GCSNet: A Galaxy Classification Algorithm Fusing Multi-Modal Features and Cosine Similarity

Siyi Zhang , Liangping Tu * , Jiawei Miao , Bing Su

Posted Date: 24 April 2026

doi: 10.20944/preprints202604.1739.v1

Keywords: galaxy classification; multimodal deep learning; emission lines galaxy; astronomy data analysis

 doi:10.20944/preprints202604.1739.v1

Article

GCSNet: A Galaxy Classification Algorithm Fusing Multi-Modal Features and Cosine Similarity

Siyi Zhang, Liangping Tu *, Jiawei Miao and Bing Su

University of Science and Technology Liaoning, Anshan 114051, China

* Correspondence: tuliangping@ustl.edu.cn

Abstract

Galaxy classification is essential for understanding the formation and evolution of cosmic structures. However, faced with the explosive growth of astronomical observation data, traditional single-modality classification methods relying solely on spectroscopy or imaging have struggled to meet high-precision demands due to insufficient feature utilization and limited generalization capability. Therefore, multimodal fusion has emerged as a promising direction by leveraging information complementarity to overcome the limitations of single data sources. Accordingly, this paper proposes a model named Galaxy CosineNet (GCSNet), which integrates imaging, spectroscopic, and tabular data for high-precision galaxy classification. Specifically, the model employs dedicated encoders to process the three modalities separately and utilizes skip connections to preserve raw features. Furthermore, it incorporates a multi-head self-attention m echanism t o d eeply mine global cross-modal complementary information. Finally, these features are concatenated and fed into a cosine similarity classification head. Experimental results demonstrate that GCSNet achieves 97.15% accuracy in classifying star-forming, composite, active galactic nuclei (AGNs), and normal galaxies. This performance outperforms the best single-modal baseline, GaSNet, by 0.76% and mainstream multi-modal models such as MB-ISTL and the Transformer by over 1.6%. Consequently, the proposed GCSNet offers an effective and novel approach for research on automatic galaxy classification.

Keywords: galaxy classification; multimodal deep learning; emission lines galaxy; astronomy data analysis

1. Introduction

Since Hubble [1] (1926) proposed the morphological sequence, the study of galaxy classification has undergone a century of development, transitioning from traditional visual inspection to modern deep learning. The current classification systems primarily rely on two complementary approaches: morphological classification, which categorizes galaxies (e.g., ellipticals, spirals, irregulars) based on geometric appearance; and physical classification, which utilizes spectral characteristics to distinguish physical states, such as star-forming galaxies, active galactic nuclei (AGNs), and their subclasses. These two systems are complementary, jointly form the cornerstone of modern galaxy research, advancing the exploration of distant galaxies.

In the realm of morphological classification, the field has shifted from traditional machine learning to deep learning. Early works, such as Odewahn et al. [2] (1992), pioneered the use of neural networks combined with photometric parameters to achieve automatic separation of stars from galaxies. With the advent of deep learning, convolutional neural networks (CNNs) became mainstream. For instance, Dai [3] (2018) applied an improved ResNet architecture to the Galaxy Zoo dataset for five galaxy morphologies (circular, intermediate, cigar-shaped, edge-on, and spiral), surpassing classical methods. Subsequently, Yang et al. [4] (2021) achieved 98.23% accuracy in a star-galaxy binary classification task. More recently, Kong et al. [5] (2024) further improved performance by introducing a channel attention mechanism for five-class galaxy classification, reaching an accuracy of 99.37%. Although the accuracy of image classification is steadily improving, methods that

 doi:10.20944/preprints202604.1739.v1

rely solely on image modalities often struggle to capture the physical processes within galaxies, leading to potential misclassification in cases of similar morphology but distinct physical properties.

Regarding spectral-based physical classification. Baldwin et al. [6] (1981) proposed the optical diagnostic map BPT for emission line galaxies. Connolly et al. [7] (1995) utilized PCA in early works to demonstrate the physical correlation between continuous spectral variations and morphological sequences. With the advancement of technology, machine learning methods have proliferated, evolving into deep learning approaches that leverage spectral physical information to enhance the accuracy of galaxy classification. Wu et al. [8] (2023) developed GalSpecNet, a 1D CNN model using emission line features to classify star-forming, composite, AGN, and normal galaxies. It achieved over 93% accuracy, with visualizations showing consistency with traditional BPT diagnostic lines. Later, Zhong et al. [9] (2024) further proposed the GasNet-II architecture. It reached 96% accuracy in star/galaxy/quasar classification and performed simultaneous redshift prediction. Although spectral data is rich in physical information, it is limited by the coverage range of optical fibers and the scarcity of labeled samples, making it difficult to meet the requirements of all-sky classification of large-scale surveys when used alone.

Given the inherent limitations of single-modal data, multi-modal fusion has become an inevitable trend to break through the classification bottleneck. There is a significant complementarity between the morphological structure provided by images and the physical parameters contained in the spectrum. Wei et al. [10] (2023) proposed that the BATMM model incorporate the Transformer self-attention mechanism, combining image and spectral data to identify blue horizontal branch stars. The results were higher than the accuracy of any single modality, effectively validating the potential of multi-modal fusion. Deng et al. [11] (2024) proposed the multimodal ensemble model MESCR, achieving 96.1% accuracy in the five-class classification of A, F, G, K, and M type stars, and constructed a predictive catalog containing over 50 million stars. These studies fully demonstrate that effectively integrating heterogeneous data can significantly enhance the generalization ability and physical interpretability of models. In the same year, Rizhko & Bloom [12] constructed a multimodal dataset comprising time-series photometry, spectra, and metadata. By leveraging the complementarity of these three modalities to classify ten types of variable stars, their model achieved an accuracy of 94.07%, offering a new perspective on multi-source astronomical data fusion. Junell et al. [13] (2025) leveraged a hybrid architecture of encoders and CNNs on four modalities (photometry, image cutouts, metadata, and spectra) to realize a unified framework for the automatic classification of transients and variables.

