Galaxy Morphology Classification with Counterfactual Explanation

Zhuo Cao[1] **Lena Krieger**[1] **Hanno Scharr**[1] **Ira Assent**[1,2]
[1]IAS-8, Forschungszentrum Jülich, Germany [2]Aarhus University, Denmark
{z.cao, l.krieger, h.scharr, i.assent}@fz-juelich.de

Abstract

Galaxy morphologies play an essential role in the study of the evolution of galaxies. The determination of morphologies is laborious for a large amount of data giving rise to machine learning-based approaches. Unfortunately, most of these approaches offer no insight into how the model works and make the results difficult to understand and explain. We here propose to extend a classical encoder-decoder architecture with invertible flow, allowing us to not only obtain a good predictive performance but also provide additional information about the decision process with counterfactual explanations.

1 Introduction

Galaxies are the primary building blocks of the universe, composed of stars, stellar remnants, interstellar gas, dust, and dark matter. A key objective in galaxy research is to elucidate how galaxies have evolved from their early stages to the diverse and large forms observed today [1]. Specifically, analyzing the morphology and structure of galaxies is essential for understanding their evolution, as these aspects are intricately linked to their evolutionary history and are crucial for exploring the physical parameters of galaxies. Morphological features are essential for interpreting its evolutionary history and determining a galaxy's current dynamic state, such as the distribution and movement of stars, gas, and dark matter.

Significant efforts have been dedicated to designing galaxy morphology classification schemes and data collection methods. For example, Galaxy Zoo [2] and its successor Galaxy Zoo 2 [3], classify galaxies from the Sloan Digital Sky Survey (SDSS) [4] into basic types. Recently, the classification of galaxy morphologies can be predicted with CNN-based models [5–8]. These automated approaches surpass previous methods and have been applied across multiple surveys [9–11]. The drawback of these methods is their black-box characteristics, limiting the application of these methods because of the lack of interpretability and explainability. In this work, we target this issue with validating and insightful counterfactual explanations, demonstrating the importance of certain features for the decision-making process.

2 Data and Methodology

Visual counterfactual explanations Visual counterfactual explanations (CEs) [12, 13] seek to make only semantically meaningful modifications to an input image in order to obtain a similar image with a target label prediction outcome.

For a given image $\mathbf{x}$, the objective is to find a counterfactual proposal $\mathbf{x}^{cf}$ that has low counterfactual (CF) loss:

$$\mathcal{L}_{cf}(\mathbf{x}^{cf}) = f(\mathbf{x}^{cf}, y^{cf}) + s(\mathbf{x}, \mathbf{x}^{cf}) \tag{1}$$

Machine Learning and the Physical Sciences Workshop, NeurIPS 2024.

where a function $s(\mathbf{x}, \mathbf{x}^{cf})$ quantifies the perceptual distance between $\mathbf{x}$ and $\mathbf{x}^{cf}$ and function f yields a lower loss when the classifier predicts a label for the counterfactual that is closer to the target label y^{cf}. In other words, a counterfactual explanation reveals what should have been different in $\mathbf{x}$ to observe a diverse outcome with label y^{cf} instead of y. Thus, the approach offers deeper insights into the features that significantly contribute to the model's decision-making. CEs accentuate class-relevant features, illustrating how alterations to these features shift the prediction from one class to another.

While counterfactual explanations provide insights into the differences between predicted classes, generating them presents challenges. First, counterfactuals must align with the data distribution, meaning they need to look realistic within the context of the original dataset. For example, generating a counterfactual image (e.g., turning a cat into a dog) requires a robust generative model capable of maintaining natural, coherent results. Second, irrelevant features must remain unchanged during the generation process. Altering irrelevant or unrelated features can lead to explanations that are misleading or uninformative. For example, if a counterfactual explanation changes the background of an image when the focus should be on the object itself, the explanation may fail to provide useful insights about the model's decision-making process. Finally, extracting meaningful representations from high-dimensional data, such as images, poses a significant challenge. In such cases, identifying the most relevant features to adjust while preserving the overall structure is difficult due to the complexity of the feature space. In this work, we address these challenges.

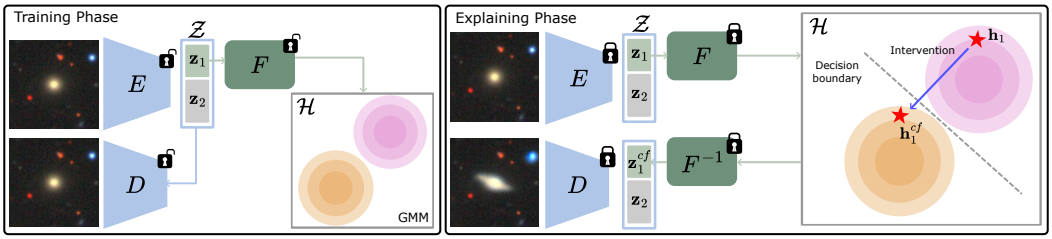

Figure 1: Architecture of our model. Left panel: training phase. Right panel: explanation phase.

Model architecture To obtain robust counterfactuals, we construct a model that exploits invertible flows such that counterfactuals obtained in latent space can be translated back to image space. The proposed model includes three components: an encoder E, decoder D, and invertible flow F [14]. Briefly, F enables bijective mapping through a specific architecture (see below). As shown in Figure 1 (left panel), the encoder maps the image to latent space $\mathcal{Z}$, and the decoder maps latent vectors back to the image space. The invertible flow converts the latent distribution to a Gaussian Mixture Model (GMM) [15], clustering vectors with the same label in a hidden space $\mathcal{H}$. The input image is classified by the closest cluster mean in $\mathcal{H}$. The entire model can be trained in an end-to-end way.

