MargFormer: Photometric Classification of Stars, Quasars and Compact Galaxies with Cross-Attention Vision Transformer

Anonymous CVPR submission

Paper ID *****

Abstract

We present MargFormer, a unified deep-learning model designed to classify stars, quasars, and compact galaxies by synergistically integrating photometric parameters and imaging data. Leveraging a cross-attention vision transformer where photometric features serve as queries to probe imaging data, MargFormer processes both heterogeneous data types within a single cohesive framework. This unified architecture contrasts with conventional approaches, where prior methods process each data modality of photometric parameters and images using corresponding architectures such as ANNs and CNNs and then stack their outputs. By enabling photometry to guide image feature extraction within this joint framework, MargFormer effectively captures intricate local/global features and cross-modal correlations. This results in a substantially lightweight model with fewer parameters and demonstrates improved generalization performance. Evaluated on data from the Sloan Digital Sky Survey (SDSS) Data Release (DR) 16, MargFormer achieves performance comparable to or exceeding state-of-the-art methods, even at fainter magnitudes. This work underscores the power and efficiency of transformer-based models, highlighting their potential as scalable solutions with strong generalization capabilities, crucial for analyzing the vast and diverse datasets from upcoming wide-field surveys like the Vera C. Rubin Observatory. Our trained models and code will made publicly available upon acceptance.

1. Introduction

Accurate classification of celestial objects is a fundamental task in observational astronomy, essential for numerous downstream scientific applications. These include precise cosmological parameter estimation, studies of galaxy evolution, mapping Galactic structure, and identifying rare or transient phenomena [17]. Reliable separation of object classes (e.g., stars, quasars, galaxies) within large astronomical datasets is therefore critical for maximizing the scientific yield of surveys, particularly those targeting sensitive cosmology and large-scale structure analyses. Modern astronomical surveys, such as the Sloan Digital Sky Survey (SDSS; [20]), the Dark Energy Survey (DES [8]), and the Zwicky Transient Facility (ZTF [4]), generate catalogues containing hundreds of millions to billions of detected sources. The forthcoming Vera C. Rubin Observatory's Legacy Survey of Space and Time (LSST [13]) will significantly increase data volumes, collecting terabytes nightly. The sheer scale of these datasets renders manual classification infeasible and necessitates the development of efficient, robust, and accurate automated methods for source classification.

Machine learning (ML), and particularly its subset deep learning (DL), has become increasingly central to astronomical data analysis over the past two decades [3]. DL models, characterized by deep neural network architectures, excel at learning complex, hierarchical representations directly from data. Their application has yielded significant results in diverse astronomical problems, such as stellar spectral classification [16], determining galaxy morphology [9], and estimating photometric redshifts [15]. A key area of impact is automated source classification. Initial ML approaches, including Random Forests [18] and Support Vector Machines [11], often relied on curated photometric features and provided valuable baselines but exhibited limitations, particularly for faint objects or ambiguous cases. Convolutional Neural Networks (CNNs) subsequently offered substantial improvements for image-based classification tasks [14], leveraging their inherent ability to effectively capture spatial hierarchies and local patterns.

Despite progress, the accurate differentiation between stars, quasars, and especially compact galaxies remains challenging, particularly at faint magnitudes (e.g., $r > 22.5$) where morphological differences diminish and signal-to-noise ratios are lower [6, 14]. While highly accurate, spectroscopic classification is observationally expensive and impractical for most survey sources. Photometric colours provide discriminatory power, especially for star-quasar separation [1], but can be degenerate or affected by redshift. Morphological classifiers struggle when galaxies appear nearly point-like. Combining information from photometric parameters and imaging data (i.e., multimodal learning) is a promising strategy to overcome these limitations. The MargNet model [7] implemented this concept using a hybrid architecture: an Artificial Neural Network (ANN) processed photometric features, while a CNN processed images, with their outputs subsequently concatenated (stacked) for final classification. This approach demonstrated improved performance on SDSS DR16 data, notably for faint, compact sources.

