Galaxy Morphological Classification with Deformable Attention Transformer

Seokun Kang
Artificial Intelligence Graduate School, UNIST
oraclemiso@unist.ac.kr

Min-Su Shin
Korea Astronomy and Space Science Institute
msshin@kasi.re.kr

Taehwan Kim
Artificial Intelligence Graduate School, UNIST
taehwankim@unist.ac.kr

Abstract

Galaxy morphological classification is an important but challenging task in astronomy. Most prior work study coarse-level morphological classification and use raster low-dynamic range images, but we are interested in high-dynamic range images commonly produced in imaging surveys. To tackle this problem, first we build a dataset with high dynamic range for fine-level multi-class classification that are even challenging to human eyes. Then we propose to use Deformable Attention Transformer for this difficult task with five-bands images and masks, and in the experimental results our model achieves about 70% and 94% for top-1 and top-2 test set accuracies, respectively. We also visualize attention maps and analysis the results with respect to different classes and mask sizes to understand the data and behavior of the model. We confirm that our model has similar confusion patterns in confusion matrix as human along with attention visualization for capturing morphological characteristics.

1 Introduction

Morphological classification of galaxies is a common task required in observational research on galaxy evolution to study correlations between morphology and other physical quantities. Previous applications of machine learning have focused on the mainly binary classification of morphology as early and late-type galaxies, including usages of vision transformer and convolution neural networks with raster images [1, 2]

We extend the validity of the attention-based deep learning models in fine-level morphological classification in terms of the Hubble sequence with full exploitation of high-dynamic range galaxy images and detection masks based on signal-to-noise levels. Compared to previous works, we directly use the high-dynamic range images instead of raster low-dynamic range images in order to make the models fully recognize important morphological features. The new models also use pixel masks of invalid values commonly produced in imaging surveys. Classification in the Hubble sequence depends on morphological features such as spiral and spheroidal structures [3]. For example, Sab and Sc-types in the Hubble sequence show strong spiral structures.

The trained model should show similar systematic patterns in the classification as professional astronomers do. For example, the model should be uncertain for classifying apparently small galaxies due to the limitation of spatial resolution and sensitivity. We also expect confusion among the classes to be identical to what humans exhibit [4].

Machine Learning and the Physical Sciences workshop, NeurIPS 2022.

Table 1: Description of our unbalanced dataset; the number of data for each class and ratio from the total amount of the dataset.

	E	Im	S0	Sab	Sc	Sdm	dE	U	Total
Train	2457 16.129%	**230** **1.510%**	2081 13.661%	**6861** **45.040%**	2580 16.937%	1024 6.722%	0 0%	0 0%	15233
Test	616 16.143%	**58** **1.520%**	522 13.679%	**1717** **44.067%**	646 16.572%	257 6.581%	0 0%	0 0%	3816
Total	3073 15.786%	**288** **1.480%**	2603 13.372%	**8578** **44.067%**	3226 16.572%	1281 6.581%	68 0.349%	349 1.793%	19466

2 Method

2.1 Data

Images of galaxies are acquired in Pan-STARRS1 stack images as cutouts for given sky positions in FITS format for five different photometric bands [1] [5]. Labeled training data consist of galaxies with known Hubble sequence morphological types in three catalogs: AMIGA [6], EFIGI [7], and Nair & Abraham's catalog [8]. Regrouping classes with a small number of training samples together, our models consider six classes: E, S0, Sab, Sc, Sdm, and Im. We do not consider a class dE because there are not many training samples compared to other classes. The statistics of our dataset are in Table 1.

The input data include galaxy images and corresponding object masks as well as masks of invalid pixels in the format of 480 by 480 pixels. Therefore, the model handles the galaxy images and masks of invalid pixels (Nan-Masks) for the five photometric bands in addition to a single object mask per galaxy (Galaxy-Mask). If there is nan-value in a single band image, the nan-value need to be changed to an appropriate value in learning. We expanded the galaxy shape mask by dilation and then fill it using a 2D interpolation function if there is nan-value in the expanded galaxy-mask. We adopt the min-max normalization method so that the image pixel value is between 0 and 1. Data augmentation includes horizontal and vertical flips and rotations.