Compared to commonly used traditional classification methods, the encoder E acts as feature extraction, with an invertible flow replacing MLP. This design reduces the dimensionality of the image data and simplifies the decision boundary to a straight line between two Gaussian means. It also ensures a bijective mapping between latent space $\mathcal{Z}$ and hidden space $\mathcal{H}$, enabling counterfactual explanations. As shown in Figure 1 (right panel), the input image maps to a latent vector $\mathbf{z}$, then to a hidden vector $\mathbf{h}$ for classification. A counterfactual latent vector $\mathbf{z}^{cf}$ is created by pushing $\mathbf{h}$ across the decision boundary and mapping it back to latent space. The decoder D then converts $\mathbf{z}^{cf}$ to image space. The latent space is regularized with Maximum Mean Discrepancy (MMD) [16] (see also Equation 2) to keep the encoding function E Lipschitz continuous, ensuring interpolability and meaningful counterfactuals [17].

This intervention enables us to achieve an arbitrarily low f loss in Equation 1. However, minimizing the distance between the original and counterfactual images $s(\mathbf{x}, \mathbf{x}^{cf})$ is still necessary, which translates to minimizing the distance between their latent vectors due to the Lipschitz continuity of the encoding function. We achieve this by splitting the latent vector into class-dependent $\mathbf{z}_1$ and class-independent $\mathbf{z}_2$ components, where only $\mathbf{z}_1$ is used by the invertible flow F for classification. Additionally, we train the invertible flow with an information bottleneck objective [18] (see also Equation 2 and 4) to reduce mutual information between latent and hidden spaces, ensuring that F focuses on essential classification information with minimal alteration to $\mathbf{z}_1$ similar to pixel-level counterfactual generation by [19]. Note that the decoder may ignore $\mathbf{z_1}$ completely if $\mathbf{z_2}$ contains

duplicated information to $\mathbf{z_1}$. Applying MMD constraints in the latent space removes the redundancy in the latent vector [20], making the generated image respond to the modification in $\mathbf{z_1}$. The details of the model and training procedure can be found in Appendix A.1.

Invertible Flow An Invertible flow, or Invertible Neural Network (INN) [14, 21, 22], is a type of neural network architecture designed so that its forward and backward operations are both computationally feasible and reversible. This means that given the output, the original input can be accurately reconstructed. INNs achieve this by using specific structures that ensure bijective (one-to-one) mappings between the input and output spaces, called coupling layers. In a coupling layer, the input data is split into two parts. One part of the data remains unchanged, while the other part is transformed using a function conditioned on the first part. This approach ensures that the transformation is invertible and the Jacobian determinant of the transformation is easy to compute.

Loss Functions As previously mentioned, the objective comprises two key aspects: firstly, the model is required to generate in-distribution images, and secondly, decision-irrelevant features should remain unchanged. To accomplish this, we design the loss functions as follows:

$$\mathcal{L} = \mathcal{L}_{\mathrm{R}}(\tilde{\mathbf{x}}, \mathbf{x}) + \mathcal{L}_{\mathrm{MMD}}(\mathbf{z}, \mathbf{n}) + \mathcal{L}_{\mathrm{IB}}(\mathbf{z_1}, y) \tag{2}$$

Here, $\mathcal{L}_{\mathrm{R}}(\tilde{\mathbf{x}}, \mathbf{x}) = \Phi(\tilde{\mathbf{x}}) - \Phi(\mathbf{x})$ denotes the reconstruction loss between the generated image ($\tilde{\mathbf{x}}$) and the input image ($\mathbf{x}$), where Φ represents a VGG16 model pre-trained on the ImageNet dataset [23]. This model captures meaningful representations, as discussed in [24].

The second term, $\mathcal{L}_{\mathrm{MMD}}$, describes the Maximum Mean Discrepancy loss, which encourages the latent vector ($\mathbf{z}$) to approximate a Gaussian distribution. This ensures the interpolability of the latent space so that the generated image with the modified latent vector is still in-distribution. The loss is empirically estimated [16] by

$$\begin{aligned}\mathcal{L}_{\mathrm{MMD}}(\mathbf{z}, \mathbf{n}) = &\frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i} k(\mathbf{z}^i, \mathbf{z}^j) - \frac{2}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} k(\mathbf{z}^i, \mathbf{n}^j) \\ &+ \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i} k(\mathbf{n}^i, \mathbf{n}^j)\end{aligned} \tag{3}$$

where m is the batch size while training. The term $\mathbf{z}^i = E(\mathbf{x}^i)$ denotes the latent vector of i-th input data $\mathbf{x}^i$, $\mathbf{n}^i$ represents the i-th sample from the target Gaussian distribution, and k is the kernel, more precisely a Radial Basis Function (RBF) in this work. Together with the reconstruction loss $\mathcal{L}_{\mathrm{R}}$, these components constitute the MMD-VAE [17], a generative model known for its strong reconstruction quality.

The last term in the loss function $\mathcal{L}$ (Equation 2) represents the information bottleneck loss [18, 25]. On high-level, it is expressed as:

$$\mathcal{L}_{\mathrm{IB}} = I(\mathcal{Z}, \mathcal{H}) - \beta I(\mathcal{H}, \mathcal{Y}) \tag{4}$$

The first term in Equation 4 on the right-hand side minimizes the mutual information I between the latent space $\mathcal{Z}$ and the hidden space $\mathcal{H}$, while the second term maximizes the mutual information between the hidden space and the class label $\mathcal{Y}$. The combination of both components ensures that only essential information is used for classification. The parameter β controls the trade-off between preserving relevant information and discarding irrelevant details. Higher β values emphasize task performance, leading to better classification accuracy but potentially less robust uncertainty quantification. Lower β values prioritize compression, resulting in improved uncertainty calibration and out-of-distribution detection, at the cost of some classification accuracy. In this work, an intermediate value of 3 is used. For the detailed implementation of this loss function, readers are referred to the original paper [18]. Note that this loss connects the input data $\mathbf{x}$ to its corresponding class label y and is applied only to the class-dependent component $\mathbf{z}_1$.