Although considerable progress has been made in multimodal galaxy classification, significant challenges remain in the fine-grained classification of four types of galaxies with distinct physical mechanisms: star-forming galaxies, composite galaxies, active galactic nuclei (AGNs), and normal galaxies. These four types of galaxies differ fundamentally in their physical nature: star-forming galaxies [14] are dominated by young massive stars and exhibit distinct spiral arms and bluer ultraviolet radiation; composite galaxies [15] feature both star formation and weak AGN activity, with spectral characteristics in a transitional state; AGNs [16] are powered by central black hole accretion and have extremely compact nuclear regions; and normal galaxies [17] lack prominent ionization sources, presenting reddened, smooth morphological structures and weak spectral emission lines. Together, they form a complete galaxy evolution sequence ranging from intense star formation activity to quiescent evolutionary states. Affected by dust extinction, redshift effects, spatial resolution, and other factors, these four types of galaxies exhibit strong degeneracy in the optical color space, making effective differentiation difficult using traditional photometric methods alone [18]. In spectral physical analysis, the emission-line ratios of composite galaxies and low signal-to-noise ratio AGNs often fall into ambiguous boundary regions in BPT diagnostic diagrams, and they are easily confused due to overlapping ionization mechanisms [19]. Early classification studies mostly relied on empirical boundaries such as BPT diagrams [20] or focused on the separation of broad celestial categories (e.g., stars vs. galaxies). For example, Cao et al. [21] (2024) proposed the FPN-ViT hybrid model for morphological classification of five galaxy types using photometric

 doi:10.20944/preprints202604.1739.v1

images. The model achieved evaluation metrics exceeding 95%, demonstrating excellent classification performance. In the same year, Moradi et al. [21] constructed the FNet II model using spectral modalities to achieve spectral classification of broad celestial objects such as stars, galaxies, and quasars. Subsequently, Deng et al. [11] fused multimodal features including spectral and photometric data and improved the classification accuracy of stellar subclasses through ensemble learning. However, the fine-grained classification mechanism for such complex physical scenarios still requires further exploration. Therefore, breaking through the limitations of single-modal data and realizing high-precision automatic classification of these four galaxy types is of great astronomical significance for revealing the laws of galaxy evolution.

Traditional single-modal galaxy classification has information limitations when characterizing complex physical mechanisms. Multimodal fusion also suffers from original feature loss and insufficient cross-modal interaction. To address these issues, this paper proposes the GCSNet model. This study conducts innovative research from three perspectives: introducing a new modality, improving feature interaction, and designing a new decision head. The main contributions are as follows:

(1) We innovatively incorporate tabular data containing redshift (z), median signal-to-noise ratio (snmedian), interstellar extinction parameter E(B-V) and the $H\alpha$ equivalent width ($EW\_H\alpha\_6562$) into the classification architecture as an independent third modality. This method supplements global physical information, thereby enhancing the model's discriminative ability regarding complex ionization mechanisms such as star formation and AGN activity.

(2) We design a skip connection structure to preserve original features, while employing a multi-head self-attention mechanism to mine deep cross-modal correlations. By concatenating these original and fused features, we achieve efficient fusion and comprehensive representation of image morphology, spectra, and tabular data.

(3) We propose a classification decision head based on cosine similarity [23]. By leveraging its focus on feature direction rather than magnitude, this approach reduces the impact of sample count discrepancies on classification boundaries, effectively mitigating the influence of class imbalance. Experiments on the Sloan Digital Sky Survey Data Release 17 (SDSS DR17) dataset demonstrate that the model achieves an overall accuracy of 97.15%. Furthermore, in the classification of composite galaxies with similar physical morphologies, it outperforms the GasNet model by 5.18%, exhibiting superior robustness.

The structure of this paper is organized as follows: Section 2 describes the dataset, data preprocessing techniques; Section 3 elaborates on the network architecture of GCSNet ; Section 4 presents and discusses the experimental results and ablation analysis; finally, Section 5 summarizes the main conclusions of the paper.

2. Data and Preprocessing

2.1. Dataset Construction

The data in this paper are derived from SDSS DR17 [24]. Through the CasJobs interface, we integrated image morphological, spectral physical, and tabular statistical features to construct a multimodal dataset containing 68,917 high-quality galaxies. The sample consists of 33,291 star-forming galaxies, 13,131 composite galaxies, 7,662 AGNs, and 14,833 normal galaxies. Given that Seyfert galaxies and low-ionization nuclear emission-line region galaxies (LINERs) exhibit similar spectral characteristics and are relatively scarce (5,636 and 2,026, respectively), we merged them into a single AGNs category to optimize class distribution balance.

To ensure the physical reliability and classification validity of the data, sample selection follows the following strict criteria:

First, we enforced strict criteria for spectral integrity and quality. The redshift range is limited to $0<z<0.3$. This range ensures that key BPT diagnostic lines such as $H\alpha 6563$ Å, [NII]6584 Å, [OIII]5007 Å, $H\beta 4861$ Å fall fully within the SDSS spectral coverage area (3800Å-9200Å). Moreover,

 doi:10.20944/preprints202604.1739.v1

we required a median signal-to-noise ratio (S/N) greater than 10. This threshold ensures the reliable detection of weak emission lines across all galaxy types, particularly for composite galaxies with ambiguous boundaries and quiescent normal galaxies, thereby guaranteeing that the classification labels are based on robust observational data.

Second, we constructed a multi-modal dataset. Images in the u, g, and r bands were selected due to their sensitivity to distinct physical properties of galaxies: the u band reveals the AGNs characteristics, the g band tracks the star-forming regions, and the r band reflects the mass distribution of stars [25]. Each galaxy is imaged under three bands, synthesizing a three-channel RGB composite image, and forming a multi-modal data pair with the corresponding spectrum. In addition, the interstellar extinction parameter E(B-V) and the $H\alpha$ equivalent width ($EW\_H\alpha\_6562$) are included. The former facilitates extinction correction and tests of the unified model, while the latter aids in distinguishing host galaxy types[26].

Regarding the selection of galaxy subclasses: SFGs, Composite galaxies, and AGNs (including Seyferts and LINERs) are easily confused due to overlapping ionization mechanisms and thus require clear differentiation; while normal galaxies, though "quiescent", are an indispensable endpoint in the galaxy evolution sequence. These two reasons determine the classification objects. The entire dataset is randomly divided into a training set, a validation set, and a test set with a ratio of 70%, 15%, and 15%, respectively, and the detailed sample distribution is shown in Table 1. The basic data of this paper can be obtained from SDSS, with the URL: https://dr17.sdss.org/ (catalog data).

Table 1. The distribution of galaxy subclasses in the dataset.

Class	Train Set	Valid Set	Test Set	Total
Star Forming	23,303	4,994	4,994	33,291
Composite	9,191	1,970	1,970	13,131
AGN	5,364	1,149	1,149	7,662
Normal	10,383	2,225	2,225	14,833
Total	48,241	10,338	10,338	68,917

2.2. Data Preprocessing

2.2.1. Image Data Preprocessing

The photometric images used in this paper were sourced from the SDSS Archive Server (SAS) and stored in FITS format [27]with three independent bands of u, g, and r. Due to the field of view differences among different bands, direct synthesis can lead to spatial misalignment and distortions in physical information. To address this, we employed the standardized RGB Synthesis program developed by He et al. [28] for multi-band registration. Specifically, the alignment pipeline was implemented using the Python reproject package, with the r -band images serving as the reference coordinate system. The u - and g -band images were rigorously aligned to this reference frame via geometric transformations. To ensure pixel-level spatial consistency, a center crop was applied based on the minimum common dimensions across all bands. Following alignment, the three-band data were stacked into (3, H, W) tensors and uniformly resampled to a spatial resolution of 300×300 pixels using bilinear interpolation, balancing computational efficiency with detail preservation. Finally, to eliminate differences in the distribution and numerical range of gray values among images of different bands, Z-score normalization is performed independently on each channel of each image. The normalization is defined as shown in Equation (1):

$$x' = \frac{x - \mu}{\sigma} \tag{1}$$

 doi:10.20944/preprints202604.1739.v1

Where x represents the original pixel value, μ and σ denote the mean and standard deviation of the specific band, respectively. Through this preprocessing pipeline, we constructed multi-band image inputs that are spatially aligned, uniform in resolution, and statistically normalized, ready for subsequent model training.