2.2 Model

Vision Transformer-based models have recently achieved superior results on many vision tasks [9, 10, 11, 12, 13, 14]. Among them, Deformable Attention Transformer(DAT) utilizes more flexible Deformable Attention [15] and has achieved state-of-the-art on image classification tasks. By utilizing the hierarchical pyramid structure validated in Pyramid Vision Transformer(PVT)[11] and Swin-Transformer[10], DAT has tried to reduce the computation cost used for Self-Attention, the biggest issue of existing Vision Transformer(ViT)[9]. Swin-Transformer uses window-based local attention. In this case, the computation cost used in the self-attention is less than a vanilla transformer, but the learning of the long-range dependency may be weaker. Therefore, DAT uses the deformable module to reduce the amount of computation but also performs self-attention on more important tokens. The overall architecture of the DAT is the almost same as that of the Swin-Transformer, except that the shift attention block in stages 3 to 4 has been replaced with the deformable attention block, as shown in Figure 1. By using this deformable attention, the model can learn powerful representations by taking more important tokens and global receptive fields.

To perform the Deformable Attention, DAT uses deformable attention equation 2 instead of vanilla attention equation 1 where feature map $x \in^{N \times C}$ and $W_q, W_k, W_v \in^{C \times C}$. By following equation 2, the keys and queries for attention are deformed where deformed input feature map $\tilde{x} \in^{H \times W \times C}$. So, final deformed attention is calculated like equation 3. By using this deformable attention, the model can learn powerful representations by taking more important tokens and global receptive fields.

$$z = \sigma(qk^\top/\sqrt{d})v, \text{ where } q = xW_q, k = xW_k, v = xW_v \tag{1}$$

$$q = \tilde{x}W_q, \tilde{k} = \tilde{x}W_k, \tilde{v} = \tilde{x}W_v, \text{ where } \tilde{x} = \phi(x; p + \Delta p) \text{ with } \Delta p = \theta_{offset}(q) \tag{2}$$

$$z = \sigma(q\tilde{k}^\top/\sqrt{d})\tilde{v} \tag{3}$$

[1] http://ps1images.stsci.edu/cgi-bin/ps1cutouts

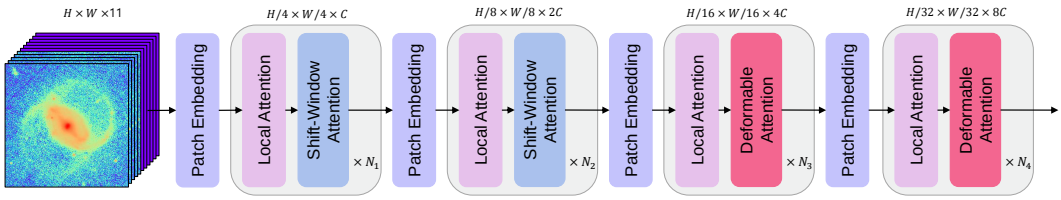


Figure 1: Overall architecture of DAT

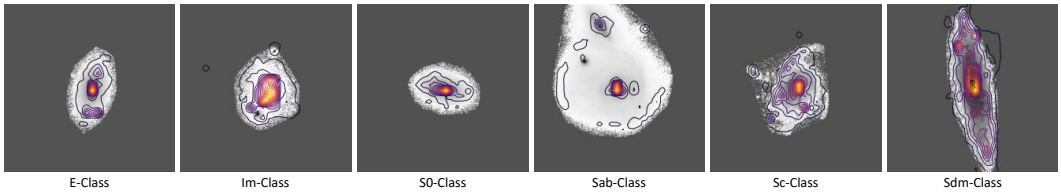


Figure 2: Distribution of the last layer's attention in the example galaxy for each class. The contours show the strength of attention, and the background gray-scale plots correspond to r-band images.

We compare the performance of DAT on galaxy morphological classification with that of prior work with another Vision Transformer-based model called Linformer [1] on Galaxy Zoo 2 dataset [16] and the experimental result shows that DAT achieves 82.222% test set accuracy and outperforms Linformer (80.427% test set accuracy in our experiment, 80.55% is reported in [1]). Therefore, we use DAT in our main experiments, expecting to perform self-attention on tokens that are more important than Linformer.

3 Results

This section will cover the various classification results on our dataset and analyze them. The hyper-parameters used to learn the base model are as follows. The warming up epoch is 10, the base learning rate is 4e-4, the minimum learning rate is 5e-6, the weight decay value is 0.05, clip gradient value is 1.0. It uses a cosine scheduler, and the batch size is 18. We observe not significant improvement after training with more epochs, so we set the epoch to 100 for all methods. Training takes about one day with A100 40GB GPU.

