# An ML Framework for Estimating Bayesian Posteriors of Galaxy Morphological Parameters

**Aritra Ghosh***
Yale Center for Astronomy and Astrophysics, and Department of Astronomy,
Yale University, New Haven, CT, USA
aritra.ghosh@yale.edu

**Meg Urry**
Yale Center for Astronomy and Astrophysics, and Department of Physics,
Yale University, New Haven, CT, USA

**Amrit Rau**
Department of Computer Science, Yale University
New Haven, CT, USA

**Laurence Perreault-Levasseur**
Department of Physics, Univesité de Montréal, Montréal, Canada
Center for Computational Astrophysics, Flatiron Institute, New York, NY, USA

## Abstract

Galaxy morphology is connected to various fundamental properties of a galaxy and studying the morphology of large samples of galaxies is central to understanding the relationship between morphology and the physics of galaxy formation & evolution. For the first time, we are able to use machine learning to estimate Bayesian posteriors for galaxy morphological parameters. To achieve this, GAMPEN, our machine learning framework, uses a spatial transformer network (STN), a convolutional neural network, and the Monte-Carlo Dropout technique. This novel application of an STN in astronomy also enables GAMPEN to crop out most secondary galaxies in the frame and focus on the galaxy of interest. We also demonstrate that by first training on simulations and then performing transfer learning using real data, we are able to achieve excellent estimates for morphological parameters of galaxies in the Hyper Suprime-Cam Wide survey, while using only a small amount of real training data.

## 1 Introduction

For almost a century, starting with Edwin Hubble's work in 1926, astronomers have linked the morphology (shape) of galaxies to the physics of galaxy formation and evolution. Morphology has been shown to be related to various fundamental properties of the galaxy and its environment – galaxy mass, star formation rate, stellar kinematics, merger history, cosmic environment, the influence of supermassive black holes, and a range of other physics (Tremaine et al., 2002; Pozzetti et al., 2010; Wuyts et al., 2011; Schawinski et al., 2014; Powell et al., 2017). Studying the morphology of large samples of galaxies at different redshifts (i.e., distances) is crucial in order to understand the physics of galaxy evolution.

*http://www.ghosharitra.com/

Fourth Workshop on Machine Learning and the Physical Sciences (NeurIPS 2021).

Driven by the fact that traditional techniques of determining galaxy morphology (visual inspection and template fitting) are not scalable to the large data volumes expected from future surveys like the Large Synoptic Survey Telescope (LSST), and the Nancy Grace Roman Space Telescope (NGRST), Convolutional Neural Networks (CNNs) have become increasingly popular for determining galaxy morphology (Ghosh et al., 2020; Hausen & Robertson, 2020; Cheng et al., 2021; Vega-Ferrero et al., 2021). However, most previous works have provided only broad morphological classifications, and they require very large training sets of pre-classified galaxies. A few works like Tuccillo et al. (2018) have produced quantitative point-estimates of single morphological parameters, but without associated uncertainties. Computation of Bayesian posteriors is crucial for drawing statistical inferences that account for uncertainty, and thus has been indispensable in deriving robust scaling relations, as well as tests of theoretical models using morphology (Bernardi et al., 2013; van der Wel et al., 2014; Schawinski et al., 2014). Additionally, if CNNs are to replace traditional methods for estimating galaxy morphological parameters in upcoming imaging surveys, there needs to be a framework that does not require a large pre-classified training set from the same survey.

In this work, we present the Galaxy Morphology Posterior Estimation Network (GAMPEN), a machine learning framework that combines a Spatial Transformer Network (STN), a CNN, and the Monte Carlo Dropout (MCD) technique to estimate the posteriors of three different morphological parameters: the bulge-to-total light ratio ($L_B/L_T$), the half-light radius ($R_e$), and the total flux. In order to avoid using a large training set of pre-classified real galaxies, we first trained GAMPEN on realistic simulations of galaxies, and then performed transfer learning on a small amount of real data. Using this technique, we demonstrate below that we are able to obtain robust estimates for the posterior distributions of morphological parameters of $z$ (redshift) $< 0.25$ galaxies in the Hyper-Suprime Cam - Wide (HSC-W) survey (Aihara et al., 2018).