2.2.2. Spectral Data Preprocessing

The SDSS DR17 original spectrum covers approximately 3800Å–9200Å, with a total of 3843 sampling points, mainly consisting of continuous spectra, absorption/emission lines, and noise [29]. As previous studies have pointed out, considering that the SDSS spectra have not undergone absolute flux calibration, the absolute intensity of the continuous spectra is unreliable, while the relative intensity ratios of emission lines (such as $[OIII]/H\beta$, $[NII]/H\alpha$) are the key basis for physical classification [30]. Directly inputting original fluxes not only introduces dimensional differences but also is vulnerable to baseline drift and noise interference. Therefore, we have constructed two spectral input strategies to explore the impact of different feature representations on classification performance:

(1) Observed Spectra and Robust Normalization

First, we extracted the traffic data from the COADD extension table of the FITS file, replacing non-numeric values (NaN) with zeros and discarding invalid samples with a standard deviation below 1×10^{-10}. To accommodate the input requirements of the convolutional neural network (CNN), all spectra were uniformly truncated or zero-padded to a fixed length of L=3843. Addressing common spectral outliers such as cosmic ray spikes and sky line residuals, we eschewed traditional Z-score standardization due to its sensitivity to extreme values. Instead, we adopted Per-sample Robust Scaling. This method computes the median and interquartile range ($IQR_{(x)}$=Q3-Q1) along the wavelength dimension (L=3843) for each sample and normalizes the data according to Equation (2):

$$x_{norm} = \frac{x - median(x)}{IQR(x)} \quad (2)$$

(2) Redshift Correction and Rest-Frame Spectra

To mitigate the effects of cosmological redshift on spectral line positions, we constructed a second dataset comprising rest-frame spectra. Utilizing the redshift values provided by SDSS, we converted the original observed wavelengths λ_{obs} and fluxes f read from the FITS file to rest-frame wavelengths according to Equation (3):

$$\lambda_{rest} = \frac{\lambda_{obs}}{1+z} \quad (3)$$

Subsequently, the fluxes were resampled onto a uniform rest-frame wavelength grid spanning [3800, 7000)Å with a step size of 2Å, yielding a total of 1600 sampling points. Samples failing to fully cover this interval (e.g., due to gaps at the blue or red ends) were excluded to prevent non-physical artifacts arising from boundary truncation. Following this, Mean Normalization was applied to mitigate the impact of varying overall brightness among galaxies, thereby enabling the network to focus on the relative morphological features of spectral lines. This process scales each spectrum by its mean flux, as defined in Equation (4):

$$f_{norm}(\lambda) = \frac{f(\lambda)}{mean(f)} \quad (4)$$

Here, $f(\lambda)$ represents the resampled flux at wavelength λ, and $mean(f)$ is the arithmetic mean of the flux values over the N=1600 grid points. This normalization ensures that differences in absolute brightness do not dominate the learning process, allowing the model to focus on spectral shape features such as line ratios and continuum slopes.

2.2.3. Tabular Data Preprocessing

Tabular data is typically organized in rows and columns, where each row represents an independent sample and each column corresponds to a physical property or metadata. Taking

 doi:10.20944/preprints202604.1739.v1

galaxies as an example, this corresponds to astronomical catalogs containing diverse information such as object identifiers, coordinates (e.g., Right Ascension and Declination), magnitudes, redshifts, effective temperatures, and morphology. Observational data derived from different surveys is uniformly stored in these catalogs, which facilitates the construction of multi-modal architectures [31]. To effectively integrate these four-dimensional physical parameters into the multi-modal model framework, this study also employs Robust Scaler for standardization. This method scales features using statistics that are robust to outliers, as defined in Equation (5):

$$x_{norm} = \frac{x - \text{median}(x)}{\text{IQR}(x)} \quad (5)$$

Where IQR(x) denotes the interquartile range (Q3-Q1). This approach is particularly suitable for astronomical data, which often contains extreme values arising from rare objects or measurement anomalies. Crucially, to prevent data leakage, the scaling parameters (median and IQR) are fitted exclusively on the training set. The validation and test sets are then standardized using these same training-derived statistics, thereby ensuring the reliability and unbiased nature of the experimental evaluation.

3. Methods

GCSNet is an end-to-end deep neural network designed for multi-modal feature fusion. As illustrated in Figure. 1, the architecture primarily consists of three modality-specific encoders, a self-attention fusion module, and a dual-task classification head. The model extracts features from each modality via independent encoders and employs a self-attention mechanism to generate fused features containing global interaction information. Building on this, a skip connection mechanism is employed to preserve original features, which are then concatenated with fused features to enhance the integrity of the feature representation. The concatenated features are then L2-normalized onto a hypersphere space, where the primary classification is performed based on cosine similarity. Furthermore, to improve the model's discriminability towards Active Galactic Nuclei (AGN), an auxiliary binary classification head is introduced to distinguish between "AGN" and "non-AGN" sources, leveraging a multi-task learning strategy to further optimize the feature distribution.

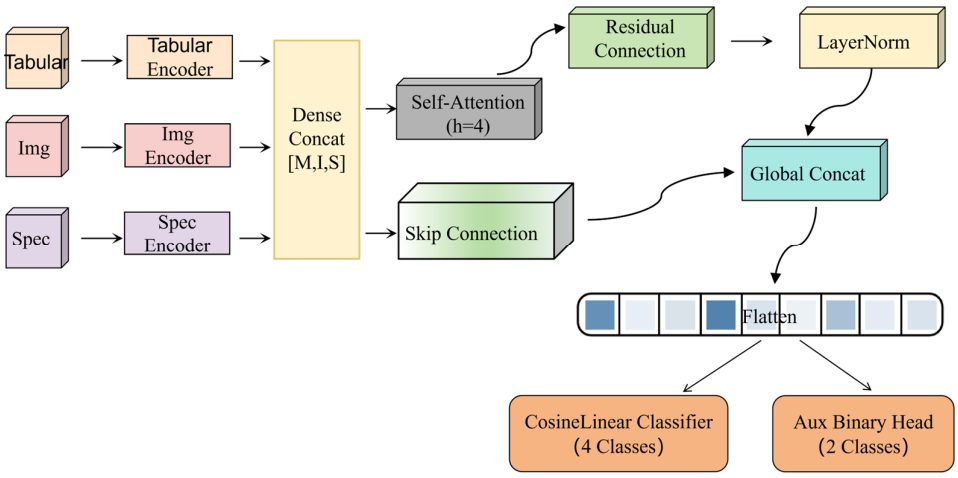

Figure 1. Overview of the multi-modal model architecture.