3.1 Classification Performance

We modify the DAT model to fit our data, and the hyper-parameters are the same in all model architectures. Since we use extremely unbalanced data, we try a number of different methods. First, weighted loss for each class was exploited according to the distribution of classes. Next, class uniform sampling was used so that all classes could be uniformly sampled in the mini-batches. In addition, we try to overcome the data issue by using data augmentation strategies called MixUp[17] and balanced-MixUp[18]. For the main results of our experiments, see Table 2. Our main model achieved 70.152% for top-1 accuracy on test set, 94.366% for top-2, and 99.319% for top-3, respectively.

We also visualize the attention maps in Figure 2. The models give high attention to the part where the galactic nucleus is located and the shape of the galaxy is traced as presented in Figure 2. See the appendix for how the attention map changes depending on the model's layer.

In the case of using weighted loss for each class, the performance was worse than base, and it can be found in [1] too. This can be attributed to the decrease in performance for classes with large numbers of data due to relatively increasing the weight of learning for the classes with smaller number of samples. And in the case of using Class Uniform Sampling, since the smallest number of data in classes (Im) is 288 and the largest number of data in classes (Sab) is 8578, we assume the data itself is not enough to learn for the classes with extremely small number of data.

By doing linear interpolation for two certain sampled data, MixUp encourages them to have linear behavior according to the decision boundary[17]. Since our dataset has continuous characteristics

Table 2: Classification performance of our models.

Model	Acc Top-1	Acc Top-2	Acc Top-3
Small	68.947%	93.842%	99.240%
Base	70.152%	94.366%	99.319%
Base-Weighted Class Loss	64.492%	91.824%	98.873%
Base-Class Uniform Sampling	64.596%	91.876%	98.716%
Base-MixUp	68.265%	93.449%	99.083%
Base-Balanced MixUp	69.418%	93.973%	99.319%

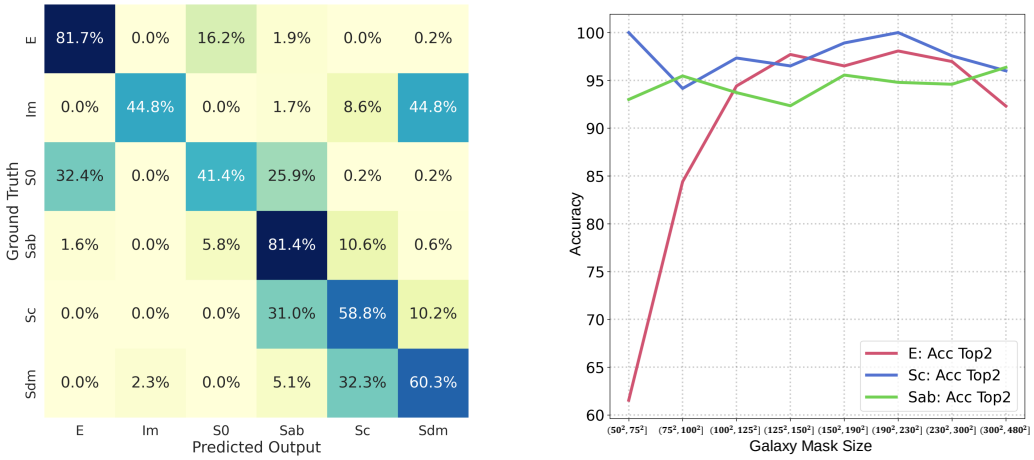


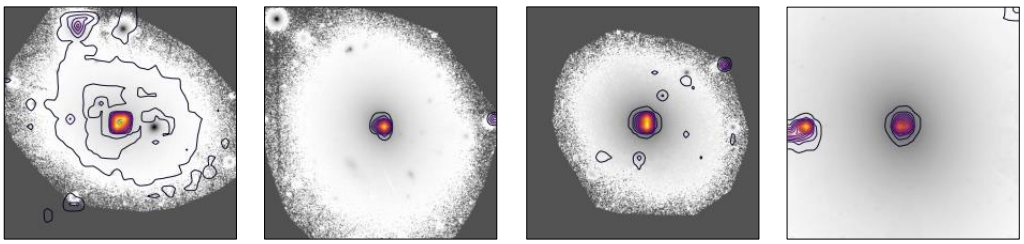
Figure 3: Confusion matrix of prediction (left) and the top-2 accuracy of E, Sab, and Sc classes with respect to the bins of galaxy mask sizes (right) for the test data in the base model.