Galaxy10 DECaLS We study Galaxy 10 DECaLS[1] with 17,736 DESI Legacy Imaging Surveys (DECaLS) [26] images (g, r and z band) and labels from Galaxy Zoo Release 2 [3] describing galaxy morphology in ten distinct classes. Figure 2 shows examples of each class. We normalize the images from [0, 255] to [0, 1], rotate the image randomly, and resize it to 256 x 256 after cropping the central region.

[1]The data is publicly available online

3 Results

Metrics We evaluate the model with regard to accuracy and similarity. Our trained model reports an accuracy of $\sim 80\%$. The accuracy for each class varies between 70% to 90% except for the disturbed galaxy, which has an accuracy of about 41% reflecting its complexity. The similarity between the counterfactual and original images is high, with Mean Squared Distance of 0.006 and Structural Similarity Index Measure (SSIM) [27] of 0.96. SSIM is designed to better align with human visual perception compared to traditional metrics. Details in Appendix A.2.

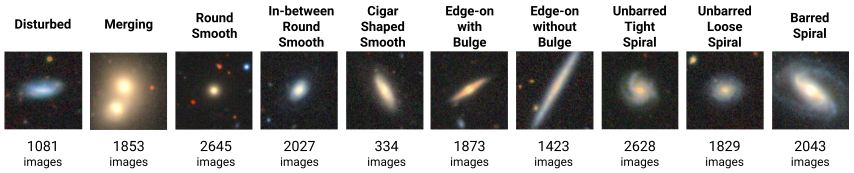

Figure 2: Sample images for each class of Galaxy 10 DECaLS and number of instances.

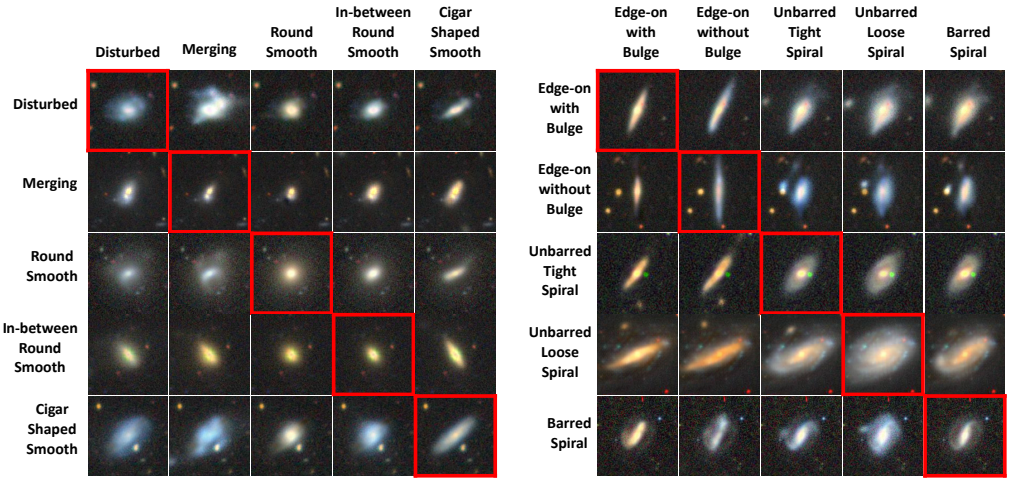

Figure 3: Original images aligned across diagonal (red boxes) with counterfactuals displayed in the same row classified according to column labels. Due to limited space, a complete image grid is shown in Appendix A.3.

Counterfactual Explanation Results are visualized in Figure 3 by overlaying the difference between counterfactual $\mathbf{x}^{cf}$ and reconstructed image $\tilde{\mathbf{x}}$ on the input image $\mathbf{x}$, i.e. $\mathbf{x} + (\mathbf{x}^{cf} - \tilde{\mathbf{x}})$. Notably, decision-relevant areas can be emphasized as in other Explainable AI methods, by calculating the difference between the counterfactual and the reconstruction, $\mathbf{x}^{cf} - \tilde{\mathbf{x}}$ (Example in Appendix A.3).

When comparing round and cigar-shaped smooth galaxies, we notice the change in shape from round to more elongated as expected. Similarly, edge-on galaxies with and without bulge are clearly distinguishable by the central galaxy bulge. Barred and unbarred spiral galaxies can be distinguished by their central structures, with barred spirals featuring elongated central objects and unbarred spirals having rounder central regions. Please note that there is no change in background during reconstruction. Thus it can be concluded that the encoder correctly distinguishes between class-dependent and class-independent latent features.

We observe that the spiral morphologies, i.e., barred tight/ loose and unbarred spiral, and their visual counterfactual explanations, displayed in the right panel, look very similar. As the image is compressed in feature space before classification, the fine structures are gradually removed. Therefore, the invertible flow and the decoder have no information about fine structures like spirals. The remaining information is likely shared in their latent features. As a result, the groups are close to each other in the latent space and the distance to the nearest sample of the neighboring class is very small (see Figure 4), corresponding to a minimal difference in the image.

Latent Space Analysis The latent space learned by our model is further analyzed with t-SNE plots [28], see Figure 4. The left and right t-SNE plots illustrate the latent space features $\mathbf{z}_1$ and $\mathbf{z}_2$. While $\mathbf{z}_1$ shows at least a few connected groups, $\mathbf{z}_2$ hardly shows any groupings corresponding to the classes. This is a desired behavior since the information in $\mathbf{z}_2$ should be independent of the classifications. Hidden space features $\mathbf{h}_1$, i.e., $\mathbf{z}_1$ transformed with F contain mostly clearly separable clusters, indicating a separation of the latent space features regarding their classes. The clusters that cannot be separated are Unbarred Loose Spiral (green) and Unbarred Tight Spiral (cyan). As already observed above, the model is challenged regarding very fine details in the images, e.g., spirals. The remaining similarities of the classes reflect on the positioning of the clusters in latent space.