3.1. Image Encoder

Designed for efficiency and robust feature extraction from two-dimensional photometric images, this module employs a lightweight 2D-CNN architecture comprising a stack of three convolutional modules. To ensure stable training dynamics and enhance non-linear representation capability, each convolutional layer is immediately followed by Batch Normalization (Batch Norm) and a ReLU

 doi:10.20944/preprints202604.1739.v1

activation function. Spatial dimensionality is reduced via a two-stage downsampling strategy, which halves the feature map resolution (height and width) at each stage. Concurrently, the channel depth is progressively expanded from 32 to 128; this design choice broadens the receptive field while capturing multi-level semantic features. Finally, following the third convolutional layer, a 4×4 adaptive average pooling layer (AdaptiveAvgPool2d) is applied to transform spatial features into 16 grid tokens (4×4). These tokens are then linearly projected into a 128-dimensional feature space to obtain the image feature sequence $F_{img} \in R^{B\times16\times128}$, as illustrated in Figure 2.

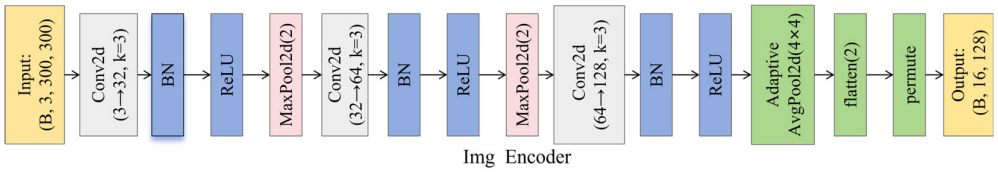

Figure 2. Schematic diagram of the Image Encoder.

3.2. Spectral Encoder

Designed specifically for one-dimensional spectral sequences, this module employs a three-layer 1D-CNN to capture local emission lines and continuous spectral characteristics. To effectively model features at multiple scales, the convolutional kernel sizes are set to 11, 7, and 5, with corresponding strides of 2, 2, and 1. The channel depth progressively increases from 1 to 128, enabling the extraction of hierarchical spectral representations. Each convolutional layer is followed by Batch Normalization and a ReLU activation function to stabilize training and enhance non-linearity. Finally, to handle variable-length input spectra, an AdaptiveAvgPool1d layer is applied to compress the features into a fixed sequence of 16 tokens, forming the final spectral feature sequence $F_{spec} \in R^{B\times16\times128}$, as illustrated in Figure 3.

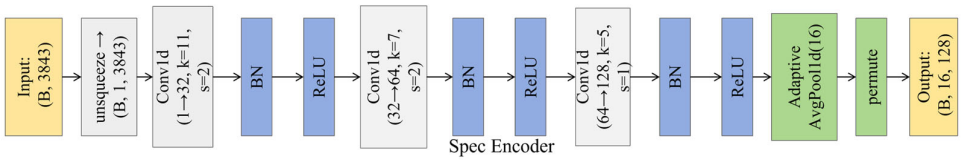

Figure 3. Schematic diagram of the Spectral Encoder.

3.3. Tabular Encoder

Designed for low-dimensional dense metadata vectors, this module employs a Multi-Layer Perceptron (MLP) to project sparse inputs into a high-dimensional latent space, effectively reshaping them into pseudo-sequences compatible with other modalities. The original metadata is first mapped to a 64-dimensional hidden space via two successive linear layers and then expanded to a total dimensionality of 2,048 (16 × 128). Following Layer Normalization and ReLU activation, this feature vector is reshaped into a sequence format of (B, 16, 128), yielding the final metadata feature sequence $F_{meta} \in R^{B\times16\times128}$, as illustrated in Figure 4.

 doi:10.20944/preprints202604.1739.v1

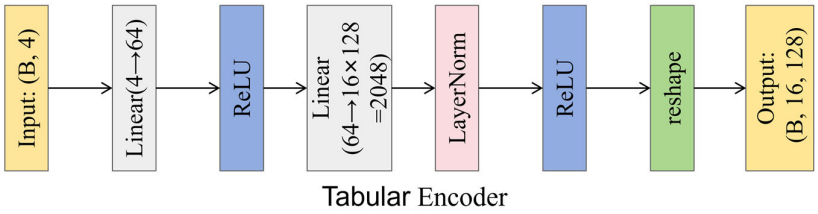

Figure 4. Schematic diagram of the Tabular Encoder.

Although the three encoders adopt different architectures to adapt to the data characteristics of their respective modalities, they all standardize their outputs into a sequence representation of (B, 16, 128) via pooling or reshaping operations, ensuring the dimensional consistency of the subsequent fusion modules. Here, B represents the batch size, 16 denotes the sequence length (number of tokens), and 128 signifies the feature dimension.

3.4. Multi-Modal Self-Attention Fusion Module

Information competition between modalities often causes simple feature concatenation to fail, where dominant modalities tend to mask the features of weaker ones [1632]. Inspired by Dense Fusion [33], this study proposes a four-way self-attention fusion module based on the concept of dense feature-level fusion via pixel-wise correspondence. Unlike traditional global fusion strategies, Dense Fusion has demonstrated that fine-grained cross-modal interaction (e.g., at the pixel or token level) significantly enhances the model's ability to capture local features and improves robustness. Specifically, the module first concatenates the output features from the three encoders along the sequence dimension, obtaining the total joint feature that contains F_{spec}, F_{img}, F_{meta} all the modal information from the three encoders. Subsequently, a 4-head multi-head self-attention mechanism is applied, enabling each token to attend to global dependencies across all other modalities. Following residual connection and Layer Normalization, the features are compressed back to 16 tokens via adaptive average pooling to obtain the global fused features $F_{fused} \in R^{B\times16\times128}$. Finally, the fused features F_{fused} are concatenated with the original output features from the three encoders to form the final complementary feature representation $F_{final} = [F_{spec}, F_{img}, F_{meta}, F_{fused}] \in R^{B\times64\times128}$. This design not only retains the original discriminative information of each mode (preventing information loss during fusion), but also introduces interactively enhanced fused features, significantly improving the expressive power of the features.