rather than traditional image datasets, the decision boundaries of the dataset may be closer and unclear. Because of this, we hypothesize that the MixUp approach did not help train on our dataset and rather hindered learning because MixUp did not consider highly unbalanced datasets. The balanced-MixUp performes a mixup between instance-based mini-batch and class-based mini-batch on a highly unbalanced dataset[18]. By doing this, it overcomes the limitations of the MixUp method that does not take into account the distribution of the highly unbalanced dataset, which is the cause of the performance degradation. We observed the performance was improved compared to MixUp by using the balanced MixUp. But the result shows the balanced MixUp method is still not helpful for training on our dataset and we conjecture the dataset itself needs more detailed investigation along with better approaches.

3.2 Confusion Pattern

In the cases of S0 and E, Sab and Sc, and Im and Sdm, morphological features are similar between them, making it a difficult classification problem even for professionally trained astronomers. As shown in Figure 3, the tendency found in the DAT model is similar to what humans generally show. The classification confusion found in our results is not surprising because the Hubble sequence represents galaxy morphology in continuous changes of morphological features (see the appendix for the results in the other models). Therefore, the top-2 accuracy presented in Table 2 is much higher than the top-1 accuracy.

The effect of limited field-of-view due to the fixed size of images appears in the classification performance for different sizes of galaxies. For small galaxies, the spatial resolution is too poor to depict important morphological features for certain morphological types. As it is difficult to see what phrases are written in distant traffic signs, it is not easy to distinguish key morphological features of galaxies when a galaxy is small. Figure 3 shows that the classification accuracy of E-class galaxies is quite low for small size, which is measured by the number of pixels in the object masks, due to the lack of strong detectable morphological features. As size increases, the classification performance



(a) PGC0022950 (b) PGC0029822 (c) PGC0051312 (d) PGC0041220

Figure 4: Large E-class galaxies where the size of object masks is between 300^2 and 480^2. The model correctly classifies (a), (b), and (c) galaxies. However, (d) galaxy is incorrectly predicted as Sab-class due to the influence of a foreground star (i.e., the bright object near the edge) which is assigned high attention.

improves for E-class galaxies. As shown in Figure4 as an example, most E-class galaxies are correctly classified even though they are large and the spatial coverage of the data is not large enough to present an entire galaxy. However, some galaxy shown in Figure 4 is affected by foreground objects, which are not part of the galaxy. As highlighted in Figure3, the top-2 accuracy of the E-class is slightly decreased for large galaxies because large E-type galaxies are more susceptible to the influence of foreground irrelevant sources affecting simple morphological features of E-type galaxies. In the case of the Sab class, the accuracy does not change significantly regardless of the galaxy size because distinguishable characteristics are recognized even when the object is small.

4 Conclusion

We conducted extensive experiments with several machine learning approaches on our new dataset. We show that our DAT models 1) utilize five different bands images(g, r, i, z, y-band), 2) does multiclass classification on the unclean high-dynamic range galaxy images, 3) not only use galaxy images but also use galaxy masks and nan-value masks, and 4) tried to overcome the unbalanced dataset by using relevant machine learning approaches.

Furthermore, we show that by analyzing predicted outputs by the model, the tendency of the person to classify and the tendency of the model to classify may have similar systematic patterns.

There are many unlabeled datasets in the field of astronomy. As a future work, we aim to overcome the unbalanced dataset issue by using semi-supervised learning that may utilize the huge unlabeled datasets.

5 Potential broader impact of this work

The inputs including masks of invalid pixels in this work are similar to what big astronomical survey projects typically produce. We expect our model to be easily adopted in astronomy community of large imaging surveys with their data reduction pipelines and products.

The systematic patterns depending on objects' size in classification performance shown here highlights the expected performance of vision tasks with limited spatial resolution in data. For example, classification and segmentation models for autonomous vehicles and medical applications must be affected by the limitation of spatial resolution in producing uncertain or wrong results in handling small distant objects found in images for vehicles and poor-resolution radiology images.

We do not find the significant potential negative societal impact of our work as we deal with astronomical data, not human or privacy-sensitive ones.

