Galaxy Image Classification and Clustering

Neil Pandya and Alejandro Orellana

Abstract

This paper will explore several machine learning methods for classifying and clustering galaxy images, incorporating dimensionality reduction techniques in the process. Our objective is to explore different methods to predict whether a galaxy is spiral or not (labeled in the data as 1 and 0, respectively) and to predict the galaxy class of the image: either ring, fan, distortion, or tail (labeled in the data as 0, 1, 2, and 3, respectively).

In a supervised learning manner, we will use and compare two different Convolutional Neural Networks (CNNs) inspired by the original AlexNet and VGG CNN architectures to predict spiral status. Additionally, we will use a similar VGG CNN architecture to predict galaxy class. Our findings show that both the AlexNet and VGG CNN architectures are highly accurate in predicting spiral status for testing data (82.7% and 81.1% accurate, respectively). Neither CNN architecture significantly outperforms the other for spiral status prediction. Similarly, a supervised VGG CNN with an accuracy of 59% was used for galaxy class prediction. However, the model also has a low macro F1 score of 0.44, which can be due to the unbalanced classes in the training data and the inability to capture particular details from the kernels.

In an unsupervised learning manner, we will use CNN feature extraction with a pre-trained VGG-16 Neural Network, followed by PCA dimensionality reduction (to 50 dimensions) to perform K-Means clustering and identify optimal 2 clusters. This will ideally represent the distinction between spiral and non-spiral galaxies. Additionally, we will perform a similar clustering instead of latent space representation via autoencoding techniques to identify an optimal 4 clusters, ideally to represent the distinction between galaxy classes. In both cases, we will visualize cluster arrangement by implementing further t-SNE or PCA dimensionality reduction to 2 dimensions. Here, our findings show that unsupervised clusters developed by VGG-16 features are unable to properly associate one cluster as spiral and the other as not spiral (Adjusted Rand Index = -0.0023, Normalized Mutual Information = 0.0044). Similarly, for the galaxy class, an unsupervised VGG Convolutional Autoencoder Neural Network showed an Adjusted Rand Index of 0.0499 and Normalized Mutual Information of 0.0374, also demonstrating poor alignment of clustered and true galaxy class labels. The feature vectors given by the Autoencoder CNN were also reduced to 50 principal components, followed by k-means clustering. It is also important to note that this model has the same limitations as the VGG CNN for galaxy classification, as well as a longer training time.

It is clear that implementing and training CNN architectures on training data to apply to testing data is a better approach than unsupervised clustering, as these models are better able to distinguish and understand complex image differences in both spiral status and galaxy class.

Distribution of Work

Neil Pandya:

- AlexNet and VGG CNNs to predict spiral status
- CNN feature extraction & clustering with PCA and t-SNE, predicting spiral status

Alejandro Orellana:

- VGG CNN to predict galaxy type
- Autoencoding latent space feature extraction & clustering with PCA, predicting galaxy type

Dataset Information

The GitHub that stores the data is from Jiaweimmiao's galaxy_classification_dataset repository. It is a subset based on the GALAXY CRUISE 1 (GC1) source catalog that already has train, validation, and test splits. It has a CSV of information for each image, with three versions of different image qualities, but an unbalanced amount of data for each class. The classes of galaxies are either ring (0), fan (1), distortion (2), or tail (3), determined based on which class had the highest number of votes (highest probability), as identified by the human classification that the data went through. There is also a column called "galaxy_type" that contains the probability that the galaxy is a spiral, as determined by another machine. Separately, we developed a column "Is_Spiral" using a threshold of 0.65 to more confidently classify true spiral galaxies and make the classification frequencies for each spiral status more balanced.

Previous Galaxy Image Classification

As a community science project, the original dataset was used in the paper *Galaxy Cruise: Deep Insights into Interacting Galaxies in the Local Universe*. The goal is to have a model to classify the Galaxy morphology, which individuals have traditionally done. The paper discussed that the human-based classification was valuable, but the image depth is limited, as people can miss smaller details. Therefore, the paper used different imaging qualities to conclude that interactions are enhanced between active galactic nuclei and star formation for distinctions of merger stages and feature types.