3.5. Cosine Similarity Classification Head

The classification decision utilizes a cosine similarity classifier head instead of the traditional Softmax classifier to effectively address class imbalance scenarios [34]. Specifically, the fused features are first flattened into a vector of dimension 2048×4=8192. Both the input feature vector x and the weight vectors w are then subjected to L2 normalization, projecting them onto a unit hypersphere. The classification score is derived in two steps: first, the cosine similarity between the normalized vectors is calculated (Eq.6); second, this value is multiplied by a learnable scaling factor σ (initialized to 30.0) to yield the final logits (Eq.7),where x denotes the input feature vector and w represents the weight vector.

$$\cos(\mathrm{x},\mathrm{w}) = \frac{\mathrm{x}\cdot\mathrm{w}}{\|\mathrm{x}\|\|\mathrm{w}\|} \tag{6}$$

$$\mathrm{logits} = \sigma\cdot\cos(\mathrm{x},\mathrm{w}) \tag{7}$$

This mechanism ensures that classification decisions rely solely on the direction of the feature vectors, effectively eliminating interference from amplitude factors (such as brightness) and enhancing the model's ability to distinguish subtle inter-class differences. Regarding data

 doi:10.20944/preprints202604.1739.v1

augmentation, to address the long-tail distribution [35] inherent in the observed galaxy data, a weighted random sampler was integrated into the training data loader. The sampling weight for each sample is determined by the reciprocal of its class frequency,where $n_{c(i)}$ denotes the total number of samples in the category c(i) to which the i - th sample belongs.

$$w_i = \frac{1}{n_{c(i)}} \tag{8}$$

Additionally, the Mixup [36] data augmentation technique is employed to generate new training samples by blending two images and their corresponding labels in specific proportions. Specifically, two samples, (X1, y1) and (X2, y2), are randomly selected, where X = (image, spectrum, tabular data), y denotes the label. To maintain modality consistency, the same mixing coefficient λ is applied across the image, spectral, and metadata inputs when generating the new samples $\widetilde{X}$. The formulation is presented in Eq. (9).

$$\widetilde{X} = \lambda X_1 + (1-\lambda)X_2, \widetilde{y} = \lambda y_1 + (1-\lambda)y_2 \tag{9}$$

4. Results and Discussion

To comprehensively evaluate the effectiveness and robustness of the proposed multimodal feature fusion framework, GCSNet, this section presents extensive performance validation experiments and ablation studies. Through a rigorous experimental design covering diverse metrics and scenarios, we demonstrate the superiority of our multi-modal fusion approach over single-modal baselines. Furthermore, the results highlight the significant advantages of our proposed architecture in terms of both classification accuracy and robustness.and robustness.

4.1. Evaluation Indicators

All experiments were conducted under identical hardware and software environments to ensure the comparability and reproducibility of the results. Specifically, the following core evaluation metrics were employed to verify the model's robustness，as defined in Eqs.(10)-(14):

(1) Accuracy: Overall classification accuracy

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{10}$$

(2) Precision: The accuracy of predicting each category as positive

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{11}$$

(3) Recall: The detection rate of actual positive cases in each category

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{12}$$

(4) F1-score: The harmonic average of precision and recall

$$\text{F1} - \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{13}$$

(5) Macro-F1

$$\text{Macro} - \text{F1} = \frac{1}{\text{N}} \sum_{\text{i}=1}^{\text{N}} \text{F1}_{\text{i}} \tag{14}$$

(6) AUC value: The area under the ROC curve, which measures the overall discriminative ability of the model. The value ranges from 0 to 1, with values closer to 1 indicating better performance.

To eliminate the influence of randomness and ensure a fair comparison, we conducted repeated experiments across multiple independent random seeds. All models were trained under identical configurations (AdamW optimizer, learning rate of 0.001, batch size of 64, and label smoothing of 0.1). The final reported results represent the average performance across all runs for each metric.

 doi:10.20944/preprints202604.1739.v1

4.2. Experimental Setup

4.2.1. Experimental Environment

All experiments were conducted under identical hardware and software configurations to ensure comparability and reproducibility. The server runs on the CentOS 7.9 operating system, equipped with 256GB of memory and 4 NVIDIA A100 (80GB) GPUs. The model training is implemented based on the Python 3.8 environment using the PyTorch 2.0.1 framework.

4.2.2. Hyperparameter Settings

We employed the AdamW optimizer for parameter updates. The models were trained from scratch for 120 epochs with a batch size of 64. The initial learning rate was set to 0.001, with a weight decay of 0.001 and a label smoothing coefficient of 0.1.

4.2.3. Baseline Models

To systematically verify the effectiveness of the proposed method, the experimental design is divided into two categories: single-modal baselines and multi-modal fusion groups.

The single-modal baseline group aims to reveal that although single-modal classification methods are mature, they are limited by the partial nature of single data sources and struggle to achieve comprehensive characterization of multi-dimensional galaxies. We selected GaSNet [37] as a highly representative pure spectral baseline model. Proposed by Zhong et al. (2023), this model is based on a multi-channel ensemble system that utilizes multiple small ResNet sub-networks with independent initial weights to process 1D spectral data, demonstrating superior performance in spectral classification tasks. In addition, we introduced the spectral branch of the transfer learning model (ISTL) [38] proposed by Wang and Hong et al. (2023) to compare and evaluate the performance differences of different spectral feature extractors. Notably, to verify the effectiveness of the spectral branch proposed in this paper and to provide a reference for the subsequent improvement in multi-modal classification capability, the GCSNet spectral branch is also included in this baseline group.

The multi-modal fusion group focuses on exploring the pros and cons of different fusion mechanisms to verify the superiority of the GCSNet fusion strategy. This group introduces MB-ISTL [38] as a representative traditional dual-modal model. This method independently extracts spectral and image features and then fuses them at the decision level via concatenation and weighting, representing a classic late fusion paradigm. Meanwhile, we referenced the TransformerFusion [39] model proposed by Gao et al. (2023). This model utilizes the self-attention mechanism of a multi-modal Transformer to achieve cross-modal interaction at the embedding sequence level, representing a state-of-the-art fusion method. On this basis, we established two core comparisons: first, the proposed end-to-end dual-modal architecture (Spectral-Image Dual-Modal), which serves as a baseline for data alignment in multi-modal tasks; and second, the final proposed tri-modal fusion model, GCSNet, which incorporates tabular data features to better supplement multi-dimensional galaxy information and further break through existing performance bottlenecks.

4.3. Comparative Experiment

4.3.1. Multi-class Galaxy Classification Performance Evaluation

Galaxy classification faces the dual challenges of high-dimensional data and feature sparsity, making it difficult to capture the full physical picture using a single feature source. Effective fusion architectures can mitigate the limitations of single-modality data by integrating heterogeneous multi-source information, thereby significantly enhancing classification accuracy. Accordingly, we conducted a systematic comparative study to evaluate the fine-grained classification performance of various models across four galaxy types: AGNs, Star-Forming, Composite, and Normal. Table 2 summarizes the quantitative results in terms of Precision, Recall, and F1-score.

 doi:10.20944/preprints202604.1739.v1

Table 2. Performance Evaluation for Four-Class Galaxy Classification.

