Research in Astronomy and Astrophysics, 25:104003 (12pp), 2025 October
https://doi.org/10.1088/1674-4527/adf716
CSTR: 32081.14.RAA.adf716
© 2025. National Astronomical Observatories, CAS and IOP Publishing Ltd. All rights, including for text and data mining, AI training, and similar technologies, are reserved.

DeepAP: Deep Learning-based Aperture Photometry Feasibility Assessment and Aperture Size Prediction

Zheng-Jun Du[1,2,4], Qing-Quan Li[1,2,4], Yi-Cheng Rui[3], Yu-Li Liu[1,2], Yu-Ting Wu[1,2], Dong Li[1,2], Bing-Feng Seng[1,2], Yi-Fan Xuan[3], and Fa-Bo Feng[3]
[1] School of Computer Technology and Applications, Qinghai University, Xining 810016, China
[2] Qinghai Provincial Key Laboratory of Media Integration Technology and Communication, Xining 810016, China
[3] State Key Laboratory of Dark Matter Physics, Tsung-Dao Lee Institute & School of Physics and Astronomy, Shanghai Jiao Tong University, Shanghai 201210, China; ruiyicheng@sjtu.edu.cn
Received 2025 May 5; revised 2025 July 14; accepted 2025 July 28; published 2025 September 12

Abstract

Aperture photometry is a fundamental technique widely used to obtain high-precision light curves in optical survey projects like Tianyu. However, its effectiveness is limited in crowded fields, and the choice of aperture size critically impacts photometric precision. To address these challenges, we propose DeepAP, an efficient and accurate two-stage deep learning framework for aperture photometry. Specifically, for a given source, we first train a Vision Transformer (ViT) model to assess its feasibility of aperture photometry. We then train the Residual Neural Network (ResNet) to predict its optimal aperture size. For aperture photometry feasibility assessment, the ViT model yields an ROC AUC value of 0.96, and achieves a precision of 0.974, a recall of 0.930, and an F1 score of 0.952 on the test set. For aperture size prediction, the ResNet model effectively mitigates biases inherent in classical growth curve methods by adaptively selecting apertures appropriate for sources of varying brightness, thereby enhancing the signal-to-noise ratio (SNR) across a wide range of targets. Meanwhile, some samples in the test set have a higher SNR than those obtained by exhaustive aperture size enumeration because of the finer granularity of aperture size estimation. By integrating ResNet with the ViT network, the DeepAP framework achieves a median total processing time of 18 ms for a batch of 10 images, representing a speed-up of approximately 5.9×10^4 times compared to exhaustive aperture size enumeration. This work paves the way for the automatic application of aperture photometry in future high-precision surveys such as Tianyu and Legacy Survey of Space and Time. The source code and model are available at https://github.com/ruiyicheng/DeepAP.

Key words: methods: data analysis – techniques: image processing – techniques: photometric

1. Introduction

With the rapid advancement of observational technology, modern robotic photometric surveys are transforming our understanding of the time-domain universe. One of the leading facilities currently in operation is the Zwicky Transient Facility (ZTF; Bellm et al. 2019), which employs a wide-field camera on the Samuel Oschin Telescope at Palomar Observatory to scan the northern sky at the optimal cadence for supernova survey. Looking to the near future, the Vera C. Rubin Observatory Legacy Survey of Space and Time (LSST; Ivezić et al. 2019) is poised to revolutionize the field further. Similarly, the upcoming Tianyu Telescope (Feng et al. 2024), a one-meter robotic telescope under development in Lenghu, Qinghai, China, is designed to detect transiting planets and transient events. These surveys, both current and forthcoming, share a common need for accurate and efficient photometric measurements, which are foundational to time-domain astrophysics and the characterization of dynamic celestial phenomena.

High-precision photometry of point sources, such as stars and asteroids, is a critical step in these surveys. For example, ZTF adopts point-spread function (PSF) photometry as its default technique (Masci et al. 2019), which is suitable for measuring stellar fluxes in crowded fields. However, the accuracy of PSF photometry is sensitive to both the detector's sampling rate and atmospheric seeing conditions. Errors in PSF modeling can introduce biases in the resulting photometry (Howell 2000).