The t-SNE plot does not reveal any information about samples belonging to the Disturbed (dark green) class. These galaxies tend to be diffuse and resemble different classes, rendering them hard to group. This is reflected in the accuracy of 41% for Disturbed galaxies. Even though there are few orange points in $\mathbf{z}_1$'s t-SNE plot due to the few included Cigar Shaped Smooth samples in the dataset, there is a well-visible cluster formed in the t-SNE plot according to $\mathbf{h}_1$. This demonstrates the model's ability to handle imbalanced classes.

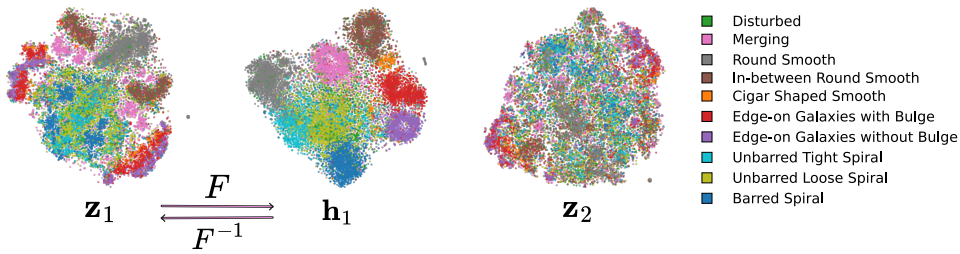

Figure 4: t-SNE plots for class-dependent ($\mathbf{z}_1$), hidden ($\mathbf{h}_1$) and background ($\mathbf{z}_2$) features.

4 Discussion

Previous works employed Explainable AI techniques to astrophysical use cases, such as identifying informative latent space representations of galaxy spectra with SHAP values [29] or detecting ultra-compact dwarfs and globular clusters using Localized Generalized Matrix Learning Vector Quantization (LGMLVQ) to provide feature importance for each class, class-wise representative samples and the possibility for non-linear visualization of the data [30]. Bhambra et al. [31] explain galaxy morphology classification with saliency maps. In contrast to our approach, they apply SmoothGrad to illustrate which pixels contribute to classification. Their findings show that the trained ensemble, consisting of the three architectures VGG16, ResNet50v2, and Xception, sometimes disagrees with the target labels in Galaxy Zoo assigned by citizen science. They show examples indicating that the ensemble might be more correct than the ground truth.

Our approach is sensitive to mislabeled samples, as the latent space feature vectors are changed and thus the bias is shifted towards the counterfactuals. This effect can be investigated by analyzing the distributions determined by invertible flow F. Closer examination of $\mathbf{h}_1$, see Figure 4, can identify problematic classes for further reviewing. This is a desirable property as it identifies possible issues of the pipeline, i.e., data or model, that might not be discovered otherwise.

5 Conclusion

We present a new approach to produce realistic counterfactual explanations for galaxy morphologies by adjusting the class-dependent latent space features. In future work, we aim to address the current limitation related to capturing fine details within the images. Additionally, we are interested in exploring the interpretability that the distributions inside F hold, as this could offer valuable insights into the relationships between the classes and reveal any limitations of the classifier or the data when the classes cannot be clearly distinguished from each other.

Acknowledgments and Disclosure of Funding

The authors gratefully acknowledge computing time on the supercomputer JURECA [32] at Forschungszentrum Jülich under grant delia-mp.

References

[1] Z. Z. Wen, X. Z. Zheng, and F. X. An, "Probing asymmetric structures in the outskirts of galaxies," *The Astrophysical Journal*, vol. 787, p. 130, may 2014.

[2] C. J. Lintott, K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M. J. Raddick, R. C. Nichol, A. Szalay, D. Andreescu, P. Murray, and J. Vandenberg, "Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey," *Monthly Notices of the Royal Astronomical Society*, vol. 389, pp. 1179–1189, Sept. 2008.

[3] K. W. Willett, C. J. Lintott, S. P. Bamford, K. L. Masters, B. D. Simmons, K. R. Casteels, E. M. Edmondson, L. F. Fortson, S. Kaviraj, W. 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*, vol. 435, no. 4, pp. 2835–2860, 2013.