Exploratory Data Analysis

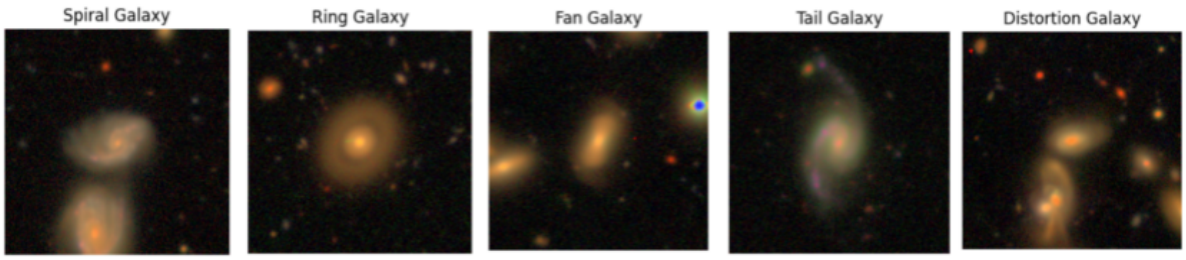

Figure 1, Representation of Different Galaxy Types

It is important to visualize the different galaxy classes and a clear example of a spiral galaxy for our analysis. Figure 1 shows the images of the most certain galaxy classifications within the entire dataset, given by the highest probability within each characteristic column. Analyzing the images at first glance, we see visual distinctions between each of the four classes (ring, fan, tail, and distortion). However, the fact that some images contain multiple galaxies may make it difficult for the model to decide which galaxy is being represented in the classification. Additionally, it may be hard to distinguish between the four classes by the nature of these galaxy images. Our machine-learning techniques will aim to explore how well these different classes can be identified.

Figure 2 shows the spread of the probabilities of each galaxy class and galaxy_type (which determines the confidence in whether it is a spiral galaxy). The distribution shows that the data is mostly confident that a galaxy is not a ring or fan, while mostly under 70% sure for any classification. Therefore, there is a classification based on the highest probability of each image, which is not determined with certainty. This leads to the use of the data for the current classification method being unreliable, while galaxy_type has more certainty in its class.

