Postprint: Galaxy Morphological Classification Based on FPN-ViT

Authors:

Date: 2024-06-03T00:00:00+00:00

Abstract

With the advancement of artificial intelligence technology, significant progress has been made in galaxy morphology classification using deep learning methods. However, there remain deficiencies in classification accuracy, automation, and the representation of spatial features of galaxies. The Vision Transformer (ViT) model currently demonstrates good robustness in galaxy morphology classification, but it has certain limitations when processing multi-scale images. Therefore, we propose introducing Feature Pyramid Networks (FPN) into the ViT model (FPN-ViT) for galaxy morphology classification research. The results show that the average accuracy, precision, recall, and F1-score of galaxy morphology classification based on the FPN-ViT model all exceed 95%, with each metric showing a certain degree of improvement compared to the traditional ViT model. Additionally, by adding varying degrees of Gaussian noise and salt-and-pepper noise to the original galaxy images, we verify that the FPN-ViT model can also achieve good classification performance on low signal-to-noise ratio data. Furthermore, to comprehensively evaluate the model, we employ the t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm for visual analysis of the classification results, which allows for a more direct observation of the FPN-ViT model' s effectiveness in galaxy morphology classification. Therefore, applying FPN networks to ViT models for galaxy morphology classification research represents a novel attempt that holds significant importance for future studies.

Full Text

Preamble

Acta Astronomica Sinica, Vol. 65, No. 3, May 2024
doi: 10.15940/j.cnki.0001-5245.2024.03.010

Research on Galaxy Morphology Classification Based on FPN-ViT

CAO Jie[1], XU Ting-ting[1], DENG Yu-he[1], LI Guang-ping[1], GAO Xian-jun[1], YANG Ming-cun[1], LIU Zhi-jing[1], ZHOU Wei-hong[1,2]

(1 School of Mathematics and Computer Science, Yunnan Minzu University, Kunming 650504)
(2 Key Laboratory of the Structure and Evolution of Celestial Objects, Chinese Academy of Sciences, Kunming 650011)

Abstract

With the development of artificial intelligence technology, significant progress has been made in galaxy morphology classification using deep learning methods. However, challenges remain in classification accuracy, automation, and the spatial feature representation of galaxies. The Vision Transformer (ViT) model currently demonstrates good robustness in galaxy morphology classification, but exhibits limitations when handling multi-scale images. Therefore, this study proposes introducing Feature Pyramid Networks (FPN) into the ViT model (FPN-ViT) for galaxy morphology classification research. The results show that all evaluation metrics—including average accuracy, precision, recall, and F1-score—exceed 95% when using the FPN-ViT model for galaxy morphology classification, representing improvements over the traditional ViT model across all indicators. Additionally, by adding varying levels of Gaussian noise and salt-and-pepper noise to original galaxy images, we verify that the FPN-ViT model achieves favorable classification performance even with low signal-to-noise ratio data. Furthermore, to comprehensively evaluate the model, we employ t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm for visual analysis of classification results, providing more intuitive insight into the effectiveness of the FPN-ViT model for galaxy morphology classification. Thus, applying FPN networks to ViT models for galaxy morphology classification represents a novel approach with significant implications for future research.

Keywords: methods: data analysis, techniques: image processing, galaxies: general

1 Introduction

Galaxy morphology classification is a systematic approach in astronomy for describing and categorizing the appearance and structure of different galaxies. By classifying galaxies according to their distinct morphological features, we can investigate physical processes of galaxy formation and evolution and reveal the large-scale structure of the universe through examination of characteristics such as shape, size, density, and structure. Observing numerous galaxies and grouping them into different morphological categories enables us to understand the diverse mechanisms of galaxy formation and evolutionary pathways. For instance, early galaxy morphology classification systems included two primary categories: spiral galaxies and elliptical galaxies, revealing two distinct formation mecha-