[4] D. G. York, J. Adelman, J. John E. Anderson, S. F. Anderson, J. Annis, N. A. Bahcall, J. A. Bakken, R. Barkhouser, S. Bastian, E. Berman, W. N. Boroski, S. Bracker, C. Briegel, J. W. Briggs, J. Brinkmann, R. Brunner, S. Burles, L. Carey, M. A. Carr, F. J. Castander, B. Chen, P. L. Colestock, A. J. Connolly, J. H. Crocker, I. Csabai, P. C. Czarapata, J. E. Davis, M. Doi, T. Dombeck, D. Eisenstein, N. Ellman, B. R. Elms, M. L. Evans, X. Fan, G. R. Federwitz, L. Fiscelli, S. Friedman, J. A. Frieman, M. Fukugita, B. Gillespie, J. E. Gunn, V. K. Gurbani, E. de Haas, M. Haldeman, F. H. Harris, J. Hayes, T. M. Heckman, G. S. Hennessy, R. B. Hindsley, S. Holm, D. J. Holmgren, C. hao Huang, C. Hull, D. Husby, S.-I. Ichikawa, T. Ichikawa, Željko Ivezić, S. Kent, R. S. J. Kim, E. Kinney, M. Klaene, A. N. Kleinman, S. Kleinman, G. R. Knapp, J. Korienek, R. G. Kron, P. Z. Kunszt, D. Q. Lamb, B. Lee, R. F. Leger, S. Limmongkol, C. Lindenmeyer, D. C. Long, C. Loomis, J. Loveday, R. Lucinio, R. H. Lupton, B. MacKinnon, E. J. Mannery, P. M. Mantsch, B. Margon, P. McGehee, T. A. McKay, A. Meiksin, A. Merelli, D. G. Monet, J. A. Munn, V. K. Narayanan, T. Nash, E. Neilsen, R. Neswold, H. J. Newberg, R. C. Nichol, T. Nicinski, M. Nonino, N. Okada, S. Okamura, J. P. Ostriker, R. Owen, A. G. Pauls, J. Peoples, R. L. Peterson, D. Petravick, J. R. Pier, A. Pope, R. Pordes, A. Prosapio, R. Rechenmacher, T. R. Quinn, G. T. Richards, M. W. Richmond, C. H. Rivetta, C. M. Rockosi, K. Ruthmansdorfer, D. Sandford, D. J. Schlegel, D. P. Schneider, M. Sekiguchi, G. Sergey, K. Shimasaku, W. A. Siegmund, S. Smee, J. A. Smith, S. Snedden, R. Stone, C. Stoughton, M. A. Strauss, C. Stubbs, M. SubbaRao, A. S. Szalay, I. Szapudi, G. P. Szokoly, A. R. Thakar, C. Tremonti, D. L. Tucker, A. Uomoto, D. V. Berk, M. S. Vogeley, P. Waddell, S. i Wang, M. Watanabe, D. H. Weinberg, B. Yanny, and N. Yasuda, "The sloan digital sky survey: Technical summary," *The Astronomical Journal*, vol. 120, p. 1579, Sept. 2000.

[5] M. K. Cavanagh, K. Bekki, and B. A. Groves, "Morphological classification of galaxies with deep learning: comparing 3-way and 4-way cnns," *Monthly Notices of the Royal Astronomical Society*, vol. 506, no. 1, pp. 659–676, 2021.

[6] J. Cao, T. Xu, Y. Deng, L. Deng, M. Yang, Z. Liu, and W. Zhou, "Galaxy morphology classification based on convolutional vision transformer (cvt)," *Astronomy & Astrophysics*, 2024.

[7] P. H. Barchi, R. de Carvalho, R. R. Rosa, R. Sautter, M. Soares-Santos, B. A. Marques, E. Clua, T. Gonçalves, C. de Sá-Freitas, and T. Moura, "Machine and deep learning applied to galaxy morphology-a comparative study," *Astronomy and Computing*, vol. 30, p. 100334, 2020.

[8] S. Pandya, P. Patel, J. Blazek, *et al.*, "E (2) equivariant neural networks for robust galaxy morphology classification," *arXiv preprint arXiv:2311.01500*, 2023.

[9] H. Domínguez Sánchez, M. Huertas-Company, M. Bernardi, D. Tuccillo, and J. L. Fischer, "Improving galaxy morphologies for SDSS with Deep Learning," *Monthly Notices of the Royal Astronomical Society*, vol. 476, pp. 3661–3676, Feb. 2018.

[10] M. Huertas-Company, J. R. Primack, A. Dekel, D. C. Koo, S. Lapiner, D. Ceverino, R. C. Simons, G. F. Snyder, M. Bernardi, Z. Chen, H. Domínguez-Sánchez, C. T. Lee, B. Margalef-Bentabol, and D. Tuccillo, "Deep Learning Identifies High-z Galaxies in a Central Blue Nugget Phase in a Characteristic Mass Range," *The Astrophysical Journal*, vol. 858, p. 114, May 2018.

[11] T.-Y. Cheng, C. J. Conselice, A. Aragón-Salamanca, N. Li, A. F. L. Bluck, W. G. Hartley, J. Annis, D. Brooks, P. Doel, J. García-Bellido, D. J. James, K. Kuehn, N. Kuropatkin, M. Smith, F. Sobreira, and G. Tarle, "Optimizing automatic morphological classification of galaxies with machine learning and deep learning using Dark Energy Survey imaging," *Monthly Notices of the Royal Astronomical Society*, vol. 493, pp. 4209–4228, Apr. 2020.

[12] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, (New York, NY, USA), p. 607–617, Association for Computing Machinery, 2020.

[13] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.

[14] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real nvp," *arXiv preprint arXiv:1605.08803*, 2016.

[15] D. A. Reynolds, "Gaussian mixture models," in *Encyclopedia of Biometrics*, 2018.

[16] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, "A kernel method for the two-sample-problem," *Advances in neural information processing systems*, vol. 19, 2006.

[17] S. Zhao, J. Song, and S. Ermon, "Infovae: Information maximizing variational autoencoders," *arXiv preprint arXiv:1706.02262*, 2017.

[18] L. Ardizzone, R. Mackowiak, C. Rother, and U. Köthe, "Training normalizing flows with the information bottleneck for competitive generative classification," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7828–7840, 2020.

[19] F. Hvilshøj, A. Iosifidis, and I. Assent, "Ecinn: efficient counterfactuals from invertible neural networks," *arXiv preprint arXiv:2103.13701*, 2021.

[20] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-VAE: Learning basic visual concepts with a constrained variational framework," in *International Conference on Learning Representations*, 2017.

[21] L. Dinh, D. Krueger, and Y. Bengio, "Nice: Non-linear independent components estimation," *arXiv preprint arXiv:1410.8516*, 2014.

[22] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," *Advances in neural information processing systems*, vol. 31, 2018.

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[24] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 694–711, Springer, 2016.