6 Acknowledgement

This work was partly supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2021-0-02068, Artificial Intelligence Innovation Hub). Min-Su Shin was partly supported by the Korea Astronomy

and Space Science Institute (KASI) grant funded by the Korea government (MSIT) (No. 2022186804, big data analysis and machine learning applications in astronomy). The authors acknowledge that the computational work reported on in this paper was partly performed on the KASI Science Cloud platform supported by KASI.

References

[1] Joshua Yao-Yu Lin, Song-Mao Liao, Hung-Jin Huang, Wei-Ting Kuo, and Olivia Hsuan-Min Ou. Galaxy morphological classification with efficient vision transformer. *arXiv preprint arXiv:2110.01024*, 2021.

[2] Ting-Yun Cheng, Christopher J Conselice, Alfonso Aragón-Salamanca, Michel Aguena, Sahar Allam, F Andrade-Oliveira, J Annis, AFL Bluck, D Brooks, David L Burke, et al. Galaxy morphological classification catalogue of the dark energy survey year 3 data with convolutional neural networks. *Monthly Notices of the Royal Astronomical Society*, 507(3):4425–4444, 2021.

[3] Ronald J. Buta. *Galaxy Morphology*, pages 1–89. Springer Netherlands, Dordrecht, 2013.

[4] Robert Geirhos, Kristof Meding, and Felix A. Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13890–13902. Curran Associates, Inc., 2020.

[5] H. A. Flewelling, E. A. Magnier, K. C. Chambers, J. N. Heasley, C. Holmberg, M. E. Huber, W. Sweeney, C. Z. Waters, A. Calamida, S. Casertano, X. Chen, D. Farrow, G. Hasinger, R. Henderson, K. S. Long, N. Metcalfe, G. Narayan, M. A. Nieto-Santisteban, P. Norberg, A. Rest, R. P. Saglia, A. Szalay, A. R. Thakar, J. L. Tonry, J. Valenti, S. Werner, R. White, L. Denneau, P. W. Draper, K. W. Hodapp, R. Jedicke, N. Kaiser, R. P. Kudritzki, P. A. Price, R. J. Wainscoat, S. Chastel, B. McLean, M. Postman, and B. Shiao. The Pan-STARRS1 Database and Data Products. , 251(1):7, November 2020.

[6] M. Fernández Lorenzo, J. Sulentic, L. Verdes-Montenegro, J. E. Ruiz, J. Sabater, and S. Sánchez. The AMIGA sample of isolated galaxies. X. A first look at isolated galaxy colors. , 540:A47, April 2012.

[7] A. Baillard, E. Bertin, V. de Lapparent, P. Fouqué, S. Arnouts, Y. Mellier, R. Pelló, J. F. Leborgne, P. Prugniel, D. Makarov, L. Makarova, H. J. McCracken, A. Bijaoui, and L. Tasca. The EFIGI catalogue of 4458 nearby galaxies with detailed morphology. , 532:A74, August 2011.

[8] Preethi B. Nair and Roberto G. Abraham. A Catalog of Detailed Visual Morphological Classifications for 14,034 Galaxies in the Sloan Digital Sky Survey. , 186(2):427–456, February 2010.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[11] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 568–578, October 2021.

[12] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[13] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021.

[14] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two pure transformers can make one strong gan, and that can scale up. *Advances in Neural Information Processing Systems*, 34:14745–14758, 2021.

[15] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4794–4803, 2022.

[16] Kyle W Willett, Chris J Lintott, Steven P Bamford, Karen L Masters, Brooke D Simmons, Kevin RV Casteels, Edward M Edmondson, Lucy F Fortson, Sugata Kaviraj, William C Keel, et al. Galaxy zoo 2: detailed morphological classifications for 304 122 galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 435(4):2835–2860, 2013.

[17] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[18] Adrian Galdran, Gustavo Carneiro, and Miguel A González Ballester. Balanced-mixup for highly imbalanced medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 323–333. Springer, 2021.