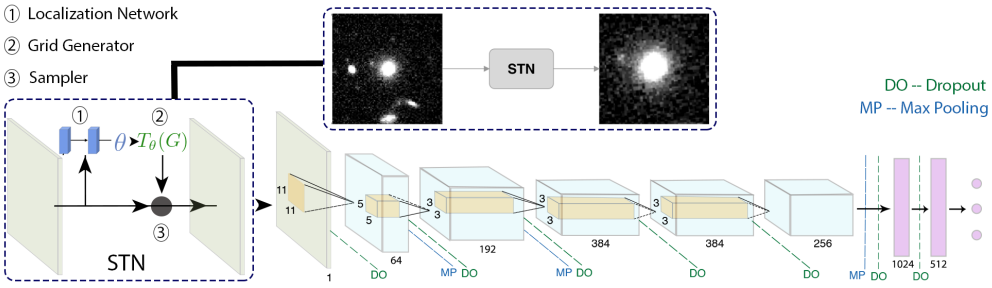


Figure 1: Schematic diagram of GAMPEN. The numbers below each layer refer to the number of filters/neurons and the numbers inside the convolutional layers refer to the kernel sizes of the convolutional layers. An example of the transformation performed by the STN on a Hyper Suprime-Cam cutout is shown in the top inset.

## 2 Description of the Framework

The architecture of GAMPEN is shown in Fig.1 and consists of a CNN preceded by an STN. The design of the CNN is based on the design of GAMORNET (Ghosh et al., 2020) and consists of 5 convolutional layers and three fully connected layers. Interspersed between these are max-pooling and dropout layers. All weight layers use the ReLU activation function except the output layer, which uses a linear activation function. The outputs of the three neurons correspond to the $L_B/L_T$, $R_e$, and total flux of the galaxy fed into the network.

As illustrated in Fig. 1, the STN included in GAMPEN correctly learns to crop out most secondary objects and focus on the galaxy at the center of the cutout. This is an extremely important feature when applying CNNs to unanalyzed surveys, as there is no robust technique to predict the correct cutout size for galaxies in the absence of $R_e$ measurements. The STN learns to perform an affine transformation which makes the downstream task of morphological parameter estimation easier and it consists of i) a localization network, ii) a parameterized grid generator, and iii) a sampler. The localization network (consisting of two convolutional layers, followed by a fully connected regression layer) takes the input image and outputs $\theta$, the six parameters of an affine transformation $\mathcal{T}_\theta$; this gives a transformation conditional on the input image. The grid generator thereafter uses the predicted transformation parameters to create a sampling grid ($G$). Finally, the sampler takes the set

of sampling points $\mathcal{T}_\theta(G)$, along with the input image, and produces the transformed input image, which is then passed on to the CNN. Since the STN can be trained with standard back-propagation, the entire GAMPEN framework can be trained end-to-end without any separate supervision required for the STN. For a more extensive discussion about STNs, please refer to Jaderberg et al. (2015).

In order to predict Bayesian posteriors, we treat the trained model, $w$, as a random variable, because, intuitively, there are many possible models that could be trained from the same training data, $D$. To predict the posterior, we need to marginalize over these possible models; and for that marginalization, we need to know how likely we were to train a particular model $w$ given the data, $p(w|D)$. In order to estimate $p(w|D)$, we use the Monte-Carlo Dropout (MCD) technique as introduced by Gal & Ghahramani (2016). In practice, this amounts to training the network with dropout before every weight layer and optimizing a cost function given by the log-likelihood with an L2 regularization term. At test time, each realization of the network's outputs — given by a forward pass with a random dropout — is a sample from the approximate parameter posterior. We do 1000 forwarded passes for every input galaxy to estimate the parameter posteriors.

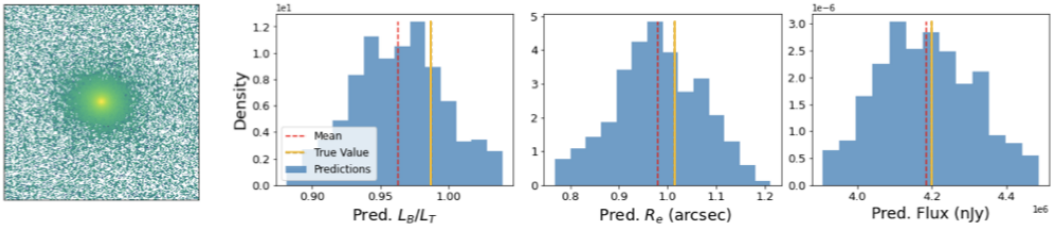


Figure 2: *(Left):* A randomly chosen simulated galaxy *(Right):* The posterior distributions for the galaxy as predicted by GAMPEN. The solid yellow line refers to the true value of each parameter.