nisms. Spiral galaxies typically form through the collapse and rotation of gas and dust clouds, while elliptical galaxies likely originate from galaxy collisions and mergers. Moreover, galaxy morphology classification correlates with both the physical properties of galaxies and their environments. Regarding physical properties, studies show that spiral galaxies are generally rich in young stars and interstellar matter, whereas elliptical galaxies tend to contain older stars and less interstellar material. Concerning environmental correlations, research indicates that in high-density environments such as galaxy clusters, interactions between galaxies are more pronounced, potentially leading to morphological transformations—for example, galaxy interactions may drive the transition to elliptical morphologies. Consequently, galaxy morphology classification provides crucial clues about galactic environments, helping us understand the distribution and evolution of matter within galaxies. Finally, by studying the distribution and clustering of different galaxy categories across the universe, we can reveal the cosmic web structure, galaxy clusters, and superclusters, which is fundamental to understanding basic physical questions about cosmic evolution. Therefore, accurate classification of galaxies by morphological features is essential for subsequent data analysis and mining.

Galaxy morphology can be divided according to different classification criteria. Among these, the Hubble sequence, proposed by Hubble in 1926, stands as one of the most famous early standards for galaxy morphology classification. The Hubble sequence remains highly relevant today, as it correlates closely with physical parameters such as neutral hydrogen mass, integrated galaxy colors, galaxy luminosity, and environment. The Galaxy Zoo project (GZ), launched in 2007, adopted classification standards based on the Hubble sequence. This citizen science project, originating from the Galaxy Zoo-the Galaxy Challenge competition on the Kaggle platform, engages volunteers to classify galaxy morphology according to the Hubble sequence. Volunteers categorize galaxies into different classes based on structural features, presence of spiral arms, bar shapes, and other attributes, making the Hubble sequence a classic framework that continues to hold significant meaning for studying galaxy evolution and formation.

In recent years, advances in astronomical observation equipment have continuously improved survey depth and detection efficiency. Projects such as the Sloan Digital Sky Survey (SDSS), the Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST), and infrared space telescopes like the James Webb Space Telescope (JWST) have generated massive amounts of spectroscopic and imaging data, necessitating more automated and intelligent classification methods to meet the demands of large-scale galaxy image processing.

With the continuous development of deep learning technology, related algorithms have been widely applied in astronomy, with galaxy morphology classification based on deep learning emerging as a research hotspot. Zhu et al. proposed an improved model based on deep residual networks (ResNet) called ResNet-26, featuring 26 layers (25 convolutional layers and one fully connected layer) that automatically extract, recognize, and classify galaxy morphological

features. Experimental results demonstrated that ResNet-26 achieved a classification accuracy of 95.12%, exhibiting superior performance compared to other popular Convolutional Neural Network (CNN) models. Ai et al. applied the EfficientNet model to galaxy morphology classification, with results showing that all evaluation metrics exceeded 96.6% when using EfficientNet-B5, representing significant improvement over the ResNet-26 model. Wei et al. proposed a contrastive learning approach that combines vision transformers with convolutional networks in the feature extraction layer to provide rich semantic representation through multi-level feature fusion, achieving test set accuracies of 94.7%, 96.5%, and 89.9% on Galaxy Zoo 2, SDSS Data Release 17, and Galaxy Zoo DECaLS datasets, respectively. He et al. designed a multi-channel deep residual network framework called ResNet-Core based on the 50-layer ResNet-50 architecture (49 convolutional layers and one fully connected layer). By incorporating convolutional kernel variance control technology tailored to spectral and galaxy image characteristics, they effectively extracted contour and detail features, improving average precision beyond the contemporary state-of-the-art ResNet-50 and demonstrating better classification performance and robustness. Hui et al. applied Densely Connected Convolutional Networks (DenseNet) to galaxy morphology classification, with DenseNet-121 (121 layers: 120 convolutional layers and one fully connected layer) achieving 91.79% accuracy—correctly classifying 2,794 out of 3,044 test images—along with 79.92% precision, 73.20% recall, and 75.48% F1-score. Li et al. proposed a Multi-Scale Convolution Capsule Network (MS-CCN) model that extracts multi-scale hidden features from galaxy images through multi-branch structures, achieving 97% accuracy, 96% precision, 98% recall, and 97% F1-score under macro-averaging on the Galaxy Zoo 2 dataset.