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]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (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)? [No]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (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]
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? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]
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, xwith 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]

A Appendix

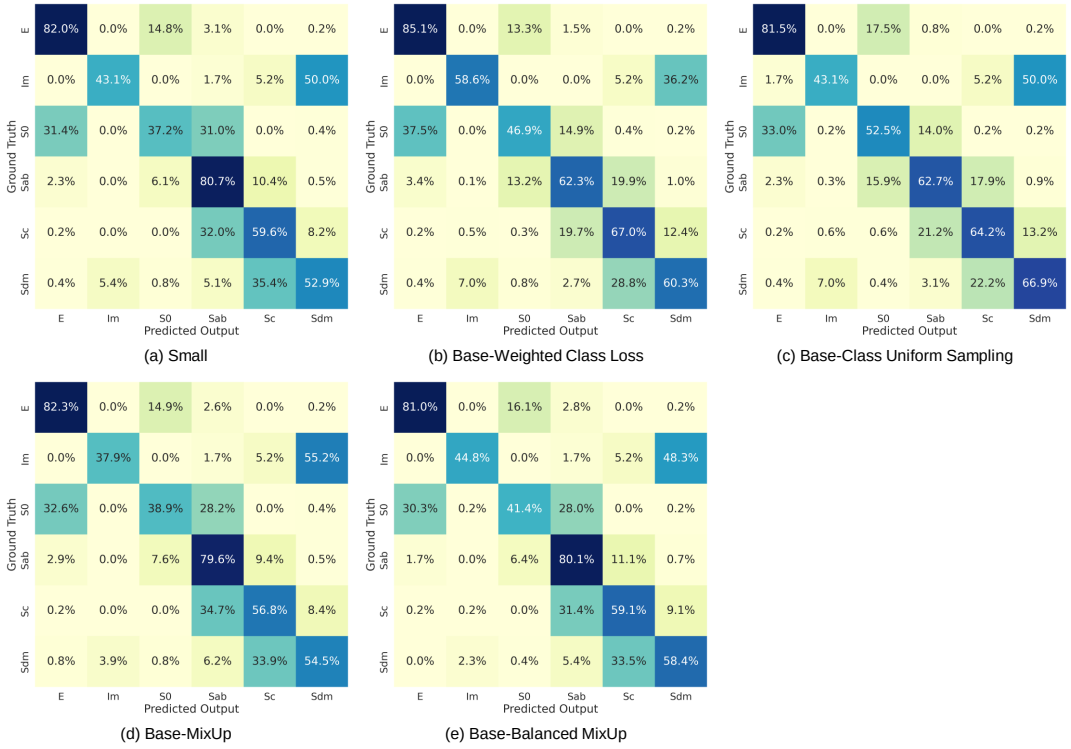


Figure 5: Confusion matrix for all models.

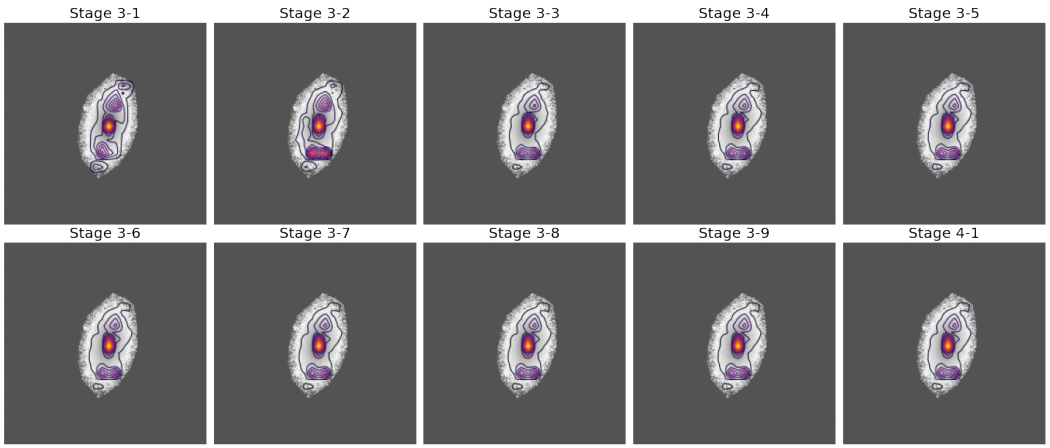


Figure 6: Attention maps for PGC0003830 in E class with prediction confidences in the base model as E: 0.863, Im: 0.000, S0: 0.132, Sab: 0.005, Sc: 0.000, and Sdm: 0.000.