[25] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[26] A. Dey, D. J. Schlegel, D. Lang, R. Blum, K. Burleigh, X. Fan, J. R. Findlay, D. Finkbeiner, D. Herrera, S. Juneau, *et al.*, "Overview of the desi legacy imaging surveys," *The Astronomical Journal*, vol. 157, no. 5, p. 168, 2019.

[27] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[28] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[29] D. Iwasaki, S. Cooray, and T. T. Takeuchi, "Extracting an informative latent representation of high-dimensional galaxy spectra," *arXiv preprint arXiv:2311.17414*, 2023.

[30] M. Mohammadi, J. Mutatiina, T. Saifollahi, and K. Bunte, "Detection of extragalactic ultra-compact dwarfs and globular clusters using explainable ai techniques," *Astronomy and Computing*, vol. 39, p. 100555, 2022.

[31] P. Bhambra, B. Joachimi, and O. Lahav, "Explaining deep learning of galaxy morphology with saliency mapping," *Monthly Notices of the Royal Astronomical Society*, vol. 511, no. 4, pp. 5032–5041, 2022.

[32] Jülich Supercomputing Centre, "JURECA: Data Centric and Booster Modules implementing the Modular Supercomputing Architecture at Jülich Supercomputing Centre," *Journal of large-scale research facilities*, vol. 7, no. A182, 2021.

[33] L. Ardizzone, T. Bungert, F. Draxler, U. Köthe, J. Kruse, R. Schmier, and P. Sorrenson, "Framework for Easily Invertible Architectures (FrEIA)," 2018-2022.

A Appendix / Supplemental Material

A.1 Experiment setting

The Encoder-Decoder architecture can be found in Table 1. The output feature has a size of 32, which is split into $\mathbf{z}_1$ and $\mathbf{z}_2$ of lengths 24 and 8, respectively.

The invertible flow is implemented with `FrEIA` [33]. The architecture has 32 All-in-One blocks, each containing the following sequences: Linear(24, 128) $\rightarrow$ BN $\rightarrow$ LeakyRelu $\rightarrow$ Linear(128, 128) $\rightarrow$ BN $\rightarrow$ LeakyRelu $\rightarrow$ Linear(128, 24).

The model is trained using a learning rate of 0.0001 with an exponential learning rate scheduler, whose factor of learning rate decay is $\gamma = 0.99$. The entire dataset is split into training, validation, and test sets with ratios of 0.7, 0.2, and 0.1, respectively. The training batch size is 128 and the number of epochs is 100. The model is trained on a single A100 80GB GPU.

Layer (Encoder)	Filter/Units	Output Shape	Layer (Decoder)	Filter/Units	Output Shape
INPUT	-	256 x 256 x 3	INPUT	-	32
DOUBLECONV1	8	256 x 256 x 8	FC1	2048	2048
MAXPOOL1	-	128 x 128 x 8	RESHAPE	-	4 x 4 x 256
6 X RESBLOCK1	8	128 x 128 x 8	UPSAMPLE1	256	8 x 8 x 256
DOUBLECONV2	16	128 x 128 x 16	DOUBLECONV1	128	8 x 8 x 128
MAXPOOL2	-	64 x 64 x 16	6 X RESBLOCK1	128	8 x 8 x 128
6 X RESBLOCK2	16	64 x 64 x 16	UPSAMPLE2	128	16 x 16 x 128
DOUBLECONV3	32	64 x 64 x 32	DOUBLECONV2	64	16 x 16 x 64
MAXPOOL3	-	32 x 32 x 32	6 X RESBLOCK2	64	16 x 16 x 64
6 X RESBLOCK3	32	32 x 32 x 32	UPSAMPLE3	64	32 x 32 x 64
DOUBLECONV4	64	32 x 32 x 64	DOUBLECONV3	32	32 x 32 x 32
MAXPOOL4	-	16 x 16 x 64	6 X RESBLOCK3	32	32 x 32 x 32
6 X RESBLOCK4	64	16 x 16 x 64	UPSAMPLE4	32	64 x 64 x 32
DOUBLECONV5	128	16 x 16 x 128	DOUBLECONV4	16	64 x 64 x 16
MAXPOOL5	-	8 x 8 x 128	6 X RESBLOCK4	16	64 x 64 x 16
6 X RESBLOCK5	128	8 x 8 x 128	UPSAMPLE5	16	128 x 128 x 16
DOUBLECONV6	256	8 x 8 x 256	DOUBLECONV5	8	128 x 128 x 8
MAXPOOL6	-	4 x 4 x 256	6 X RESBLOCK5	8	128 x 128 x 8
6 X RESBLOCK6	256	4 x 4 x 256	UPSAMPLE6	8	256 x 256 x 8
FLATTEN	-	2048	DOUBLECONV6	8	256 x 256 x 8
FC1	32	32	6 X RESBLOCK6	8	256 x 256 x 8
			CONV&SIGMOID	3	256 x 256 x 3

Table 1: Layer-wise architecture of the encoder and decoder. The DOUBLECONV layer consists of two sequences of Convolution, Batch Normalization, and ReLU activation layers. A RESBLOCK comprises the following sequence: Conv(1) $\rightarrow$ ReLU $\rightarrow$ Conv(3) $\rightarrow$ ReLU $\rightarrow$ Conv(3) $\rightarrow$ ReLU $\rightarrow$ Conv(1). Upsampling is performed using linear interpolation.

A.2 Quantitative Analysis

The classification report and the confusion matrix can be found in Table 2 and Figure 5.

Based on the results, the model performs best for the round smooth class, achieving an F1-Score of 0.93 (and 0.91 for the entire dataset). Conversely, the disturbed class has the lowest performance, with an F1-Score of 0.41 (0.45 for the entire dataset). This outcome is expected since the morphology of a round smooth galaxy is relatively simple, resembling a compact object, whereas a disturbed galaxy's structure is much more complex. The confusion matrix indicates that disturbed galaxies are frequently misclassified as unbarred loose spiral galaxies. Additionally, unbarred loose and tight spiral galaxies are often difficult to distinguish from one another.