Since 2021, Transformer models in deep learning have achieved tremendous success in Natural Language Processing (NLP) by introducing self-attention mechanisms for global context modeling of sequential data. Inspired by this, the Google team developed a new image classification architecture called Vision Transformer (ViT). The ViT model has since been widely applied to classification tasks across various domains. Gheflati et al. applied ViT to medical imaging for breast ultrasound classification, demonstrating superior performance compared to CNN models. Gao et al. employed ViT in an AI medical image analysis competition for COVID-19 diagnosis, classifying COVID-19 versus non-COVID-19 cases from CT scans, with ViT outperforming contemporary DenseNet models and achieving an F1-score of 0.76. Tanzi et al. used ViT architecture to classify different fracture types, comparing it with classical CNNs and multi-level CNN structures, with ViT correctly predicting 83% of test images and outperforming CNN models.

Following the release of Vision Transformer, many researchers have proposed improvements. Chu et al. introduced Conditional Position Encodings for Vision Transformers (CPVT), using Conditional Position Encodings (CPE) instead of ViT' s predefined position embeddings, enabling Transformers to handle arbitrarily sized images without interpolation. Han et al. proposed the Transformer-iN-Transformer (TNT) model, which models patch-level and pixel-level repre-

sentations using external Transformer modules for image patch embeddings and internal Transformer modules for modeling relationships between pixel embeddings. Yuan et al. proposed the Tokens-To-Token (T2T) model, primarily improving ViT by concatenating multiple tokens within sliding windows into a single token. Wang et al. introduced the Pyramid Vision Transformer (PVT), adopting a multi-stage design for Transformers (without convolution) similar to multi-scale approaches in CNNs, which benefits dense prediction tasks. Wu et al. proposed the Convolutional Vision Transformer (CvT), introducing convolutions into the ViT model to enhance performance. The CvT model achieved 87.7% Top-1 accuracy on the public ImageNet-1k dataset, surpassing ViT' s 76% accuracy on the same dataset.

Building upon these ViT improvements, this paper proposes introducing Feature Pyramid Networks (FPN) into the ViT model to enhance performance. The paper is organized as follows: We first discuss FPN and traditional ViT network structures, then present the basic framework and principles of the FPN-ViT architecture in Section 2. Section 3 introduces the dataset used in our experiments and describes data augmentation for categories with fewer samples. Section 4 analyzes and discusses classification results from the FPN-ViT model, compares them with similar work, and presents visualization analysis of the classification results. Finally, Section 5 summarizes our work.

2 Methodology

This paper proposes a method that introduces FPN into the ViT model for galaxy image classification. Traditional ViT models have limitations when processing multi-scale images because they can only handle fixed-size input images. Feature pyramids provide a multi-scale feature representation method that captures local and global information in images by applying filters of different scales to convolutional feature maps at different levels. We combine FPN with the ViT model to improve its performance, with the overall structure shown in [Figure 1: see original paper].

2.1 Feature Pyramid Networks (FPN)

Feature Pyramid Networks (FPN) are multi-scale feature extractors proposed by Lin et al. in 2017. FPN constructs a pyramid-structured feature representation by adding lateral connections and upsampling operations to fuse feature maps from different network levels. Specifically, FPN performs upsampling on low-level feature maps to match the dimensions of high-level feature maps, then combines them through element-wise addition to obtain fused feature maps. This approach preserves high-level contextual information while effectively utilizing low-level detail features. The FPN structure is illustrated in [Figure 2: see original paper], where Conv2d denotes 2D convolution operations, $1\$\times 1 indicates a convolution kernel of 1 pixel height and 1 pixel width, 3\times 3 indicates a 3\times\$3$ kernel, s=1 represents a stride of 1, and numbers like 112 and 96 refer to feature map dimensions and channel counts.