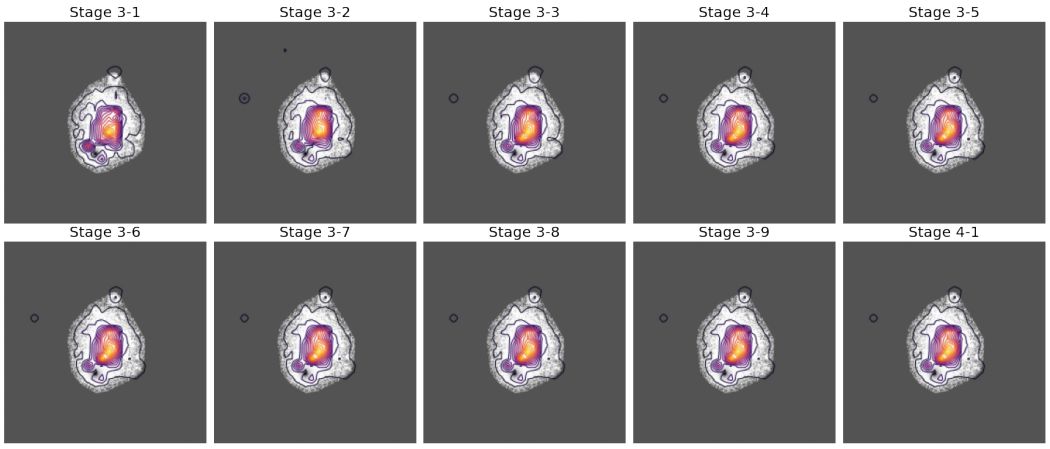


Figure 7: Attention maps for PGC0009530 in Im class with prediction confidences in the base model as E: 0.000, Im: 0.852, S0: 0.000, Sab: 0.000, Sc: 0.001, and Sdm: 0.147.

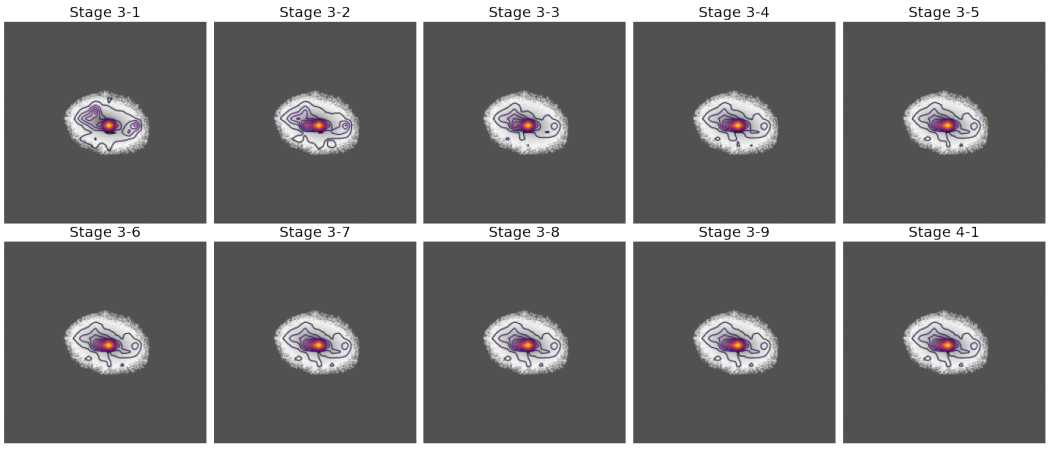


Figure 8: Attention maps for PGC0036944 in S0 class with prediction confidences in the base model as E: 0.168, Im: 0.000, S0: 0.630, Sab: 0.201, Sc: 0.000, and Sdm: 0.000.