However, conventional hybrid architectures like MargNet process distinct data modalities through separate, parallel streams, deferring fusion until later stages. This architectural separation inherently constrains the model's ability to capture complex, low-level inter-dependencies between photometric properties and spatial image features during the critical initial feature extraction phases. Recent advancements in deep learning, particularly the emergence of Transformer architectures [19] built upon attention mechanisms and their successful application to vision (Vision Transformers, ViTs [10]), present a compelling alternative paradigm. With their capacity for modelling global context via self-attention, transformers offer more flexible and potentially deeper mechanisms for integrating heterogeneous data streams. While the application of ViTs to this multimodal problem has been explored (e.g., MM ViT [5]), existing strategies have limitations; for instance, MM ViT primarily utilized photometric features only to inform the final classification (CLS) token, rather than leveraging them to actively guide the spatial feature extraction process within the ViT itself. This leaves untapped potential for more synergistic fusion.

Motivated by the potential for multimodal integration, we propose MargFormer, a novel deep-learning architecture designed for the unified classification of stars, quasars, and compact galaxies. MargFormer integrates photometric parameters and imaging data within a cohesive framework using a cross-attention Vision Transformer. It employs photometric features as queries within the cross-attention mechanism to dynamically guide extracting and integrating relevant information from image representations. This unified processing methodology allows for the direct learning of cross-modal interactions, differing fundamentally from the separate-stream, late-fusion paradigm of prior hybrid models. Key advantages of this approach include a significantly more parameter-efficient architecture and demonstrably improved generalization performance, which is critical for reliable application across the diverse and large-scale datasets expected from future surveys like LSST.

The main contributions of this paper are:

- We introduce MargFormer, a unified transformer architecture that jointly processes heterogeneous astronomical data (photometry and images) by employing photometric features as queries within a cross-attention mechanism to guide image feature extraction for source classification effectively.
- We evaluate MargFormer using SDSS DR16 data, performing a direct comparison with the baselines MargNet [7], MM ViT [5], and show that it achieves comparable or superior performance, particularly for challenging faint and compact objects.

This paper is organized as follows. Section 2 describes the SDSS DR16 dataset used. Section 3 details the data preprocessing steps, the MargFormer model architecture, and comparative methods. Section 4 presents the experimental results and discussion. Finally, Section 5 provides conclusions and discusses potential future work.

2. Dataset and Experimental Setup

The data utilized in this study are sourced from the Sloan Digital Sky Survey Data Release 16 (SDSS DR16) [2]. We use a set of 24 derived photometric parameters, associated $ugriz$ FITS images [12], and ground-truth classifications (star, quasar, galaxy) obtained from the official SDSS spectroscopic pipeline. Crucially, to ensure rigorous and direct performance comparison with previous work, we adopt the exact dataset construction methodology, data partitioning strategy, and experimental framework established in [7] (hereafter C23). This involves employing the two primary datasets defined and curated in C23 to target challenging populations: the “Compact source dataset” and the “Faint and Compact source dataset.” The specific magnitude and compactness selection criteria defining these datasets are detailed comprehensively in C23.

Following C23, these datasets were partitioned into training, validation, and test subsets, maintaining identical class distributions (star, quasar, compact galaxy) across splits. We replicate the three experimental scenarios defined in C23 to evaluate performance under different conditions, particularly focusing on generalization: **Experiment 1:** Training, validation, and testing performed solely on the Compact source dataset. **Experiment 2:** Training, validation, and testing performed solely on the Faint and Compact source dataset. **Experiment 3:** Training and validation performed using the Compact source dataset, with testing conducted on the Faint and Compact source dataset to evaluate generalization to fainter, more challenging objects explicitly. By strictly adhering to the data selection, preparation, and experimental procedures detailed in C23, we establish a direct baseline for comparing the performance of the MargFormer model presented herein against the results reported for MargNet [7] and MM ViT [5]. For exhaustive details regarding data retrieval, selection cuts, processing, and splits, we refer the reader to C23.