Model	Class	Precision	Recall	F1-score
GaSNet	AGN	0.947461	0.941688	0.944566
	STAR FORMING	0.983333	0.968763	0.975994
	Composite	0.891654	0.927411	0.909181
	Normal	0.995959	0.996854	0.996406
	Average	0.954602	0.958679	0.956537
MB-ISTL	AGN	0.905281	0.939948	0.922289
	STAR FORMING	0.978256	0.936924	0.957144
	Composite	0.823117	0.892893	0.856586
	Normal	0.996854	0.996854	0.996854
	Average	0.925877	0.941655	0.933218
Transformer	AGN	0.947141	0.937627	0.942265
	STAR FORMING	0.966346	0.967227	0.966741
	Composite	0.882856	0.884264	0.883247
	Normal	0.997606	0.998502	0.998054
	Average	0.948487	0.946905	0.947577
GCSNet	AGN	0.938031	0.961706	0.949721
	STAR FORMING	0.977205	0.987185	0.98217
	Composite	0.943416	0.905584	0.924113
	Normal	1	0.999551	0.999775
	Average	0.964663	0.963507	0.963945

The results indicate that model classification performance is highly correlated with the distinctiveness of galactic physical features and sample abundance: categories with unique features and sufficient samples yield superior classification results, whereas high spectral feature overlap significantly increases classification difficulty.

Normal Galaxies: All models performed exceptionally well on this category, attributable to its distinct physical characteristics. Normal galaxies exhibit weak or absent emission lines, resulting in clear boundaries from other categories in the BPT diagnostic diagram. Notably, GCSNet, leveraging its multi-modal fusion strategy, further eliminated misclassifications, achieving a precision of 100%.

Star-Forming Galaxies: Benefiting from abundant samples and prominent spectral features, all models demonstrated strong learning capabilities in this category. The best performance was observed in GCSNet and the single-spectrum model GaSNet. Although GCSNet's accuracy was marginally lower than GaSNet's (by 0.6%), it achieved a 2% higher recall and a 1% higher F1-score, indicating superior comprehensive performance.

Composite and AGN Galaxies: Classification performance for these categories was relatively poor, primarily due to their prominent strong emission line features. Particularly for composite galaxies, which reside in the transition region between star formation and AGN on the BPT diagram, the ambiguous spectral boundaries make them highly prone to confusion. The proposed GCSNet effectively mitigates this issue, achieving a classification accuracy of 94.34%, which outperforms the second-best model by 5.18%.

4.3.2. Overall Model Performance Evaluation

Based on the detailed analysis of individual galaxy types, this section presents a comprehensive comparison of the overall performance of all models, with a focus on evaluating the impact of different fusion strategies on classification accuracy. The results are summarized in Table 3.

 doi:10.20944/preprints202604.1739.v1

Table 3. Comparison of experimental results across different models.

Model	Accuracy	Precision	Recall	F1-score
GaSNet	0.963920	0.954602	0.958679	0.956537
ISTL（Spectral）	0.935191	0.914952	0.941538	0.927011
our-Spectral	0.961405	0.956844	0.962084	0.959457
MB-ISTL	0.941768	0.925877	0.941655	0.933218
TransformerFusion	0.954859	0.948487	0.946905	0.947577
our-Spec_Img	0.960239	0.955738	0.955744	0.955741
GCSNet	0.971464	0.964663	0.963507	0.963945

Results analysis and experimental data demonstrate two key findings: (1) Single-modality classification methods are highly mature, with GaSNet achieving a high accuracy of 96.39% using only spectral features; (2) Existing multi-modality methods perform even worse than the single-modality baseline due to improper fusion strategies, such as shallow concatenation or modality inconsistency. In contrast, GCSNet, by designing a reasonable and effective fusion strategy and leveraging multi-dimensional feature data of galaxies, achieves optimal performance across all metrics in terms of overall performance.

The AUC curves in Figure. 5 show that GCSNet lies closest to the top-left corner, with values nearest to 1. This indicates better class discrimination than the compared models, confirming the model's robustness and generalization in classification tasks.

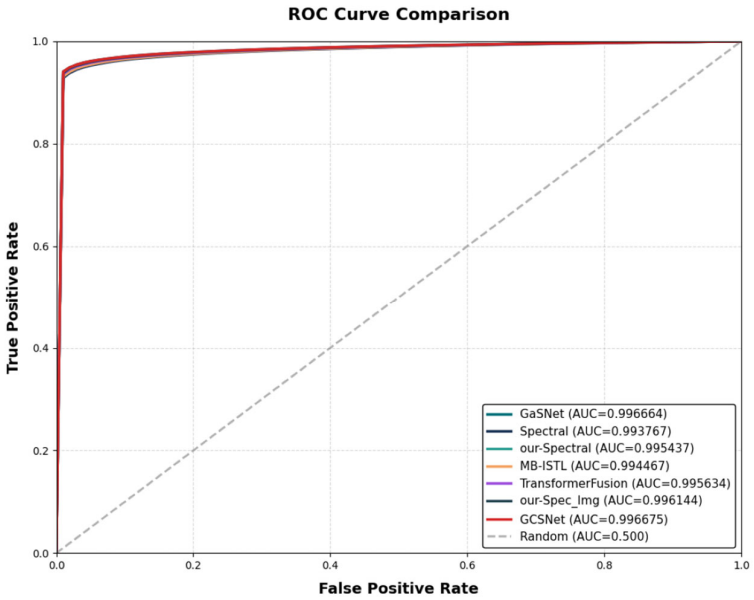

Figure 5. ROC curves of all models.

4.3.3. Model Stability Evaluation

Based on the aforementioned experiments, it is evident that GCSNet demonstrates superior overall performance compared to existing comparative models. To further investigate the model's fine-grained discriminative capability across various categories and its stability, this study conducted three independent repeated experiments under different random seeds. Table 4 (four-class

 doi:10.20944/preprints202604.1739.v1

classification metrics) and Figure 6 (confusion matrix) provide a detailed presentation of its classification performance.

Table 4. Evaluation performance of the GCSNet model for four-class galaxy classification.

Class	Precision	Recall	F1-score
AGN	0.9409±0.0067	0.9603±0.0105	0.9505±0.0061
STAR FORMING	0.9833±0.0027	0.9792±0.0042	0.9813±0.0009
Composite	0.9250±0.0156	0.9232±0.0102	0.9240±0.0053
Normal	1.0000±0.0003	1.0000±0.0000	0.9999±0.0001

Note: All metrics are reported as the mean ± standard deviation of three independent experiments (random seeds: 42, 100, 3407).

As shown in Table 4, GCSNet demonstrates excellent performance in the four-class galaxy classification task. The standard deviations ranging from 0.006 to 0.016 confirm the model's high stability and strong robustness.