Aperture photometry is used to achieve higher precision, especially in exoplanet detection missions like Kepler (Borucki et al. 2010), the Transiting Exoplanet Survey Satellite (TESS; Ricker et al. 2014), the Wide Angle Search for Planets (Pollacco et al. 2006), and the Next Generation Transit Survey (Wheatley et al. 2018). For space-based missions such as Kepler and TESS, fixed-pixel apertures are used due to the spacecraft's excellent pointing stability (Morris et al. 2020). These apertures are selected based on the pixel-wise signal-to-

[4] First and second authors contribute equally to this work.

noise ratio (SNR) tests. However, in ground-based surveys, the need to include two additional apertures for accurate sky background subtraction makes this approach computationally expensive and operationally challenging. Furthermore, variations in the PSF, caused by changing atmospheric conditions, telescope optics, and tracking issues, complicate the task of determining the optimal aperture shape. As a practical solution, ground-based instruments typically adopt a fixed aperture shape, such as a circle or square, and adjust only the size to optimize performance.

The key parameters in aperture photometry are the sizes of three apertures: one for the target source and two for estimating the surrounding sky background. Traditionally, the inner aperture size is selected using a method called the "growth curve," which measures the SNR across a series of increasing aperture sizes and selects the size that yields the highest SNR (Stetson 1990). However, this method has several limitations. First, it only optimizes the size of the inner aperture and does not account for the background annuli. Second, it operates on a per-image basis, meaning that the optimal aperture for a single image may not generalize to the full night's data, even if that image is a stacked composite. The growth curve method maximizes the SNR for a single source in a single frame, not for the entire light curve over time. Third, it is prone to contamination from nearby stars; the algorithm may inadvertently increase the aperture size to include nearby sources, thus corrupting the target signal. In some cases, it even fails to find an optimal aperture within the tested range. In this case, aperture photometry is not feasible because the vicinity is too crowded. These limitations highlight two key challenges in aperture photometry: (1) determining whether a target is feasible for aperture photometry, and (2) if so, determining the optimal sizes for the apertures.

Recent advances in computer vision, particularly Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), provide powerful tools for astronomical image analysis. CNNs excel at extracting hierarchical local features (e.g., edges, gradients) and preserving spatial relationships, making them ideal for regression and classification tasks. For example, in denoising, Liu et al. (2025) proposed a self-supervised TDR method to reduce solar magnetogram noise from 8 G to 2 G while preserving faint galaxy structures in Hubble images. Elhakiem et al. (2021) optimized Astro U-net for multi-noise denoising via hybrid fusion, achieving superior PSNR/SSIM improvements. Vojtekova et al. (2021) demonstrated that Astro U-net emulates doubled exposure times, boosting SNR by 1.63× for Hubble Space Telescope (HST) data. Gheller & Vazza (2022) employed denoising autoencoders to detect faint radio cosmic webs in Low Frequency Array observations through cosmological simulations. For galaxy morphology classification, Dieleman et al. (2015) pioneered rotation-invariant CNNs for Galaxy Zoo data, achieving >99% accuracy on high-agreement samples. Large-scale morphology catalogs were enabled by Gravet et al. (2015) (CANDELS, 50k galaxies) and Domínguez Sánchez et al. (2018) (SDSS, 670k galaxies), with the latter reducing misclassification to <1%. To address limited labeled data, Luo et al. (2025a) proposed semi-supervised GC-SWGAN, achieving 75% accuracy on Galaxy10 DECals using only 20% labeled samples. For survey data enhancement, Luo et al. (2025b) developed Pix2WGAN, a CNN-based neural network, to transform SDSS/DECaLS images to HSC quality, improving structural visibility. Sandeep et al. (2021) implemented multiple CNN architectures for simultaneous galaxy classification (92.3% accuracy) and redshift prediction. For rare object identification, Primack et al. (2018) detected "blue nugget" galaxies in CANDELS, revealing their characteristic mass range ($10^{9.2-10.3}$ $M_{\odot}$), while Davies et al. (2019) achieved 77% recall for gravitational lenses in Euclid-like simulations. Advanced techniques include Burke et al. (2019)'s Mask R-CNN for deblending (98% galaxy precision) and Jia et al. (2020)'s real time Faster R-CNN for wide-field surveys.