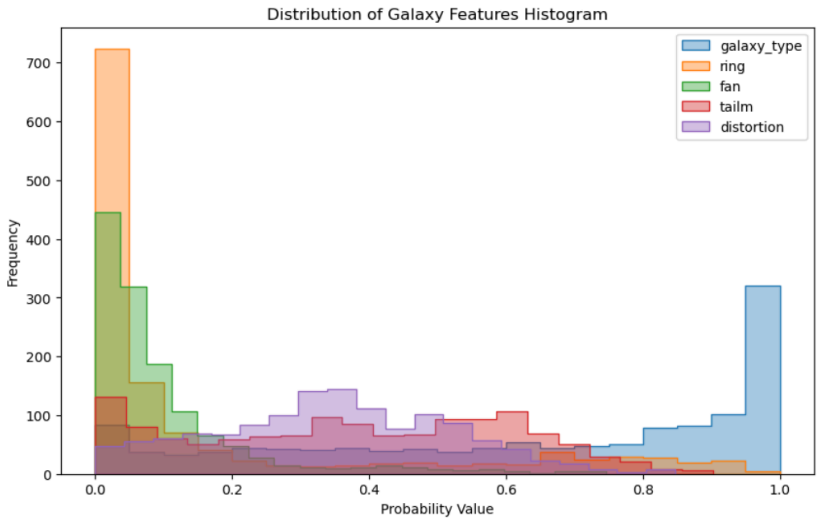

Figure 2, Histogram of Probabilities by Galaxy Type

Detecting Galaxy Spiral

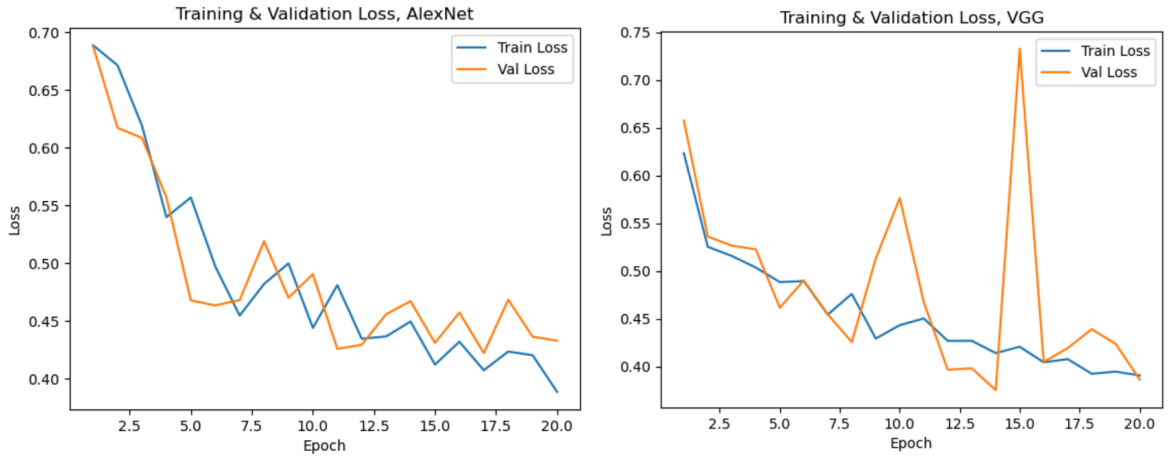

Figure 3: Training and Validation Loss by AlexNet and VGG Architecture

Two important CNN architectures for image classification include AlexNet and VGG (supervised learning mechanisms). The AlexNet architecture, as described in the original paper (Krizhevsky et al.), featured several techniques to process many high-resolution ImageNet images into 1000 categories, such as including larger kernels earlier on, dropout in fully connected layers to prevent network overfitting, and an overall deep architecture with five convolutional layers compared to other networks at the time. Our version of the AlexNet model incorporates the above. It mimics a similar architecture to the paper, initially reducing spatial dimensions and later using smaller kernels and more filters to capture detailed image features. Max Pooling is incorporated after some ReLU activations to reduce spatial size and retain the most important features. We also used normalization and dropout, as described in the original paper, to help improve generalization and prevent overfitting.

On the other hand, some key aspects of VGG architectures (Simonyan et al.) include consistently stacked 3x3 convolution kernels, increased filters as depth increases, and uniform 2x2 max pooling. Similarly, we incorporated these principles into our own VGG model, using multiple small kernels stacked to emulate larger receptive fields and a deeper network compared to AlexNet. The consistent architecture and fixed kernel sizes allow for a more uniform design, simplifying implementation while achieving strong performance on image classification tasks.

We performed a 70-15-15 train-validation-test split due to a limited set of 1302 images for the neural networks. Figure 3 demonstrates the relationship between average training loss (binary cross entropy with logits) up to 20 epochs, which is as computationally expensive as our best computer GPU can handle. In both architectures, the training and validation losses converge in a roughly similar pattern. Some spikes in validation loss for the VGG architecture model reflect random and natural fluctuations due to some batches potentially containing unusual examples or unstable parameters early on in the training process.

In addition to supervised learning, we implemented an unsupervised learning method using the entire dataset to identify whether spiral status can be identified through the natural separation of image features into 2 clusters. To do so, we used a fully pre-trained VGG-16 neural network (trained on over 1.2 million images and around 1000 object categories) that will assist in detecting specific features in image data, which is useful for developing a feature vector for an image in an unsupervised manner. In our case, feature vectors can help identify various galaxy patterns, such as intricate edges and curves. We generated a feature vector for each galaxy image using the pre-trained neural network, followed by PCA with 50 components to reduce dimensionality in the features. We can properly implement K-Means clustering at this point without the curse of dimensionality from highly-dimensional feature vectors.