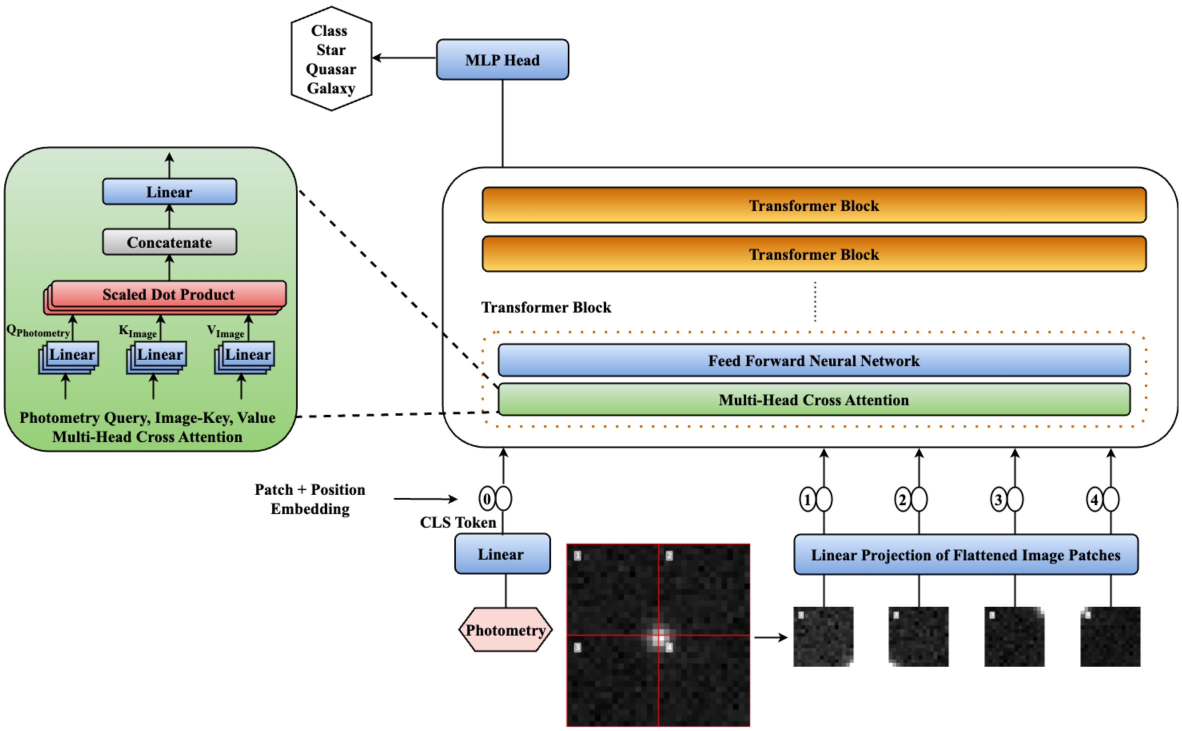

Figure 1. Architecture of MargFormer, illustrating cross-attention fusion using photometric features as queries to probe imaging data.

3. Methodology

We introduce MargFormer (Fig. 1), a novel deep learning architecture for unified astronomical source classification (stars, quasars, galaxies) that intrinsically fuses photometric and imaging data. Unlike conventional hybrid models (e.g., MargNet [7]) with separate streams and late fusion limiting cross-modal learning, MargFormer leverages a cross-attention mechanism [10, 19]. Photometric embeddings serve as queries to probe image patch representations (keys and values), enabling photometric context to guide visual feature extraction throughout the network. This design facilitates deeper, synergistic fusion aimed at capturing complex inter-dependencies.

3.1. Input Processing

The input processing stage prepares the two distinct data modalities for the cross-attention mechanism. Multi-band (ugriz) FITS images are partitioned into a sequence of non-overlapping patches, each linearly projected into an embedding vector. Learnable positional embeddings are added to these patch embeddings to preserve spatial context. These processed image embeddings form the basis for the Keys (K_I) and Values (V_I) in the subsequent cross-attention layers. Concurrently, the corresponding vector of 24 derived photometric parameters (as used in C23 [7]) is linearly projected into a compatible embedding space. These photometric embeddings are designated to serve as the crucial Queries (Q_P), enabling them to probe the image representations within the transformer blocks.

