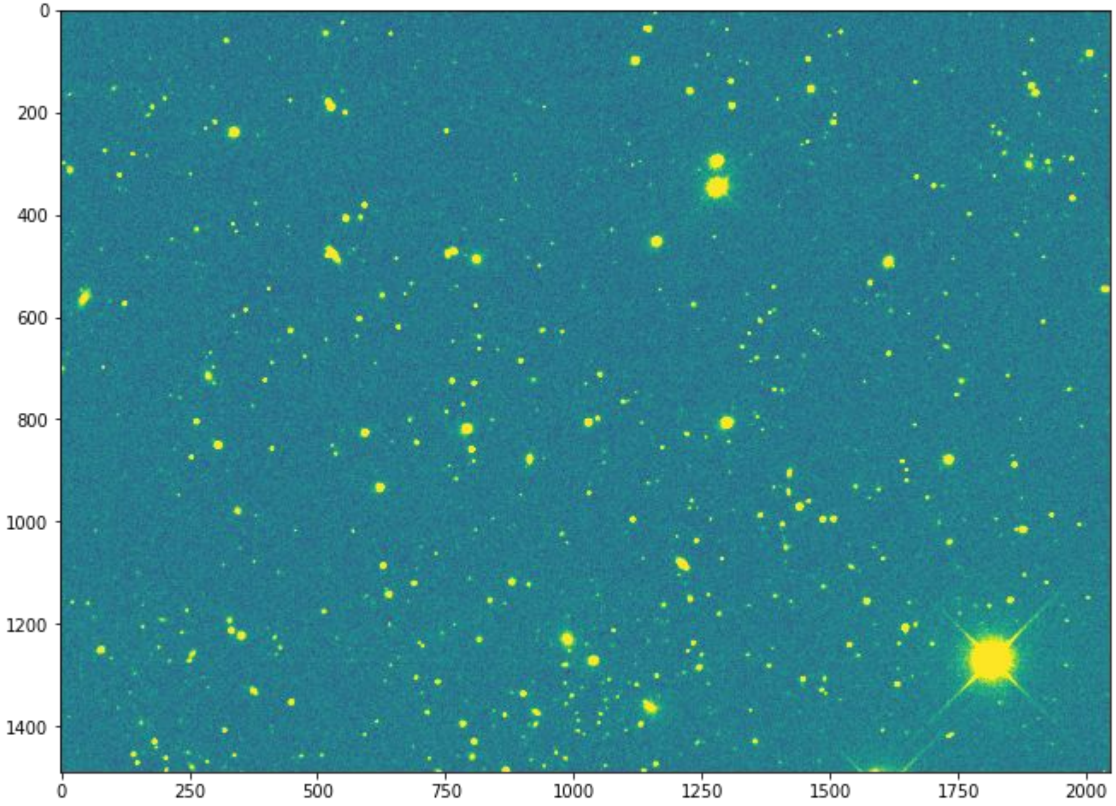
Galaxy
Star



Examples of BLISS output on real data

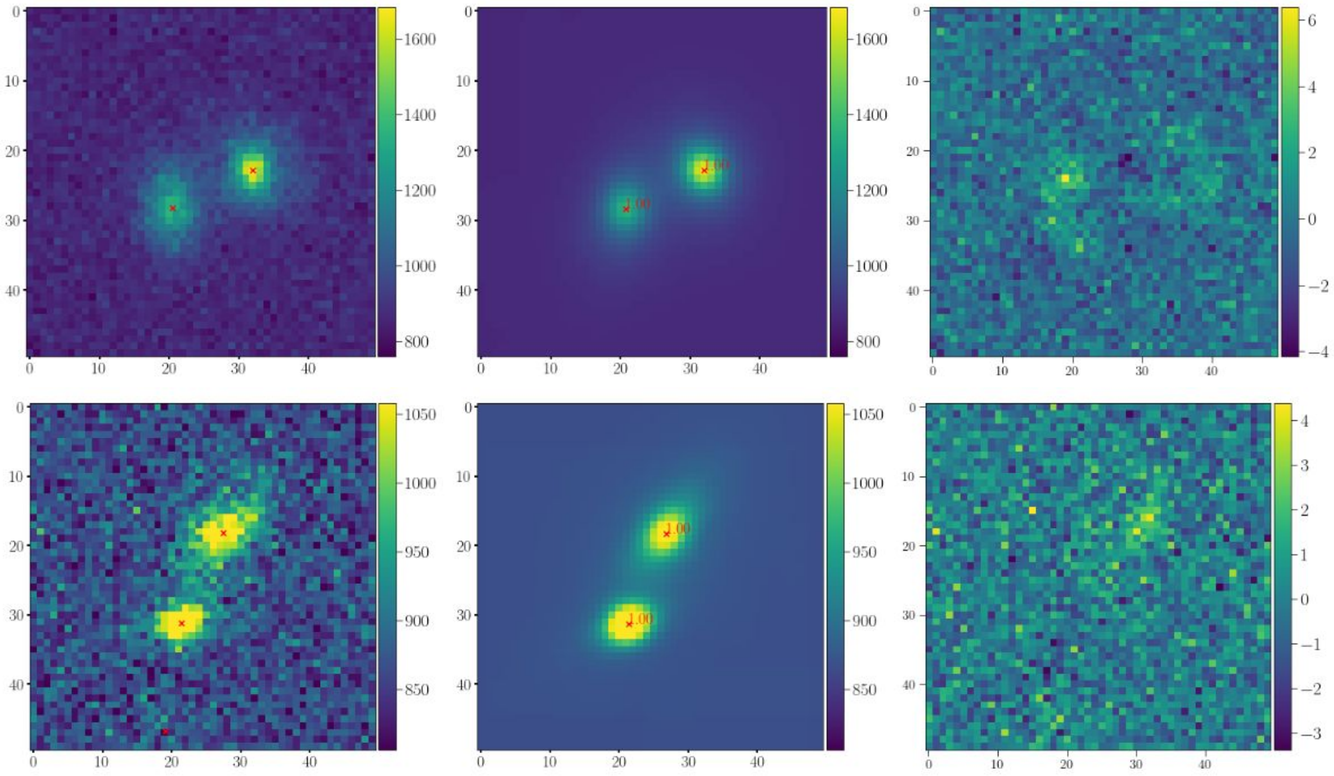
Fast inference on large survey scenes

- SDSS frame 1500x2100 inference (locations, galaxy latents, …) ~**20s on a gpu** after training.



Reconstruction of individual real galaxy blends from SDSS