Spiral Detection Model Performances and Comparisons

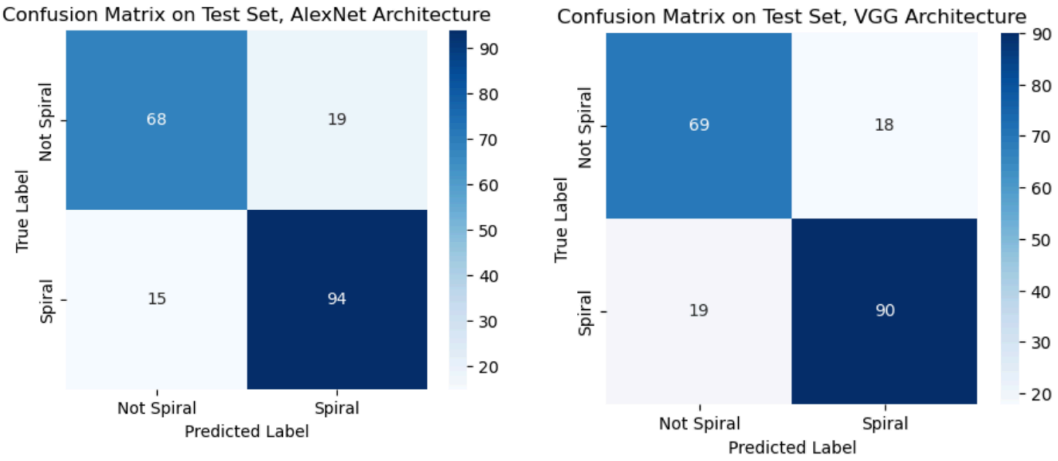

Figure 4: Confusion Matrices by CNN Model

Our classification results for spiral detection from both CNN architecture models show strong performance, as the accuracies on the testing images are roughly 82.7% (F1 Score = 0.8468) and 81.1% (F1 Score = 0.8295) for the AlexNet and the VGG architectures, respectively. There is likely to be some error due to the fact that we used a threshold of 0.65 to deem an image as 1 = "Spiral" versus 0 = "No Spiral," where images close to this threshold are likely to be confused as the other. However, this error is minimal as both CNN architectures demonstrate similar performances that indicate a spiral galaxy can be distinguished from non-spiral galaxies. Figure 4 shows the confusion matrix for each network's performance on unseen testing data, where the majority of values are on the true diagonal.

On the other hand, the results from our unsupervised model do not provide evidence that the VGG-16 model can distinguish between two feature clusters representing "spiral" and "no spiral" elements. We achieved an Adjusted Rand Index (ARI) of approximately -0.0023 and a Normalized Mutual Information (NMI) value of 0.0044. Both scores indicate poor and minimal alignment between the two generated clusters and the true spiral status, as visualized in Figure 5 through further t-SNE dimensionality reduction of the numerical features, which shows significant overlap between the true labels. Even though the image features are heavily compressed to two dimensions, the resulting clusters still show poor separation of true spiral statuses, as reflected by the ARI and NMI metrics. Overall, we expect a CNN to work better with spiral detection because the feature vectors may not work well with galaxy images, as they were trained on everyday objects. In addition, having a properly trained neural network on complex and intricate galaxy data will be more helpful in identifying similar patterns of galaxies compared to unsupervised learning.

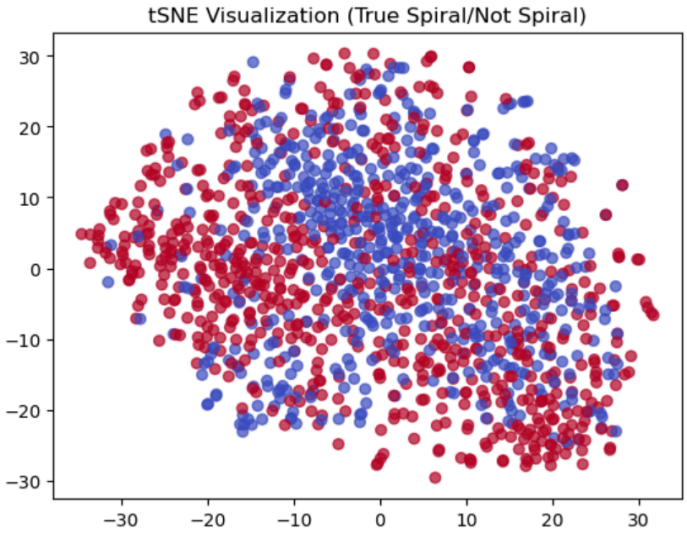

Figure 5: tSNE Visualization of spiral status, 2 components

Galaxy Morphology Classification

VGG Convolutional Neural Network (Supervised Learning)