Class	Precision	Recall	F1-Score	Support
Disturbed	0.40 (0.48)	0.41 (0.42)	0.41 (0.45)	94 (1081)
Merging	0.79 (0.81)	0.90 (0.90)	0.84 (0.85)	188 (1853)
Round Smooth	0.94 (0.91)	0.93 (0.90)	0.93 (0.91)	254 (2645)
In-between	0.91 (0.92)	0.85 (0.85)	0.88 (0.88)	200 (2027)
Cigar	0.88 (0.86)	0.54 (0.59)	0.67 (0.70)	28 (334)
Edge-on with Bulge	0.87 (0.85)	0.91 (0.90)	0.89 (0.88)	191 (1873)
Edge-on without Bulge	0.94 (0.91)	0.82 (0.88)	0.88 (0.89)	147 (1423)
Unbarred Loose Spiral	0.64 (0.65)	0.75 (0.75)	0.69 (0.70)	277 (2628)
Unbarred Tight Spiral	0.79 (0.80)	0.63 (0.66)	0.70 (0.72)	185 (1829)
Barred Spiral	0.81 (0.82)	0.82 (0.84)	0.81 (0.83)	209 (2043)
Accuracy			0.80 (0.80)	
Macro Avg	0.80 (0.80)	0.75 (0.77)	0.77 (0.78)	1773 (17736)
Weighted Avg	0.81 (0.81)	0.80 (0.80)	0.80 (0.80)	1773 (17736)

Table 2: Classification report showing precision, recall, f1-score, and support for different classes. The numbers in and out of the parentheses are for the entire dataset and test set, respectively.

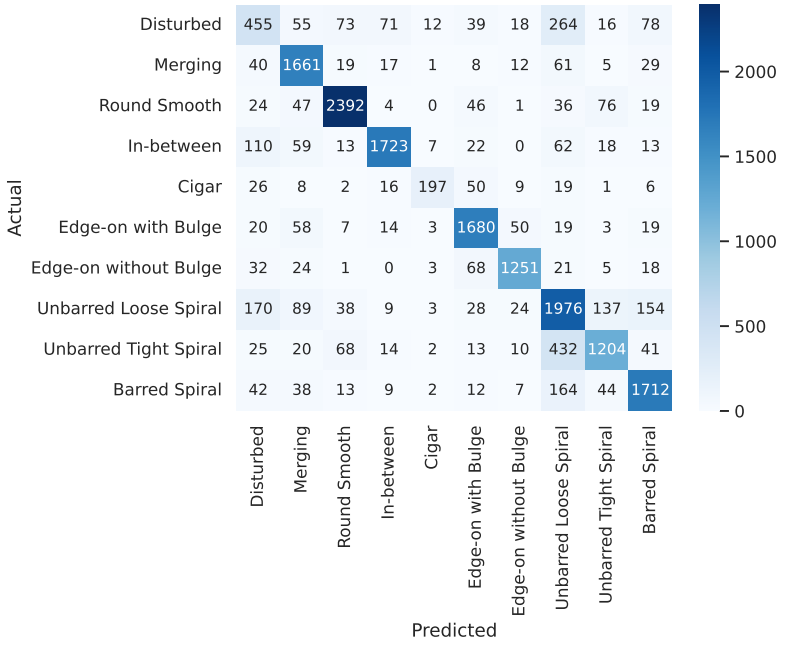

Figure 5: Confusion matrix of the trained model for the entire dataset.

A.3 Complete Counterfactual Visualization

The complete version of the counterfactual visualization is illustrated in Figure 6. The diagonal entries display the original input images from different galaxy classes, while the off-diagonal entries illustrate how these images are transformed into counterfactuals that resemble other galaxy classes. For instance, in the first row, the original "Disturbed" galaxy is modified to appear as though it belongs to the "Merging," "Round Smooth," "In-between Round Smooth," and other classes. It demonstrates the changes required for each galaxy to be reclassified as a different type, offering insights into the decision boundaries and the alterations necessary to cross them. The diversity and realism of the transformations reflect the model's understanding of the morphological features that differentiate galaxy classes.