The specific computational steps of the FPN structure are as follows:

1. **Input**: An image is taken as input.
2. **Feature extraction**: CNN is used for feature extraction, with appropriate network levels selected as starting points—typically deeper levels that have larger receptive fields but lower resolution.
3. **Top-level features**: Starting from the deepest level of feature extraction, a 1$×$1 convolution (pointwise convolution) is applied to generate feature maps with fewer channels, reducing computational cost and preparing for subsequent operations.
4. **Upsampling**: For feature maps with lower resolution than the top-level features, upsampling operations enlarge their dimensions to match the adjacent shallower feature maps. Common upsampling methods include bilinear interpolation and transposed convolution.
5. **Fusion**: The upsampled feature maps are combined with adjacent shallower feature maps through element-wise addition, merging detail features with contextual features to create a multi-scale feature pyramid.
6. **Iteration**: Steps 4 and 5 are repeated until reaching the topmost feature map.
7. **Output**: The final feature pyramid can be used for tasks such as object detection and semantic segmentation. For object detection, additional network layers typically predict object locations and categories.

Examples of feature maps at each FPN level and the final merged feature map are shown in [Figure 3: see original paper], where Level 1 to Level 4 correspond to feature maps generated by different depths of the CNN layers in the FPN architecture, which are ultimately combined through upsampling and fusion operations to produce the merged feature map.

2.2 Vision Transformer

In 2017, Google' s machine translation team published "Attention is All You Need" at the Conference on Neural Information Processing Systems (NIPS), pioneering a simple network architecture that completely abandoned CNNs and Recurrent Neural Networks (RNNs) in sequence transduction, relying solely on attention mechanisms, which they named Transformer. In 2021, inspired by Transformer' s tremendous success in NLP, the Google team developed a new image classification architecture called ViT, with its basic structure shown in [Figure 4: see original paper].

Input images in the ViT model are first divided into fixed-size patches, after which the flattened patches undergo linear mapping (dimension transformation through matrix multiplication). To preserve positional information for each patch, position encodings are added before feeding the patches into the Transformer encoder. The specific calculation formula is:

$$z_0 = [x_{\text{class}}; x_1^{pE}; x_2^{pE}; \ldots; x_N^{pE}] + E_{\text{pos}}$$

where $E \in \mathbb{R}^{(P^2 \cdot C) \times D}$, $E_{\text{pos}} \in \mathbb{R}^{(N+1) \times D}$. Here, z_0 represents the input to the entire Transformer Encoder, x_{class} is a [CLS] token added during serialization of input image $x \in \mathbb{R}^{H \times W \times C}$ (where H and W are image height and width, C is the number of channels) to accumulate and contain information from the entire sequence. x_i^p $(i = 1 \ldots N)$ is the i-th image patch, E is the linear projection layer, E_{pos} is the position encoding, P is the resolution of each patch, D is the patch dimension, and $N = HW/P^2$ is the sequence length of patches input to the Transformer encoder.

The Transformer encoder consists of L standard Transformer blocks, each comprising Layer Normalization (LN), Multi-head Self-Attention (MSA), a multi-layer perceptron (MLP), and residual connections (R-C). The computational process is as follows:

$$
\begin{aligned}
z_l' &= \text{MSA}(\text{LN}(z_{l-1})) + z_{l-1}, \quad l = 1 \ldots L \\
z_l &= \text{MLP}(\text{LN}(z_l')) + z_l', \quad l = 1 \ldots L \\
y &= \text{LN}(z_0^L)
\end{aligned}
$$

where z_l is the image patch sequence input to the Transformer Encoder, l is the iteration count, z_l' is the intermediate representation after applying the MSA module and residual connection. The previous layer' s output z_{l-1} undergoes layer normalization, then passes through the MSA module to obtain z_l'. Subsequently, z_l is the representation after applying the MLP module and the second residual connection. After obtaining z_l, it again goes through layer normalization, the MLP module, and residual connection to produce z_l. z_0^L is the [CLS] token at the first position in the final output sequence z_L after L Transformer Encoder cycles, and y is the final output of the Transformer Encoder.

It is important to note that ViT can only achieve performance comparable to state-of-the-art convolutional structures when pre-trained on large-scale datasets and then transferred to medium- or small-scale datasets.

3 Data and Preprocessing

3.1 Dataset Description

The dataset used in this study is Galaxy Zoo 2 (GZ2), created based on the large-scale volunteer classification work of the Galaxy Zoo project and adopting its data and classification standards. The Galaxy Zoo project, originating from the Galaxy Zoo-the Galaxy Challenge competition on the Kaggle platform, is a crowdsourced citizen science initiative aimed at studying and understanding galaxy morphology and evolution through volunteer classifications.