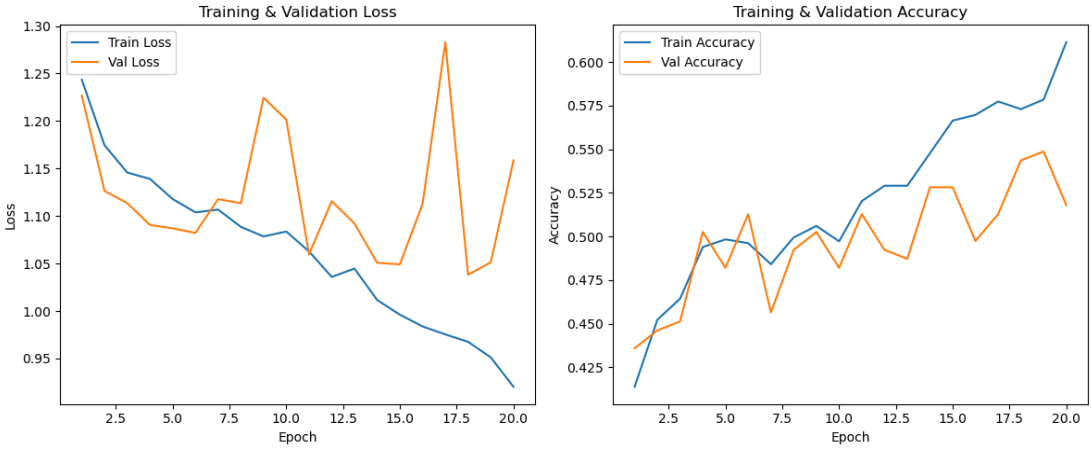

Figure 6: Epochs over Loss and Accuracy for VGG Convolutional Neural Network

The sub-dataset from Jiaweimmiao's galaxy_classification_dataset repository had a column called 'label1', which was used for the convolutional neural network model to classify images of galaxies for morphology. VGG architecture was chosen for the CNN because it has been discussed that, compared to AlexNet, it has a deeper network, allowing it to recognize patterns and details better. The differences between the four classes (ring, fan, distortion, and tail) should be better distinguished with VGG, although the characteristics may not be as distinct. Figure 6 shows that the training loss could converge with further epochs, but the validation loss has inconsistent spikes for reasons similar to those of the first VGG model. As the accuracy slowly increases, there is a slight increase in performance for each epoch.

The same process was used as the VGG before, but there were also preprocessing transformations of the images to highlight the characteristics. It produced a model that predicted the true ring, distortion, and tail as themselves but failed to identify any fan galaxies. The model predicted most fan-classified galaxies as the distortion class and had no false positives, as shown in Figure 7. Although the failure of fan classification was significant, there was moderate success in the other classes, with the best-performing class being tail, having an ~80% correct classification. However, it may be due to how many images have been classified as tail, with the following individual F1 scores: ring: 0.55, fan: 0, distortion: 0.5, tail: 0.7, and an overall accuracy of 59.2% with a macro F1 score of 0.44. Providing a probability of correctly classifying an image is larger than double that of random selection (25%). However, there are no fan classifications and unsatisfactory F1 scores, so this model is slightly reliable for the ring, distortion, and tail class.

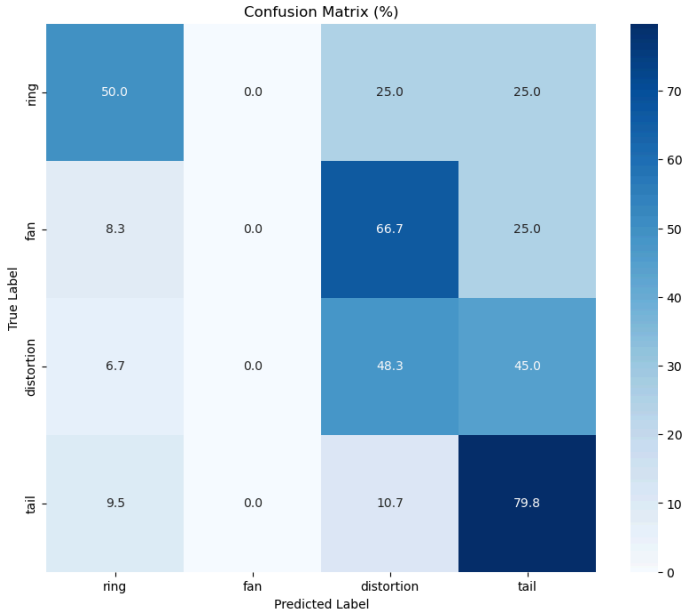

Figure 7: Confusion Matrix of Convolutional Neural Network

VGG Convolutional Autoencoder Neural Network (Unsupervised Learning)