## 3 Training & Preliminary Results

We first train GAMPEN on realistic simulations of galaxies and then fine-tune the already trained network using a small amount of real data. We use GalSim (Rowe et al., 2015) to generate 150,000 realistic galaxies matching the properties of $z < 0.25$ galaxies in the HSC-W survey. In order to have a diverse training set, $75\%$ of the simulated galaxies consisted of a bulge + disk component; and the remaining $25\%$ had either a bulge or a disk component. The parameters required to generate the galaxies are drawn from uniform distributions with ranges representative of typical galaxies at this redshift (Binney & Merrifield, 1998). We convolved these simulated galaxies with a representative point spread function (PSF) and added representative noise based on real HSC-W images. Using an 80-10-10 train-validation-test split, we trained GAMPEN on these simulated galaxies. The network was trained using stochastic gradient descent and its hyperparameters were tuned using the loss of the validation set. A randomly chosen simulated galaxy and its parameters as predicted by GAMPEN after training is shown in Fig. 2

Thereafter, we fine-tuned the network trained on simulations using a small amount of real data. We select $z < 0.25$ HSC-W g-band galaxies with secure redshifts and no imaging issues (such as cosmic ray hits), and then cross-match these with the Simard et al. (2011) catalog, which had performed bulge + disk decomposition using Sloan Digital Sky Survey imaging. We use the Simard et al. (2011) fits to represent the correct parameters for this set of $20,000$ real galaxies, of which $30\%$ are used to fine-tune the network trained on simulations, and the rest for validation and testing.

We regard the dropout rate used in GAMPEN as a variational parameter of the model because while using MCD, greater dropout rates lead to higher estimated uncertainties (on average). Thus, we train GAMPEN with different dropout rates and then calculate their coverage probabilities, defined as the fraction of the validation set galaxies where the true value lies within a particular confidence interval of the predicted distribution. From our experiments, a dropout rate of 0.1 yields a coverage probability roughly equal to the confidence level of the interval for which it was calculated. Thus, we choose this model as it should generate accurate uncertainties.

Finally, we calculate the residuals for the galaxies in our testing set. We define the residual as the difference between the most probable value of the predicted posterior distribution and the true value. The histogram of residuals for all three output parameters is shown in Fig. 3. All the histograms

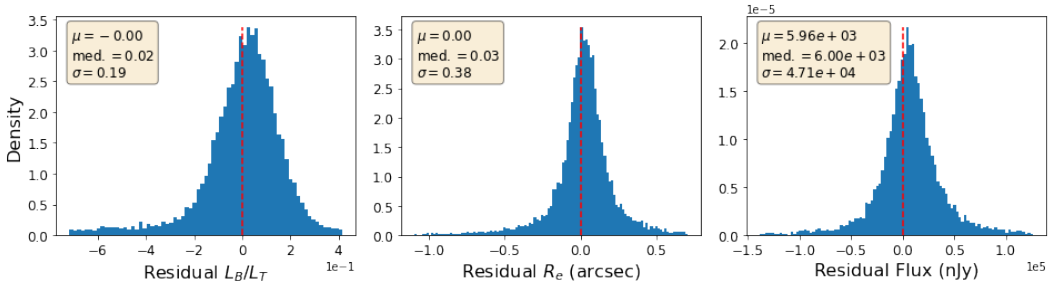


Figure 3: Histograms of residuals for the three output variables for HSC-W g-band $z < 0.25$ galaxies. The red line corresponds to $x = 0$ and the numbers in the top-left box refer to the mean ($\mu$), the median (med.), and the standard deviation ($\sigma$) of the residual distribution.