The competition training set contains 61,578 labeled galaxy observation images from SDSS Data Release 7. SDSS galaxy observation data includes five optical bands (u, g, r, i, and z), though research commonly uses the first three bands to synthesize corresponding RGB galaxy images. Each image measures

424$\times$424$\times$3*pixelsandhasa*1$\times$$37 label vector derived from corrected cumulative frequency values of volunteer voting scores. GZ2 divides galaxy morphology into 11 questions with 37 possible answers. Following related work, this paper selects five galaxy categories for classification research using the FPN-ViT model: In-between smooth galaxy, Completely round smooth galaxy, Edge-on galaxy, Spiral galaxy, and Cigar-shaped smooth galaxy.

3.2 Sample Selection

Selecting clean samples (well-sampled galaxies) from the GZ2 dataset requires following rules from the data release white paper:

1. **Volunteer requirement**: Each galaxy image must be classified by at least 20 volunteers to ensure sufficient classification opinions.
2. **Corrected cumulative vote score threshold**: The corrected cumulative vote score calculated for each image must meet a certain threshold based on comprehensive volunteer classification results.
3. **Threshold conditions**: To classify an image into a specific galaxy category, corresponding threshold conditions must be met. For spiral galaxies, an image must satisfy three threshold conditions: frequency of being classified as featured/disk ($f_{\text{features/disk}} > 0.430$), frequency of being classified as non-edge-on ($f_{\text{edge-on;no}} > 0.715$), and frequency of being classified as spiral ($f_{\text{spiral;yes}} > 0.619$).
4. **Special case for smooth galaxies**: Due to the strict clean sample selection rules in GZ2, relatively few samples are available for the three smooth galaxy categories (round, in-between, and cigar-shaped).

To ensure sufficient samples for model training and testing, this study moderately relaxed the threshold selection criteria for smooth galaxies from 0.8 to 0.5, while maintaining the default thresholds from the GZ2 white paper for edge-on and spiral galaxies.

Following these rules, the GZ2 dataset yielded a total of 28,790 clean sample galaxy images. [Figure 5: see original paper] shows examples of the five galaxy categories randomly selected from the clean sample dataset.

The clean samples were divided into training and test sets at a 9:1 ratio. [Figure 6: see original paper] shows the number of images in the training and test sets for each category, demonstrating that the five galaxy types satisfy the requirement of identical distribution and proportional representation.

Additionally, to balance sample quantities across categories, this study augmented the training data for the cigar-shaped galaxy category, which had fewer samples. Cigar-shaped galaxy images were rotated by 45°, 90°, 120°, and 180°, with examples shown in [Figure 7: see original paper], where r denotes the rotation angle.

4 Experiments and Results

4.1 Experimental Environment

Computations were performed on a server equipped with a V100-SXM2-32GB GPU and a 12-vCPU Intel(R) Xeon(R) Platinum 8255C CPU, using PyCharm Professional 2021.1 as the compiler and CUDA 11.3. The implementation was based on PyTorch 1.11.0 framework using Python, with additional libraries including sklearn, Scikit-image, and transforms.

4.2 Results Analysis

To validate the classification performance of the FPN-ViT model, this work employed the base FPN-ViT B/16 model and evaluated performance using accuracy, precision, recall, and F1-score. presents the best classification results for each galaxy category using FPN-ViT. Except for cigar-shaped galaxies, all categories achieve classification accuracy above 98%, with precision, recall, and F1-score all exceeding 97%. The cigar-shaped galaxy category shows lower performance, with all metrics below 90% but above 82%, primarily due to insufficient data quantity. On average across the five categories, the model achieves 95.2% accuracy, 95.2% precision, 95.0% recall, and 95.2% F1-score, confirming the robustness of FPN-ViT for galaxy morphology classification.