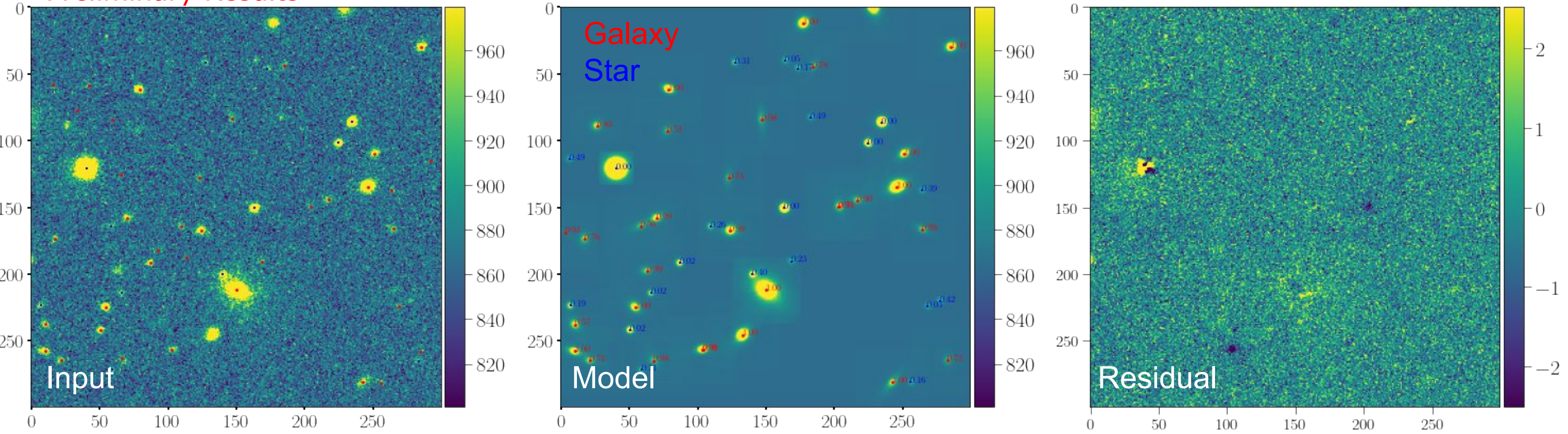
Preliminary Results



Extensible to large frames: Reconstruction of an SDSS frame

- We are able to perform inference with BLISS on a 300x300 pixel chunk of an SDSS frame in **~1s on a gpu.**

Preliminary Results



What's next for BLISS?