3.2. Cross-Attention Transformer Blocks

$$\text{Attention}(Q_P, K_I, V_I) = \text{softmax}\left(\frac{Q_P K_I^T}{\sqrt{d_k}}\right) V_I \quad (1)$$

The core of MargFormer consists of stacked transformer blocks utilizing cross-attention to explicitly model interactions between modalities, diverging from standard ViT self-attention. Within each block, attention scores are computed via scaled dot-product attention (Eq. 1). d is the key dimension ensuring stable gradients [19]. This mechanism allows the photometric queries (Q_P) to selectively weight and integrate the most relevant visual information from the image values (V_I). Stacking these cross-attention layers enables the learning of progressively complex, deeply integrated cross-modal representations, ensuring photometric context actively guides visual feature extraction throughout the network depth. The output representation corresponding to a dedicated CLS token from the final block is then processed by a Multi-Layer Perceptron (MLP) head, which maps the learned features to the final class probabilities (star, quasar, or compact galaxy).

Experiment	Model	Accuracy	Precision	Recall
Ex1 - SG	MargNet	98.1 ± 0.1	98.1 ± 0.1	98.1 ± 0.1
	MM ViT	98.1 ± 0.1	98.1 ± 0.1	98.1 ± 0.1
	MargFormer	98.1 ± 0.1	98.1 ± 0.1	98.1 ± 0.1
Ex1 - SGQ	MargNet	93.3 ± 0.2	93.3 ± 0.2	93.3 ± 0.2
	MM ViT	93.2 ± 0.2	93.2 ± 0.2	93.2 ± 0.2
	MargFormer	93.1 ± 0.2	93.1 ± 0.2	93.1 ± 0.2
Ex2 - SG	MargNet	96.9 ± 0.1	96.9 ± 0.1	96.9 ± 0.1
	MM ViT	96.9 ± 0.1	96.9 ± 0.1	96.9 ± 0.1
	MargFormer	97.1 ± 0.1	97.1 ± 0.1	97.1 ± 0.1
Ex2 - SGQ	MargNet	86.7 ± 0.2	86.8 ± 0.2	86.7 ± 0.2
	MM ViT	86.3 ± 0.2	86.2 ± 0.2	86.3 ± 0.2
	MargFormer	86.6 ± 0.2	86.7 ± 0.2	86.6 ± 0.2
Ex3 - SG	MargNet	92.0 ± 0.1	92.7 ± 0.1	92.0 ± 0.1
	MM ViT	91.8 ± 0.1	92.5 ± 0.1	91.8 ± 0.1
	MargFormer	92.7 ± 0.1	93.2 ± 0.1	92.7 ± 0.1
Ex3 - SGQ	MargNet	73.4 ± 0.2	76.5 ± 0.2	73.4 ± 0.2
	MM ViT	71.8 ± 0.2	75.4 ± 0.2	71.8 ± 0.2
	MargFormer	75.2 ± 0.2	77.8 ± 0.2	75.2 ± 0.2

Table 1. Comparative performance evaluation for MargFormer, MargNet [7], and MM ViT [5] across the three experimental setups.

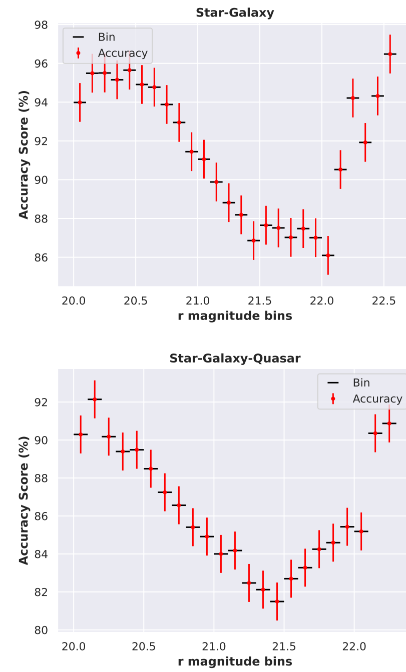

Table 2. Classification accuracy achieved by MargFormer plotted against r-band magnitude.