[Figure 8: see original paper] uses Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) to evaluate model performance. The results show good classification performance for each category, with AUC values above 0.98 for all categories except the cigar-shaped galaxies, which has an AUC of 0.975. The confusion matrix for FPN-ViT classification on the GZ2 dataset is shown in [Figure 9: see original paper], where 0–4 represent different galaxy morphologies. The results indicate high classification accuracy for in-between and round galaxies, while some errors occur for edge-on, spiral, and particularly cigar-shaped galaxies. Specifically, 7 cigar-shaped galaxies were misclassified as in-between, 12 as edge-on, and 3 as spiral, likely because the limited sample size prevented the model from adequately learning their morphological features.

FPN was proposed to address recognition difficulties caused by image size variations. To validate FPN' s effectiveness, we performed scaling operations on galaxy images from the GZ2 dataset by adjusting the scale parameter. Experiments used scale values of 0.35, 0.5, 0.65, and 0.8, with scaled images shown in [Figure 10: see original paper]. Classification results for these resized images using FPN-ViT are presented in . The experiments demonstrate average accuracies around 90% for galaxy images of different sizes and resolutions, confirming that FPN-ViT maintains good classification performance and validating the effectiveness of introducing FPN into the ViT architecture.

Furthermore, we added varying levels of Gaussian noise and salt-and-pepper noise to original galaxy images to verify FPN-ViT' s generalization capability for low signal-to-noise ratio images. Gaussian noise intensity was controlled

by adjusting the sigma value (standard deviation of the Gaussian distribution), with experimental values of 5, 15, and 25. Salt-and-pepper noise adds black and white pixels, controlled by the amount parameter representing the noise proportion, with experimental values of 0.05, 0.1, and 0.2. Larger sigma and amount values indicate more severe image degradation. Noisy galaxy images are shown in [Figure 11: see original paper], with classification results in . While overall classification performance decreases compared to clean images, FPN-ViT maintains accuracy above 70% under noise conditions, demonstrating stable performance and good generalization for low signal-to-noise ratio galaxy image classification.

In galaxy morphology classification research, brightness variations correlate closely with observation distance. Since galaxy brightness changes at different distances, using images with varying brightness can simulate observations at different distances and study the robustness of morphology classification algorithms against distance effects. Experiments adjusted brightness values to 0.5, 0.75, 1.5, and 2.0, with adjusted images shown in [Figure 12: see original paper]. Classification results for different brightness levels using FPN-ViT are presented in , confirming the model' s robustness for galaxy images with varying brightness.

In actual astronomical observations, observation distance relates to image size, brightness, and signal-to-noise ratio in several ways: (1) Image size generally decreases with increasing distance due to reduced angular size; (2) Brightness weakens with distance because of attenuation of radiative energy during propagation; (3) Signal-to-noise ratio decreases with distance as the received signal weakens while noise effects become more significant.

To accurately describe these relationships, we extended experiments by simulating real observation distances through a distance parameter. Greater distance values produce smaller galaxy images with lower brightness and signal-to-noise ratio. Experiments used distance values of 0.5, 0.75, 1.5, and 2.0, with resulting images shown in [Figure 13: see original paper]. For these simulated distances, we calculated image brightness and Peak Signal-to-Noise Ratio (PSNR), where higher PSNR indicates less noise impact. Classification results using FPN-ViT for these simulated observation distances are shown in , with accuracies above 75% for all distance simulations, demonstrating good robustness.

We compared FPN-ViT classification performance with the traditional Vision Transformer model using identical data sources and train-test split ratios. shows that both Transformer-based models achieve high accuracy for galaxy system classification, with FPN-ViT showing improvements across all metrics compared to the base ViT model, confirming its superior performance for galaxy morphology classification tasks.

4.3 Visualization of Classification Results

To explore morphological feature information from classification results, we visualized test set outcomes using t-SNE algorithm. t-SNE is a nonlinear dimensionality reduction algorithm for multi-dimensional data scaling that preserves local structure and obtains low-dimensional data with high similarity to the original high-dimensional data. By converting similarity between data points into probabilities—using Gaussian distributions for high-dimensional space and t-distributions for embedding space—t-SNE maps high-dimensional data to low-dimensional space for visualization.

[Figure 14: see original paper] visualizes FPN-ViT galaxy morphology classification results. The visualization shows clearly defined clusters for each galaxy type, indicating effective classification. Edge-on and cigar-shaped galaxies show minor boundary connections due to their smaller sample sizes and similar shapes, leading to some misclassification.