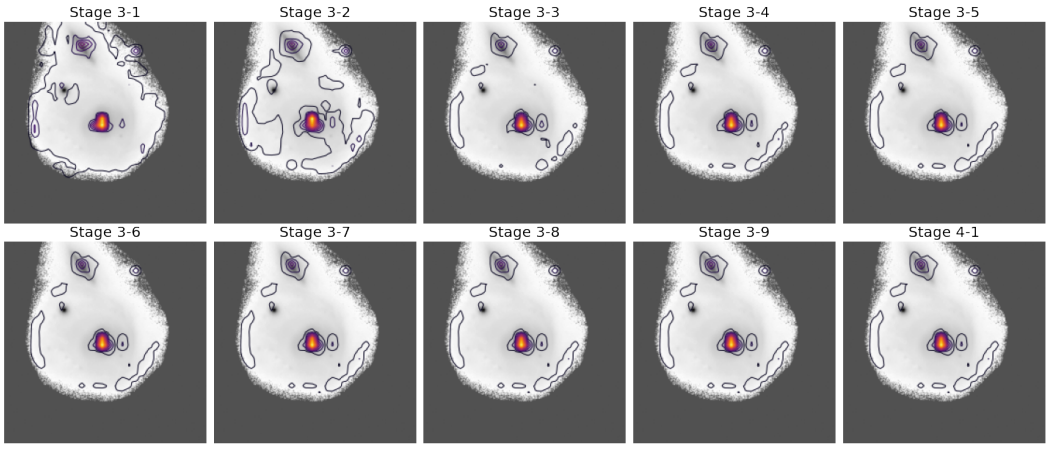


Figure 9: Attention maps for PGC0002331 in Sab class with prediction confidences in the base model as E: 0.004, Im: 0.000, S0: 0.092, Sab: 0.891, Sc: 0.013, and Sdm: 0.000.

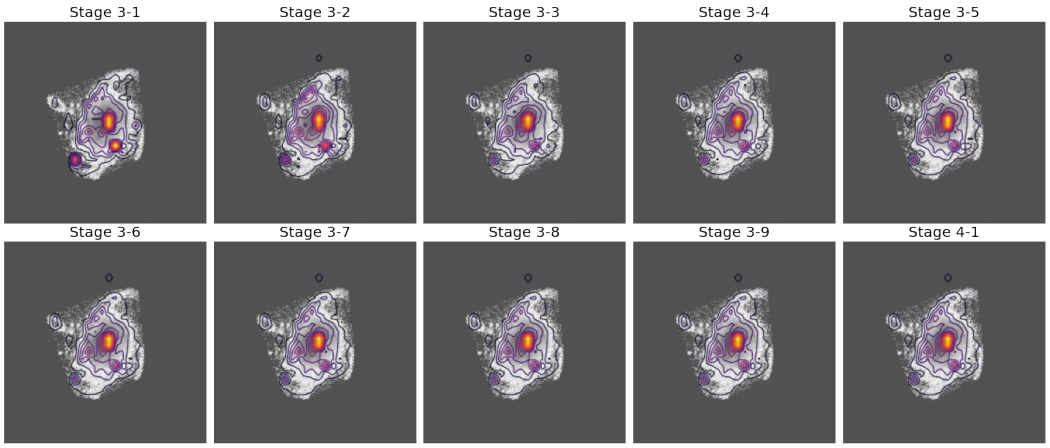


Figure 10: Attention maps for PGC0021443 in Sc class with prediction confidences in the base model as E: 0.000, Im: 0.000, S0: 0.000, Sab: 0.094, Sc: 0.888, and Sdm: 0.018.

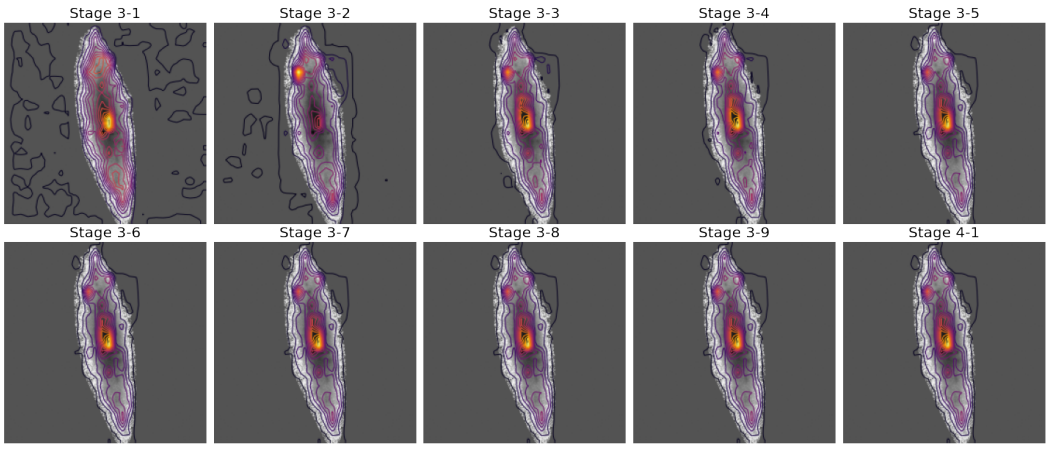


Figure 11: Attention maps for PGC0035900 in Sdm class with prediction confidences in the base model as E: 0.000, Im: 0.113, S0: 0.000, Sab: 0.000, Sc: 0.026, and Sdm: 0.861.