4. Results and Discussion

We demonstrate the performance of MargFormer for astronomical source classification using the SDSS DR16 dataset. The evaluation follows the three experimental scenarios defined in Section 2, which replicate the methodology of Chaini et al. (2023)[7] to enable direct and rigorous comparison with baseline models MargNet[7] and MM ViT [5]. We assess performance under binary (Star-Galaxy (SG)) and ternary (Star-Galaxy-quasar (SGQ)) classification settings, using overall accuracy, precision, and recall as primary metrics.

Table 1 summarizes the comparative performance. In Experiment 1 (Compact dataset), all models achieved high and comparable accuracy for both SG (98.1%) and SGQ (93.1-93.3%) tasks. Experiment 2 (Faint/Compact dataset) showed reduced performance overall, with MargFormer achieving competitive results (SG: 97.1%, SGQ: 86.6%), performing comparably or slightly better than MargNet and MM ViT. The advantage of MargFormer is most evident in Experiment 3, the generalization test (Train Compact, Test Faint/Compact). Here, MargFormer significantly outperformed both baselines, achieving higher accuracy in the SG setting (92.7% vs. 92.0% (MargNet) / 91.8% (MM ViT)) and particularly in the challenging SGQ setting (75.2% vs. 73.4% (MargNet) / 71.8% (MM ViT)). These results highlight MargFormer's superior generalization capability compared to the baseline architectures when faced with distribution shifts towards fainter objects. Table 2 details MargFormer's classification accuracy as a function of r band magnitude. As anticipated, classification for both binary (star-galaxy) and ternary (star-galaxy-quasar) classification generally declines with increasing magnitude (fainter sources) up to $r = 21.5$. This is consistent with decreasing signal-to-noise ratios in the photometric data. However, for magnitudes $r > 21.5$, we observe a counter-intuitive increase in accuracy. This unexpected trend reversal at faint magnitudes is not unique to MargFormer; it is consistently observed across all baselines. Notably, MargFormer demonstrates this behavior while being significantly more parameter-efficient than MargNet (using only 5.8% of its parameters) and comparable in size to MM ViT.

5. Conclusion and Future Work

We introduced MargFormer, a unified cross-attention model efficiently fusing photometric and imaging data. It achieves strong generalization, outperforming baselines on challenging faint objects while using significantly fewer parameters (5.8% vs MargNet). This demonstrates the power of unified multimodal transformers for large surveys. Future work will extend MargFormer to other key tasks leveraging both data types, such as photometric redshift estimation, galaxy parameter estimation, and strong lens identification.

References

[1] Sheelu Abraham, Ninan Sajeeth Philip, Ajit Kembhavi, Yogesh G Wadadekar, and Rita Sinha. A photometric catalogue of quasars and other point sources in the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 419(1):80–94, 2012. 2

[2] Romina Ahumada, Carlos Allende Prieto, Andrés Almeida, Friedrich Anders, Scott F Anderson, Brett H Andrews, Borja Anguiano, Riccardo Arcodia, Eric Armengaud, Marie Aubert, et al. The 16th data release of the sloan digital sky surveys: first release from the apogee-2 southern survey and full release of eboss spectra. *The Astrophysical Journal Supplement Series*, 249(1):3, 2020. 2

[3] Dalya Baron. Machine learning in astronomy: A practical overview. *arXiv preprint arXiv:1904.07248*, 2019. 1

[4] Eric Bellm. The zwicky transient facility. In *The Third Hot-wiring the Transient Universe Workshop*, 2014. 1

[5] Srinadh Reddy Bhavanam, Sumohana S Channappayya, Srijith P. K, and Shantanu Desai. Enhanced astronomical source classification with integration of attention mechanisms and vision transformers. *Astrophysics and Space Science*, 369 (8):92, 2024. 2, 4