5 Summary and Outlook

As sky surveys deepen, astronomical observation coverage expands, and detection technology advances, the volume of astronomical data from survey projects continues to grow, rendering traditional data processing methods inadequate for large-scale data processing demands. Motivated by deep learning' s broad application in astronomical data and Transformer' s tremendous success in NLP, this paper applied the FPN-ViT model to galaxy morphology classification. Transformer-based classification models achieved high accuracy in galaxy morphology classification, with FPN-ViT attaining overall average accuracy of 95.2%, average precision of 95.2%, average recall of 95.0%, and average F1-score of 95.2%—representing improvements over CNN-based models and proving the applicability of Transformer-based models to galaxy morphology classification. Additionally, FPN-ViT maintained classification accuracy above 70% for low signal-to-noise ratio galaxy images, demonstrating good generalization capability. The t-SNE visualization of classification results provides intuitive insight into FPN-ViT' s effectiveness for galaxy morphology classification.

In the coming years, major telescope projects such as the China Space Station Telescope (CSST) and the Large Synoptic Survey Telescope (LSST) will launch, providing broader observational coverage and more detailed astronomical data. The FPN-ViT model adopted in this paper offers possibilities for subsequent data analysis, meaning it can be applied to broader astronomical datasets beyond the GZ2 dataset used here. Future work will continue exploring and researching the FPN-ViT model, applying it to classify galaxy images with morphologies not described in this paper, and focusing on investigating how network structure affects morphology classification performance to further validate the model' s effectiveness in galaxy morphology classification.

The user wants me to translate a "References" section from Chinese to English. The section contains a list of references, some of which have Chinese characters

in the author names and journal titles.

Let me analyze the specific issues: - Chinese characters found: 杜, 婷, 学, 徐, 艾, 等, 霖, 天, 峰, 权, 报, 文, 嫔, 利 - These appear to be: author names and journal titles

The translation rules are: - Author names: Romanize using pinyin (张三 → Zhang San) - Journal names: Use English title if available, otherwise translate the meaning - Institution names: Use official English names when known - Technical terms: Use standard English equivalents from the field - DO NOT leave any Chinese characters - not even one

Looking at the references, I need to identify which ones contain Chinese characters: - Reference [10] has Chinese characters: 艾霖嫔, 徐权峰, 杜利婷, 等. 天文学报, 2022, 63: 44

Let me romanize the names: - 艾霖嫔 → Ai Linpin (or Ai Lin-pin, but typically we don't hyphenate) - 徐权峰 → Xu Quanfeng - 杜利婷 → Du Liting

The journal name "天文学报" is "Acta Astronomica Sinica" in English.

So reference [10] should be: [10] Ai L P, Xu Q F, Du L T, et al. Acta Astronomica Sinica, 2022, 63: 44

Wait, let me check the standard format for Chinese names in academic references. Typically, we use the full pinyin for the surname and then initials for the given names. So: - 艾霖嫔 → Ai L P (or Ai L.) - 徐权峰 → Xu Q F (or Xu Q.) - 杜利婷 → Du L T (or Du L.)

But looking at the other references, they use formats like "Hubble E P", "Dressler A", "Blanton M R". So the pattern is: Surname, then space, then initials with spaces.

So: - 艾霖嫔 → Ai L P - 徐权峰 → Xu Q F - 杜利婷 → Du L T

The journal "天文学报" (Tianwen Xuebao) is known in English as "Acta Astronomica Sinica".

Let me double-check the other references to make sure I haven't missed any Chinese characters. Scanning through the list: - [1] through [9] are all English - [10] has Chinese characters - [11] through [27] are all English

So only reference [10] needs translation.

The original format is: [10] 艾霖嫔, 徐权峰, 杜利婷, 等. 天文学报, 2022, 63: 44

The romanized version should be: [10] Ai L P, Xu Q F, Du L T, et al. Acta Astronomica Sinica, 2022, 63: 44

Wait, I should check if "等" should be translated. In Chinese academic writing, "等" means "et al." (et alii), so it should be replaced with "et al."