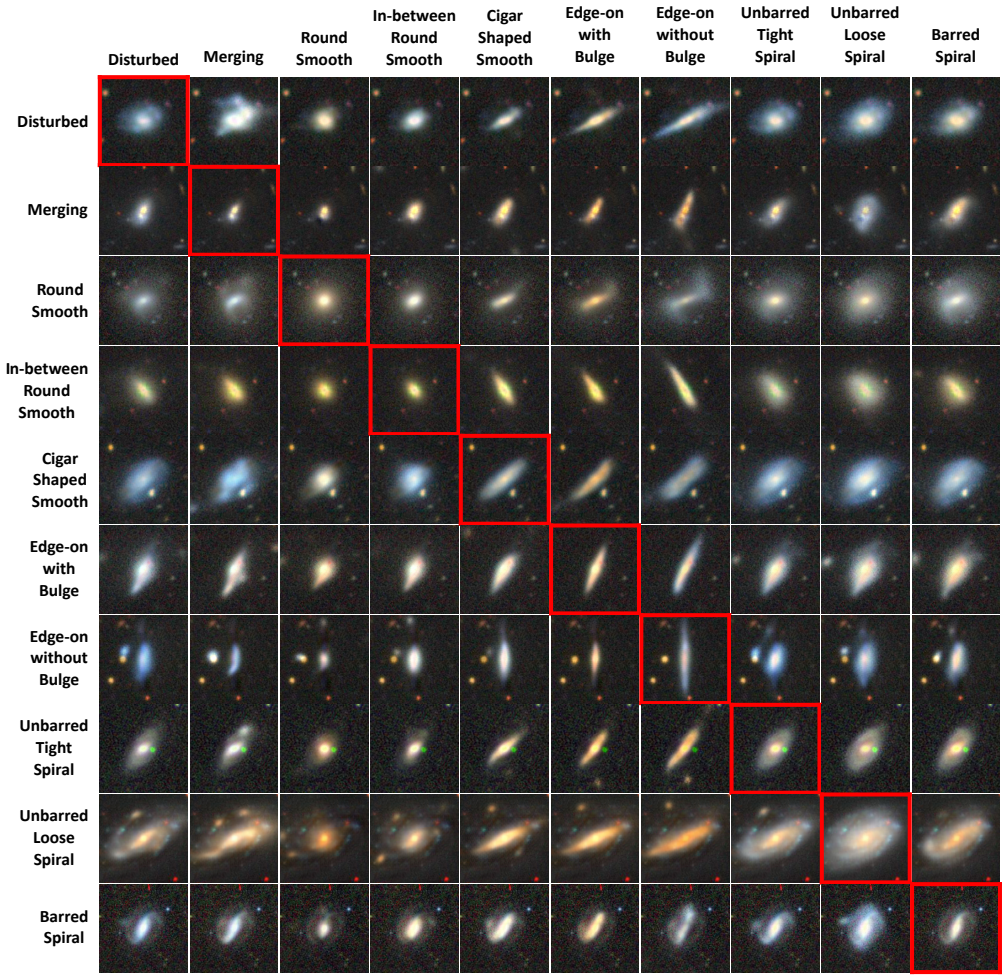

Figure 6: The original images are aligned across the diagonal (red boxes), and the counterfactual images are displayed in the same row their classification according to the labels above.

To emphasize the differences between the original and counterfactual images, we superimpose the difference heatmap $\mathbf{x}^{cf} - \tilde{\mathbf{x}}$ onto the counterfactuals. This approach further highlights the characteristic features of each class. For instance, converting any class to a 'Round Smooth' galaxy (shown in the third column) involves reducing the surrounding details while enhancing the central object, indicating that round smooth galaxies are compact in nature.

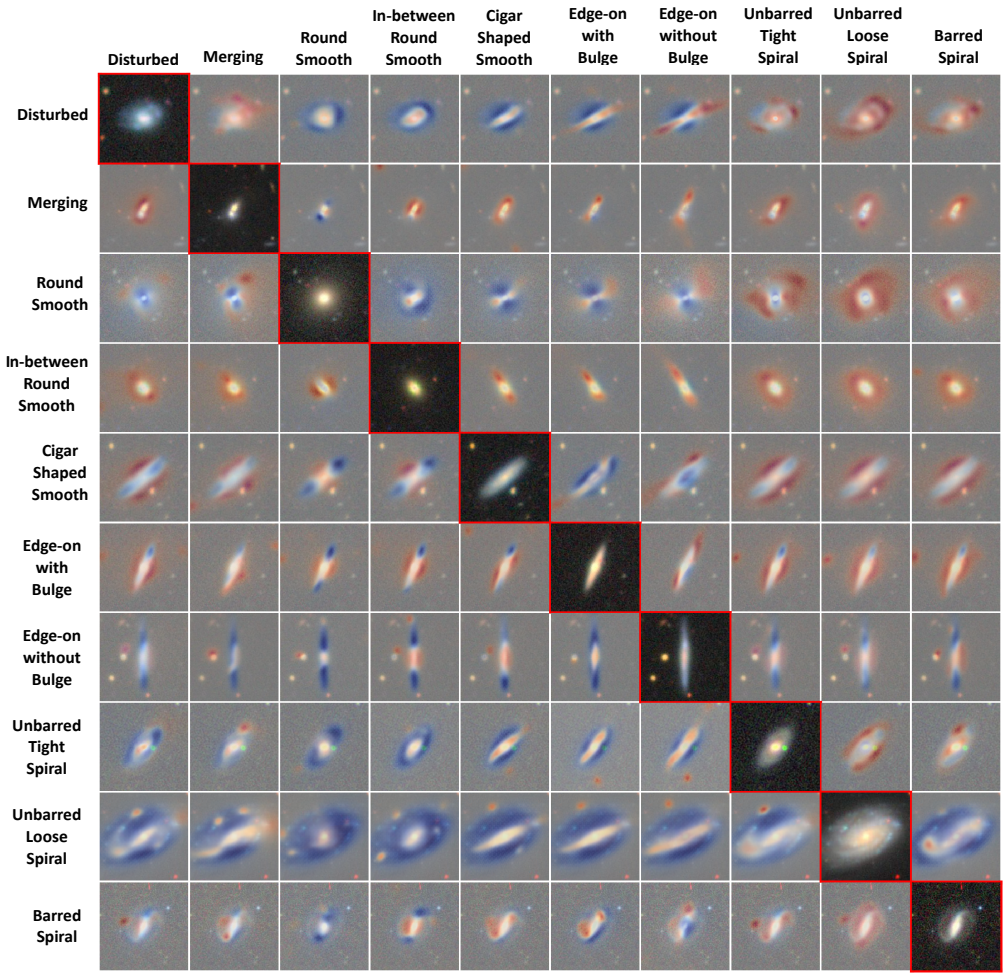

Figure 7: The original images are arranged along the diagonal (highlighted in red boxes), while the counterfactual images are shown in the same row corresponding to their classification as indicated by the labels above. A difference heatmap $\mathbf{x}^{cf} - \tilde{\mathbf{x}}$ is overlayed on the counterfactual images. Reddish colors represent positive values, while bluish colors indicate negative values.