[6] Laura Cabayol, Ignacio Sevilla-Noarbe, Enrique Fernández, Jorge Carretero, Martin Eriksen, Santiago Serrano, Alex Alarcon, Adam Amara, Ricard Casas, Francisco Javier Castander, et al. The pau survey: star–galaxy classification with multi narrow-band data. *Monthly Notices of the Royal Astronomical Society*, 483(1):529–539, 2019. 1

[7] Siddharth Chaini, Atharva Bagul, Anish Deshpande, Rishi Gondkar, Kaushal Sharma, M Vivek, and Ajit Kembhavi. Photometric identification of compact galaxies, stars, and quasars using multiple neural networks. *Monthly Notices of the Royal Astronomical Society*, 518(2):3123–3136, 2023. 2, 3, 4

[8] University of Chicago Lawrence Berkeley National Laboratory Cerro-Tololo Inter-American Observatory Dark Energy Survey Collaboration: Fermilab, University of Illinois at Urbana-Champaign and Brenna Flaugher. The dark energy survey. *International Journal of Modern Physics A*, 20 (14):3121–3123, 2005. 1

[9] Sander Dieleman, Kyle W Willett, and Joni Dambre. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly notices of the royal astronomical society*, 450(2):1441–1459, 2015. 1

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3

[11] Ross Fadely, David W Hogg, and Beth Willman. Star–galaxy classification in multi-band optical imaging. *The Astrophysical Journal*, 760(1):15, 2012. 1

[12] M Fukugita, K Shimasaku, T Ichikawa, JE Gunn, et al. The sloan digital sky survey photometric system. Technical report, SCAN-9601313, 1996. 2

[13] Željko Ivezić, Steven M Kahn, J Anthony Tyson, Bob Abel, Emily Acosta, Robyn Allsman, David Alonso, Yusra AlSayyad, Scott F Anderson, John Andrew, et al. Lsst: from science drivers to reference design and anticipated data products. *The Astrophysical Journal*, 873(2):111, 2019. 1

[14] Edward J Kim and Robert J Brunner. Star-galaxy classification using deep convolutional neural networks. *Monthly Notices of the Royal Astronomical Society*, page stw2672, 2016. 1

[15] Johanna Pasquet, Emmanuel Bertin, Marie Treyer, Stéphane Arnouts, and Dominique Fouchez. Photometric redshifts from sdss images using a convolutional neural network. *Astronomy & Astrophysics*, 621:A26, 2019. 1

[16] Kaushal Sharma, Harinder P Singh, Ranjan Gupta, Ajit Kembhavi, Kaustubh Vaghmare, Jianrong Shi, Yongheng Zhao, Jiannan Zhang, and Yue Wu. Stellar spectral interpolation using machine learning. *Monthly Notices of the Royal Astronomical Society*, 496(4):5002–5016, 2020. 1

[17] Maayane Tamar Soumagnac, Filipe B Abdalla, Ofer Lahav, Donnacha Kirk, I Sevilla, Emmanuel Bertin, Barnaby TP Rowe, J Annis, MT Busha, LN Da Costa, et al. Star/galaxy separation at faint magnitudes: application to a simulated dark energy survey. *Monthly Notices of the Royal Astronomical Society*, 450(1):666–680, 2015. 1

[18] EC Vasconcellos, RR De Carvalho, RR Gal, FL LaBarbera, HV Capelato, H Frago Campos Velho, Marina Trevisan, and RSdR Ruiz. Decision tree classifiers for star/galaxy separation. *The Astronomical Journal*, 141(6):189, 2011. 1

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 3

[20] Donald G York, Jennifer Adelman, John E Anderson Jr, Scott F Anderson, James Annis, Neta A Bahcall, JA Bakken, Robert Barkhouser, Steven Bastian, Eileen Berman, et al. The sloan digital sky survey: Technical summary. *The Astronomical Journal*, 120(3):1579, 2000. 1