The journal name "天文学报" - let me verify the standard English name. The Chinese Astronomical Society publishes "天文学报" which is translated as "Acta

Astronomica Sinica" . Sometimes it' s also referred to as "Chinese Journal of Astronomy and Astrophysics"but the official English name is "Acta Astronomica Sinica" .

Actually, looking at the reference style used in the other entries, they use abbreviated journal names like "ApJ" , "ARA&A" , "MNRAS" , "AJ" , "RAA" , "PASP" , "Ap&SS" . So I should check if "Acta Astronomica Sinica" has a standard abbreviation.

The standard abbreviation for Acta Astronomica Sinica is "Acta Astron. Sin." or sometimes "AAS" . But looking at the context, they seem to use full journal names or well-known abbreviations. Let me check the style more carefully:

- ApJ = The Astrophysical Journal
- ARA&A = Annual Review of Astronomy and Astrophysics
- MNRAS = Monthly Notices of the Royal Astronomical Society
- AJ = The Astronomical Journal
- RAA = Research in Astronomy and Astrophysics
- PASP = Publications of the Astronomical Society of the Pacific
- Ap&SS = Astrophysics and Space Science

So they use standard astronomical journal abbreviations. For "天文学报" , the standard abbreviation is "Acta Astron. Sin." or "ChJAA" (Chinese Journal of Astronomy and Astrophysics). However, the journal "天文学报" is actually the Chinese version, while the English version is "Chinese Journal of Astronomy and Astrophysics" (ChJAA).

Let me reconsider. The reference is: [10] 艾霖嫔, 徐权峰, 杜利婷, 等. 天文学报, 2022, 63: 44

If the article was published in the Chinese journal "天文学报" , then the English name is "Acta Astronomica Sinica" . But if it was published in the English language version, it might be "Chinese Journal of Astronomy and Astrophysics" .

Given that the title is in Chinese characters, it' s most likely the Chinese version. So "天文学报" should be "Acta Astronomica Sinica" .

However, I notice that reference [7] is "Cui X Q, Zhao Y H, Chu Y Q, et al. RAA, 2012, 12: 1197" where RAA stands for "Research in Astronomy and Astrophysics" , which is the English journal. And reference [9] is "Zhu X P, Dai J M, Bian C J, et al. Ap&SS, 2019, 364: 55" .

Given the context, I think the safest approach is to translate "天文学报" as "Acta Astronomica Sinica" and keep the format consistent with the other references.

So the final translation for reference [10] is: [10] Ai L P, Xu Q F, Du L T, et al. Acta Astronomica Sinica, 2022, 63: 44

Wait, I should double-check the name romanization. In Chinese academic contexts, sometimes the given names are written as: - 艾霖嫔 → Ai Linpin (full given

name) - 徐权峰 → Xu Quanfeng - 杜利婷 → Du Liting

But looking at the style of other references, they use initials. For example: - "Hubble E P" (Edwin Powell Hubble) - "Dressler A" (Alan Dressler) - "Blanton M R" (Michael R. Blanton) - "Lintott C J" (Chris J. Lintott)

So the format is: Surname, space, first initial, space, middle initial (if any).

For Chinese names, the surname is the first character(s) and the given name is the rest. So: - 艾霖嫔: Surname = 艾 (Ai), Given name = 霖嫔 (Linpin) → Ai L - 徐权峰: Surname = 徐 (Xu), Given name = 权峰 (Quanfeng) → Xu Q - 杜利婷: Surname = 杜 (Du), Given name = 利婷 (Liting) → Du L

But wait, the given name might be two characters, so we might need two initials. For example, if 霖嫔 is Lin Pin, then it would be L P. Similarly, 权峰 is Quan Feng → Q F, and 利婷 is Li Ting → L T.

So: - 艾霖嫔 → Ai L P - 徐权峰 → Xu Q F - 杜利婷 → Du L T

This matches the style of "Hubble E P" where E P are two initials.

Therefore, the final translation is: [10] Ai L P, Xu Q F, Du L T, et al. Acta Astronomica Sinica, 2022, 63: 44

Let me verify the journal name once more. "天文学报" (Tianwen X

Note: Figure translations are in progress. See original paper for figures.