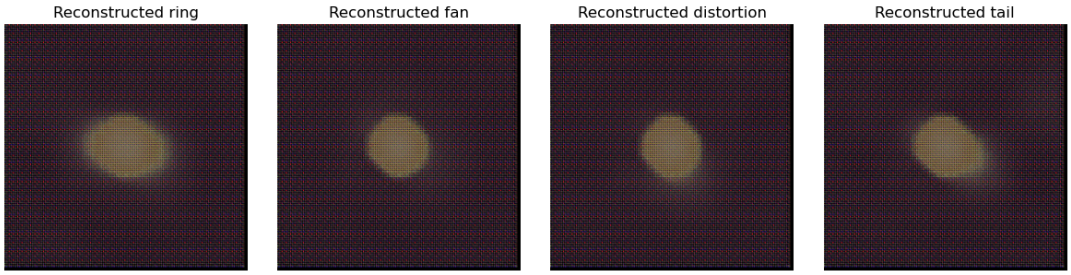

Figure 8: Reconstruction of each class

To explore other methods for galaxy morphology classification, unsupervised learning was used to find whether other approaches are better. Using the VGG architecture as a basis, it was expanded as a convolutional autoencoder neural network (CNNAE) to reduce dimensionality and then deconvoluted to create a reconstruction of the images. The images in Figure 8 are the preprocessed images and the reconstructed versions from the CNNAE model. Generally, each classification is similar, with minor differences that could be mistaken for noise.

The second stage of the model involves using the latent features after the CNNAE was trained to extract them from the data to be used in k-means, but due to the curse of dimensionality, PCA was used to reduce it to 50 dimensions from 911. There are some bulges for the ring and tail, but through the performance metrics, it is not enough to distinguish between the galaxies. To visualize the performance of the latent space in Figure 9, there is a plot of the first two principal components (PCA 1: 33.59% variance explained, PCA 2: 9.63% variance explained), with a total variance explained being 43.22%. Therefore, the shape of the plot would be reasonable to interpret that each cluster's spatial relation has no classification structure. This is confirmed by the Adjusted Rand Index of 0.0499 and Normalized Mutual Information of 0.0374, which shows weak cluster alignment with the true galaxy class labels. This low performance is likely due to the uneven distribution of each class in the data and the similarity between classes due to insufficient data. There may be improvement if there is more data for each class with more certainty in the classifications, with adjustments to the kernel to capture more details.

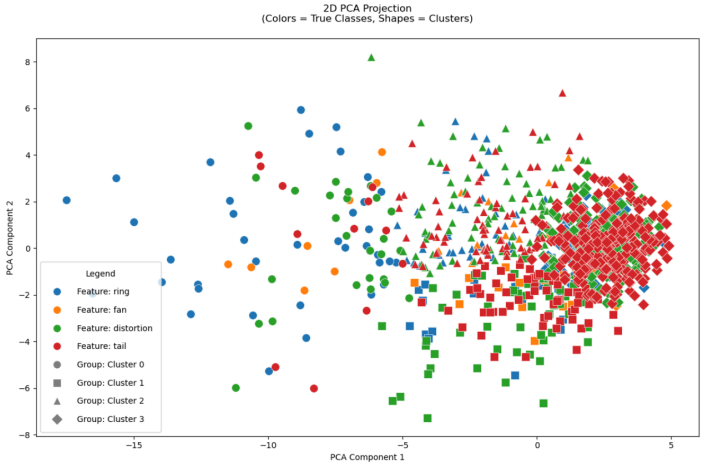

Figure 9: PCA 1 and PCA 2 of autoencoder latent space scatter plot

Model Performance Comparison

The classes were more distinguished in the supervised CNN than the unsupervised CNNAE because the supervised CNN could better understand complex patterns in the image data. In contrast, the CNNAE is prone to noisy features with limited data. However, we found the performance of each not to be reliable enough, as proven by the low macro F1 score, Adjusted Rand Index, and Normalized Mutual Information. Therefore, the issue may lie in the data, as it was not representative of all classes. Overall, the supervised model training went faster than the unsupervised model and did moderately well in predicting three of the four classes. In conclusion, the supervised model would be more effective, accounting for all the limitations and time.

Bibliography

Tanaka, Masayuki, et al. "GALAXY CRUISE: Deep Insights into Interacting Galaxies in the Local Universe." Publications of the Astronomical Society of Japan, vol. 75, no. 5, 2023, pp. 986-1010, doi:10.1093/pasj/psad055.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *ImageNet classification with deep convolutional neural networks*. *Advances in neural information processing systems*, 25.

Simonyan, K., & Zisserman, A. (2015). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. *International Conference on Learning Representations (ICLR)*.