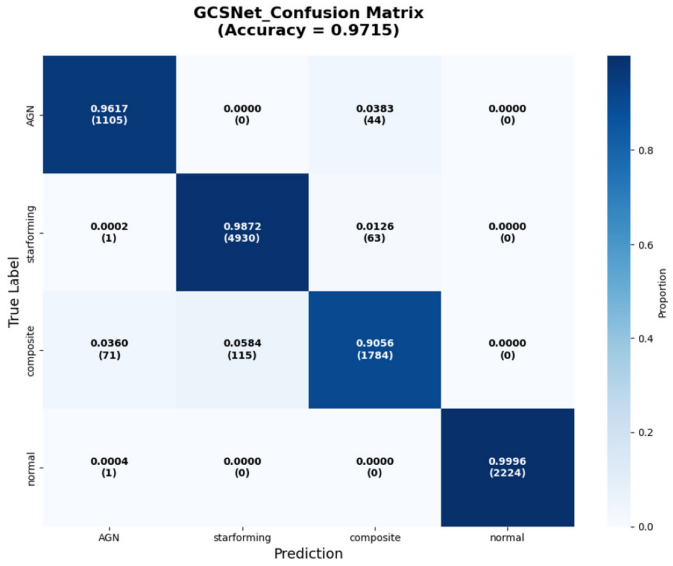

Figure 6. GCSNet confusion matrix.

The clear diagonal distribution observed in the confusion matrix (Figure 6) confirms the model's effectiveness in distinguishing morphologically similar galaxies, with minimal inter-class confusion. The sparse misclassifications in the off-diagonal regions are primarily concentrated on boundary samples with ambiguous features (e.g., between AGNs and composite galaxies). This pattern indicates that the errors stem from the inherent ambiguity of the samples themselves rather than any systematic bias in the model. Overall, the combined chart data further demonstrates the model's superior performance in the galaxy classification task.

4.4. Ablation Experiment

To systematically evaluate the contribution of each module to the final performance of the model, we designed multiple sets of ablation experiments. We observed their influence on the best validation

 doi:10.20944/preprints202604.1739.v1

accuracy and the final test accuracy in two scenarios: obs (original observed spectra) and rest (rest-frame spectra).

(1) Observation Scene（Obs）

As shown in Table 5, the full metadata input, including redshift information, is retained.

Table 4. Verification experiments for observing the original spectrum.

frame	ablation	best_val_acc	final_test_acc
obs	all	**0.972142**	**0.971851**
obs	all_noz	0.965467	0.969046
obs	spec_img	0.962082	0.963242
obs	spec_tabular_ewonly	0.967982	0.967402
obs	spec_tabular_noew	0.966435	0.967692
obs	spec_tabular_noz	0.967015	0.970014
obs	spec_only	0.961985	0.963726

Note: Bold entries in the table highlight the best results in the obs scenario, demonstrating the superiority of full modality integration.

(2) Redshift removal scene (Rest)

Table 6 illustrates an extreme scenario simulating low signal-to-noise ratio (SNR) spectra where redshift cannot be measured, or cross-domain transfer to datasets without redshift labels. This setup aims to examine whether the model overly relies on redshift while neglecting image and spectral features.

Table 5. Verification Experiments of redshift spectra.

frame	ablation	best_val_acc	final_test_acc
rest	all	0.951435	0.948430
rest	all_noz	0.952598	0.950465
rest	spec_img	0.947945	0.946976
rest	spec_tabular_ewonly	0.950368	0.947169
rest	spec_tabular_noew	0.948914	0.944940
rest	spec_tabular_noz	0.950271	0.950465
rest	spec_only	0.948236	0.947460

As can be seen from the comparison of Table 5 and Table 6, the full model using the original observed data from all three sources (image + spectrum + tabular data) achieved the best performance of 0.9721, indicating significant synergy among the modules. In contrast, the overall classification performance decreased after redshift removal. This confirms that redshift, as the strongest physical prior in galaxy classification, can effectively constrain the high-dimensional feature space and help the model distinguish galaxy types with similar morphology but different evolutionary stages.

5. Conclusions

This paper proposes GCSNet, a multi-modal deep learning framework designed to overcome the limitations of single-source data in characterizing complex galaxy structures. By deeply fusing image morphology, spectral physics, and tabular statistical features, the model achieves an accuracy of 97.15%, successfully surpassing the performance ceiling of unimodal approaches and demonstrating significant superiority over existing multimodal models.

 doi:10.20944/preprints202604.1739.v1

This study further clarifies the impact of galactic physical features and sample abundance on classification performance. It is found that the distinctiveness of galactic spectral features is a key factor determining classification difficulty. Normal galaxies, characterized by unique spectral features (weak or absent emission lines) and clear boundaries in the BPT diagnostic diagram, along with sufficient samples, perform excellently across various models, with GCSNet achieving 100% precision. In contrast, composite galaxies and AGNs, due to their strong emission line features and transitional positions in the BPT diagram, exhibit blurred spectral boundaries, leading to significantly increased classification difficulty. Notably, the model proposed in this study exhibits significant advantages in the composite galaxy classification task, achieving an accuracy of 94.34%. Compared with the GasNet model, which is the second-best performing method in this task, the performance is improved by 5.18%.

In conclusion, GCSNet exhibits robust classification capabilities, validates the immense potential of multimodal learning in astronomical data mining, and provides a high-precision solution for automated galaxy classification.

Author Contributions: Methodology, S.Z., L.T., and .J.M; data curation, S.Z. writing—original draft preparation, S.Z.; writing—review and editing, S.Z., L.T., J.M., and B.S.; supervision, S.Z., L.T., J.M., and B.S.; project administration, L.T.; funding acquisition, L.T. All authors have read and agreed to the published version of the manuscript

Funding: This work was supported by the National Natural Science Foundation of China (NSFC, grant No. U1731128).