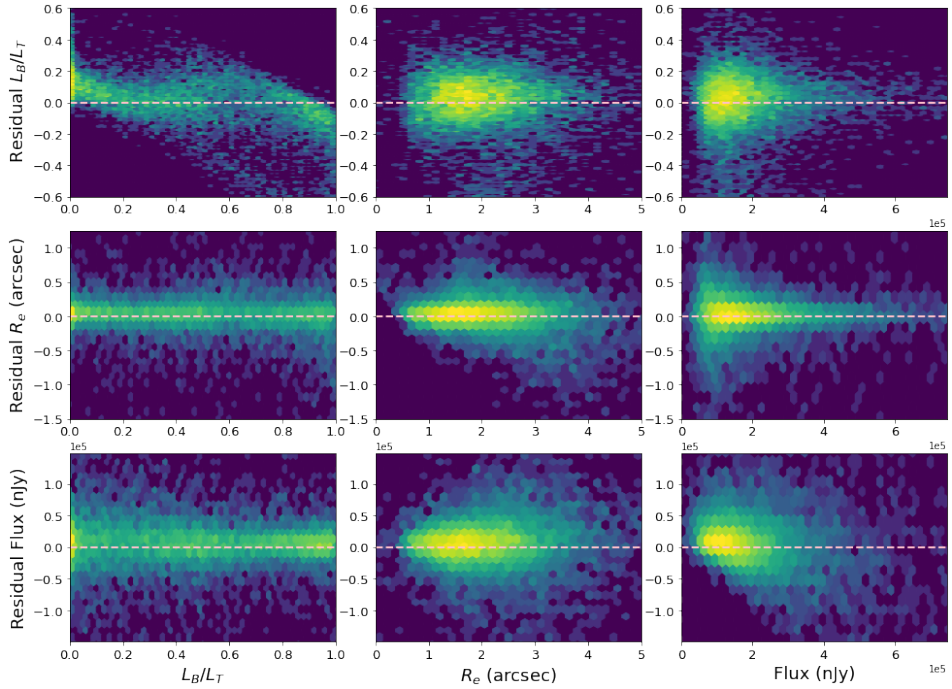


Figure 4: Two dimensional density plots for the prediction residuals against the true values of the output variables for the HSC-W g-band $z < 0.25$ galaxies

are centered at $\sim 0$ and have reasonable sigmas ($\sigma_{L_B/L_T} = 0.19$; $\sigma_{R_e} = 0.38$"; $\sigma_{Flux} = 4.7 \times 10^4$ nJy). Fig. 4 shows the residuals for each variable when plotted against the true values of the output variables. With some exceptions, the residuals are largely uniformly distributed about the $y = 0$ line.

In the top-left plot, we can see that for very small values of $L_B/L_T$ (i.e., a completely disk-dominated galaxy), GAMPEN over-predicts the amount of light in the bulge; and for very high values of $L_B/L_T$ (i.e., a completely bulge dominated galaxy), GAMPEN under-predicts the light in the bulge. Therefore, in these "edge" cases, when one component completely dominates over the other, GAMPEN under-predicts the amount of light in the weaker component. In the top-center plot, we can see that there is a slight decrease in $L_B/L_T$ residuals for galaxies with $R_e > 2.5$" — that is, the network measures larger galaxies more accurately. Lastly, from the right column, it is clear that all three residuals increase for fainter galaxies, which is not surprising as fainter galaxies are inherently more difficult to analyze due to their low signal-to-noise ratios.

## 4 Conclusions & Broader Impact

In this work, we have demonstrated that we can use GAMPEN, our machine learning framework, to predict accurate posterior distributions for morphological parameters of galaxies while using minimal

real training data. We demonstrate GAMPEN's capabilities by applying it on $z < 0.25$ HSC-Wide galaxies and showing that it can accurately recover the $L_B/L_T$, $R_e$, and total flux for these galaxies. The galaxies with the largest residuals are smaller, fainter, and/or have one morphological component completely dominating over the other — situations where morphological analysis is inherently difficult. We have also outlined how the use of an STN allows GAMPEN to crop out secondary galaxies present in the cutout and focus on galaxies at the center. This is the first time an STN has been used in astronomical analysis, and GAMPEN is the first ML framework that can provide Bayesian posteriors for morphological analysis of galaxies. After further testing, we aim to make GAMPEN and its associated models public by mid-2022

We believe that our use of publicly available datasets and catalogs encourages the development of community tools and facilitates more accessibility. Our general technique can also be easily applied to problems in other fields. We recognize that deep learning models have sometimes propagated existing biases in our society, or have been used in disastrous wars and mass surveillance. However, because the information available in astrophysical images is very different from that of human-centric "daily-life" imaging data, we believe that our work is much less prone to misuse or abuse.

## Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See §3
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See §4
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The relevant details for our algorithm is given in §2 and Appendix A
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] The relevant details for training and hyper-parameter tuning is given in §3 and Appendix A
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] This information is provided in Appendix A
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] This is mentioned in Appendix A
   (c) Did you include any new assets either in the supplemental material or as a URL? [No] GAMPEN will be publicly available in mid-2022 after further testing.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] All the data-sets used in this work are publicly available
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] All the data-sets used in this work are astrophysical in nature and doesn't pertain to personally identifiable information in any way
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

## References


Aihara, H., Armstrong, R., Bickerton, S., et al. 2018, Publications of the Astronomical Society of Japan, 70, doi: 10.1093/pasj/psx081

Bernardi, M., Meert, A., Sheth, R. K., et al. 2013, Monthly Notices of the Royal Astronomical Society, 436, 697, doi: 10.1093/mnras/stt1607

Binney, J., & Merrifield, M. 1998, Galactic astronomy (Princeton University Press), 796. https://press.princeton.edu/titles/6358.html

Cheng, T.-Y., Conselice, C. J., Aragón-Salamanca, A., et al. 2021, Monthly Notices of the Royal Astronomical Society, 507, 4425, doi: 10.1093/mnras/stab2142

Gal, Y., & Ghahramani, Z. 2016, in 33rd International Conference on Machine Learning, ICML 2016, Vol. 3 (PMLR), 1651–1660. https://proceedings.mlr.press/v48/gal16.html

Ghosh, A., Urry, C. M., Wang, Z., et al. 2020, The Astrophysical Journal, 895, 112, doi: 10.3847/1538-4357/ab8a47

Hausen, R., & Robertson, B. E. 2020, The Astrophysical Journal Supplement Series, 248, 20, doi: 10.3847/1538-4365/ab8868

Hubble, E. P. 1926, The Astrophysical Journal, 64, 321, doi: 10.1086/143018

Jaderberg, M., Simonyan, K., Zisserman, A., & Kavukcuoglu, K. 2015, Advances in Neural Information Processing Systems, 28, 2017 . http://proceedings.neurips.cc/paper/2015/file/33ceb07bf4eeb3da587e268d663aba1a-Paper.pdf

Powell, M. C., Urry, C. M., Cardamone, C. N., et al. 2017, The Astrophysical Journal, 835, 22, doi: 10.3847/1538-4357/835/1/22

Pozzetti, L., Bolzonella, M., Zucca, E., et al. 2010, Astronomy & Astrophysics, 523, A13, doi: 10.1051/0004-6361/200913020

Rowe, B., Jarvis, M., Mandelbaum, R., et al. 2015, Astronomy and Computing, 10, 121, doi: 10.1016/j.ascom.2015.02.002

Schawinski, K., Urry, C. M., Simmons, B. D., et al. 2014, Monthly Notices of the Royal Astronomical Society, 440, 889, doi: 10.1093/mnras/stu327

Simard, L., Trevor Mendel, J., Patton, D. R., Ellison, S. L., & McConnachie, A. W. 2011, Astrophysical Journal, Supplement Series, 196, 11, doi: 10.1088/0067-0049/196/1/11

Tremaine, S., Gebhardt, K., Bender, R., et al. 2002, The Astrophysical Journal, 574, 740, doi: 10.1086/341002

Tuccillo, D., Huertas-Company, M., Decencière, E., et al. 2018, Monthly Notices of the Royal Astronomical Society, 475, 894, doi: 10.1093/mnras/stx3186

van der Wel, A., Franx, M., van Dokkum, P. G., et al. 2014, The Astrophysical Journal, 788, 28, doi: 10.1088/0004-637X/788/1/28

Vega-Ferrero, J., Domínguez Sánchez, H., Bernardi, M., et al. 2021, Monthly Notices of the Royal Astronomical Society, 506, 1927, doi: 10.1093/mnras/stab594

Wuyts, S., Förster Schreiber, N. M., van der Wel, A., et al. 2011, The Astrophysical Journal, 742, 96, doi: 10.1088/0004-637X/742/2/96

# A Appendix

**Further Training Information** In order to increase reproducibility, we outline further information about the training process of GAMPEN.

We used a learning rate of $5 \times 10^{-6}$, and a batch size of 32. For the stochastic gradient descent optimizer, we used a momentum of 0.99 enabled Nesterov momentum.

All the cutouts used in this work were square cutouts of 239 pixels on each side. An inverse hyperbolic sine function (arcsinh) stretch was applied to both the simulated as well as the real galaxy images before being fed into the network. All the target labels were also normalized using standard scaling.

We used two NVIDIA Tesla P100 GPUs at the Yale Center for Research Computing for training GAMPEN. 40 epochs of training took $\sim 15$ hours.

**Data-Set Used** We used the publicly available Hyper Suprime-Cam Data Release 2 which is available at `https://hsc-release.mtk.nao.ac.jp/doc/index.php/sample-page/pdr2/`. The HSC collaboration didn't attach a specific public-use license along with their dataset. However, the data is available for use freely as long as the relevant data release publication is cited.