Data Availability Statement: The data supporting this study consist of a specific subset derived from the publicly available Sloan Digital Sky Survey 17th Data Release (SDSS DR17), obtained through rigorous selection and matching processes. The detailed criteria for this data curation are described in Section 2 (Data and Preprocessing), and the corresponding code is available in the Supplementary Materials.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Pang W. Hubble galaxy classification method[J]. Scientific American, 2021(14): 2.
2. Odewahn S C, Stockwell E B, Pennington R L, et al. Automated star/galaxy discrimination with neural networks[J]. Astronomical Journal, 1992, 103(1): 318-331.
3. Dai J M. Research on galaxy morphological classification based on deep convolutional neural networks[D]. Beijing: University of Chinese Academy of Sciences (National Space Science Center, Chinese Academy of Sciences), 2018.
4. Yang Y, Wen Z L, Xia J Q. Star-galaxy classifier based on residual neural networks[J]. Journal of Beijing Normal University (Natural Science), 2021, 57(4): 450-457.
5. Kong X Y, Dou J P. Galaxy morphology classification model based on SE-Inception-v3[J]. Acta Astronomica Sinica, 2024, 65(2): 30-43.
6. Baldwin J A, Terlevich M M P. Classification parameters for the emission-line spectra of extragalactic objects[J]. Publications of the Astronomical Society of the Pacific, 1981, 93(551): 5-19.
7. Connolly A J, Szalay A S, Bershady M A, et al. Spectral classification of galaxies: an orthogonal approach[J]. The Astronomical Journal, 1994, 110(3): 1071.
8. Wu Y, Tao Y, Fan D, et al. Galaxy spectral classification and feature analysis based on convolutional neural network[J]. Monthly Notices of the Royal Astronomical Society, 2023, 527(1): 1163-1176.
9. Fucheng Z, Napolitano N R, Caroline H, et al. Galaxy Spectra neural Network (GaSNet). II. Using deep learning for spectral classification and redshift predictions[J]. Monthly Notices of the Royal Astronomical Society, 2024(1): 1.
10. Jiaqi W, Bin J, Yanxia Z. Identification of Blue Horizontal Branch Stars with Multimodal Fusion[J]. Publications of the Astronomical Society of the Pacific, 2023, 135(1050): 1-15.

 doi:10.20944/preprints202604.1739.v1

11. Deng Z J, Yu S Y, Luo A, et al. Ensemble Learning for Stellar Classification and Radius Estimation from Multimodal Data[J]. Research in Astronomy and Astrophysics, 2024, 24(11): 211-224.
12. Rizhko M, Bloom J S. AstroM3: A self-supervised multimodal model for astronomy[J]. 2024.
13. Junell A, Sasli A, Nunes F F, et al. Applying multimodal learning to Classify transient Detections Early (AppleCiDEr) I: Data set, methods, and infrastructure[J]. 2025.
14. The physical properties of star-forming galaxies in the low-redshift Universe[J]. Monthly Notices of the Royal Astronomical Society, 2010, 351(4): 1151-1179.
15. Kauffmann, Guinevere, Heckman, et al. Stellar masses and star formation histories for 10[sup5] galaxies from the Sloan Digital Sky Survey.[J]. Monthly Notices of the Royal Astronomical Society, 2003.
16. Wang B Z. Discussion on characteristics, classification and unified model of active galactic nuclei[J]. Journal of Jiaozuo Teachers College, 2025, 41(2): 74-76.
17. Baldry I K, Glazebrook K, Brinkmann J, et al. Quantifying the Bimodal Color-Magnitude Distribution of Galaxies[J]. ApJ, 2004, 600.
18. Kauffmann, Guinevere, Heckman, et al. Stellar masses and star formation histories for 10[sup5] galaxies from the Sloan Digital Sky Survey[J]. Monthly Notices of the Royal Astronomical Society, 2003.
19. Chamorro-Cazorla M, De Paz A G, Castillo-Morales A, et al. MEGADES: MEGARA galaxy disc evolution survey. Ionised gas diagnosis[J]. 2025.
20. Kewley L J, Groves B, Kauffmann G, et al. The Host Galaxies and Classification of Active Galactic Nuclei[J]. 2006.
21. Cao J, Xu T T, Deng Y H, et al. Research on galaxy morphological classification based on FPN-ViT[J]. Acta Astronomica Sinica, 2024, 65(3): 120-133.
22. Moradi R, Rastegarnia F, Wang Y, et al. FNet II: spectral classification of quasars, galaxies, stars, and broad absorption line (BAL) quasars[J]. Monthly Notices of the Royal Astronomical Society, 2024(2): 2.
23. Salton G. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613-620.
24. The Seventeenth Data Release of the Sloan Digital Sky Surveys: Complete Release of MaNGA, MaStar and APOGEE-2 Data[J]. 2021.
25. Strateva I, Ivezic Z, Knapp G R, et al. Color Separation of Galaxy Types in the Sloan Digital Sky Survey Imaging Data[J]. The Astronomical Journal, 2001, 122(4): 1861-1874.
26. Cardelli J A, Clayton G C, Mathis J S. The relationship between infrared, optical, and ultraviolet extinction[J]. Astrophysical Journal, 1989, 345.
27. Liu M Y, Zhao Z X, Wang W, et al. Analysis and information extraction of LAMOST stellar spectra FITS files[J]. Computer Programming Skills & Maintenance, 2019(8): 5.
28. Zhendong H, Bo Q, A-Li L, et al. Deep Learning Applications Based on SDSS Photometric Data: Detection and Classification of Sources[J]. Monthly Notices of the Royal Astronomical Society, 2021.
29. THE MULTI-OBJECT, FIBER-FED SPECTROGRAPHS FOR THE SLOAN DIGITAL SKY SURVEY AND THE BARYON OSCILLATION SPECTROSCOPIC SURVEY[J]. The Astronomical Journal, 2013, 146(2).
30. L, J, Kewley, et al. Theoretical Modeling of Starburst Galaxies[J]. The Astrophysical Journal, 2001, 556(1): 121-121.
31. Collaboration T M U, Audenaert J, Bowles M, et al. The Multimodal Universe: Enabling Large-Scale Machine Learning with 100TB of Astronomical Scientific Data[J]. 2024.
32. Baltrusaitis T, Ahuja C, Morency L P. Multimodal Machine Learning: A Survey and Taxonomy[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, PP(99): 1-1.
33. Wang C, Xu D, Zhu Y, et al. DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion[J]. IEEE, 2020.
34. Liu W, Wen Y, Yu Z, et al. SphereFace: Deep Hypersphere Embedding for Face Recognition[J]. IEEE, 2017.
35. Cui Y, Jia M, Lin T Y, et al. Class-Balanced Loss Based on Effective Number of Samples[J]. arXiv, 2019.
36. Zhang H, Cisse M, Dauphin Y N, et al. mixup: Beyond Empirical Risk Minimization[J]. 2017.
37. Fucheng Z, Napolitano N R, Caroline H, et al. Galaxy Spectra neural Network (GaSNet). II. Using deep learning for spectral classification and redshift predictions[J]. Monthly Notices of the Royal Astronomical Society, 2024(1): 1.

 doi:10.20944/preprints202604.1739.v1

38. Bingjun W, Shuxin H, Zhiyang Y, et al. A Multimodal Transfer Learning Method for Classifying Images of Celestial Point Sources[J]. Publications of the Astronomical Society of the Pacific, 2023, 135(1052): 1-12.
39. Jialin G, Jianyu C, Jiaqi W, et al. Deep Multimodal Networks for M-type Star Classification with Paired Spectrum and Photometric Image[J]. Publications of the Astronomical Society of the Pacific, 2023, 135(1046): 1-9.