In contrast, ViTs capture global contextual patterns through self-attention mechanisms, enhancing robustness for complex astronomical image classification. For example, Donoso-Oliva et al. (2023) introduced ASTROMER, a transformer-based model that generates light curve representations via self-supervised pretraining, improving downstream classifiers in limited-label scenarios. Similarly, Yang & Li (2024) applied ViT ("stellar-ViT") to stellar classification using SDSS photometric images, achieving 83.9%–86.3% accuracy across seven spectral classes and outperforming CNNs. For galaxy morphology, ViT-based approaches show strong performance: Bhavanam et al. (2024) used ViT-CNN hybrids to improve faint SDSS object classification. Cao et al. (2024) proposed Convolutional Visual Transformer (CvT) for galaxy morphology classification. Compared with other five class classification models, CvT outperforms >98% in terms of average accuracy, precision, recall, and F1 score. Yeganehmehr & Ebrahimnezhad (2025) further demonstrated ViT's capability, attaining 99.85% classification accuracy on galaxy images. Transformers also excel in time-series astronomy: Allam & McEwen (2024) developed a time-series transformer for photometric transient classification, achieving competitive performance (log-loss 0.507, AUC 0.98) on LSST challenge data with minimal feature engineering.

To fully exploit the strengths of ViTs in modeling global contextual information and CNNs in efficient local feature extraction for our specific task, we propose a robust two-stage framework to overcome the limitations of classical aperture photometry. Specifically, stage 1 employs the ViT classifier (Dosovitskiy et al. 2020) to assess the global feasibility of applying aperture photometry to a given source. If a source is deemed feasible, Stage 2 utilizes the ResNet-18 network (He et al. 2016) to predict the optimal sizes for the three apertures. This integration leverages ViT's capability in global source viability assessment and CNN's efficiency in precise aperture size prediction, thereby enhancing the overall photometric performance.

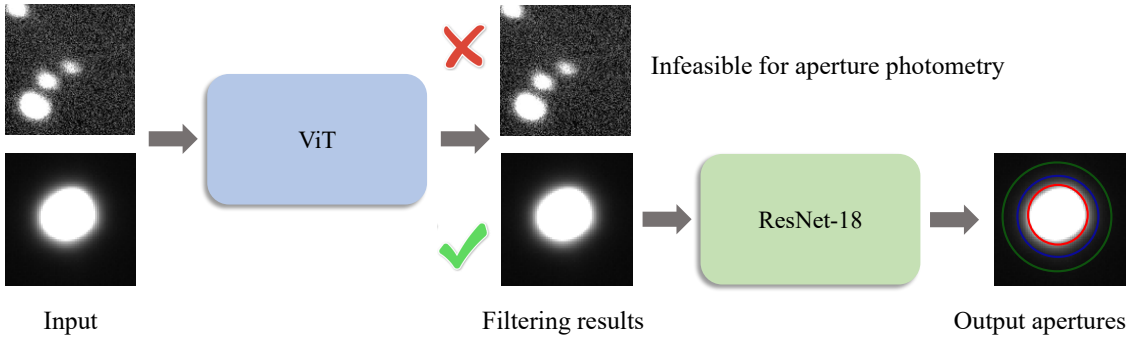


Figure 1. The pipeline of our method. Our method contains two stages. In stage 1, we filter out chopped images that are infeasible for aperture photometry via a ViT network. In stage 2, we train a ResNet, to obtain the accurate apertures of a given chopped image.

The remainder of the paper is organized as follows. Section 2 presents the data set and architecture design; Section 3 evaluates the performance of our method on real survey data and shows the comparison with other methods; and Section 4 concludes with implications for future improvements.

2. Method

2.1. Overview

Our goal is to achieve efficient yet accurate aperture photometry. To this end, we employ a two-stage deep learning framework. First, we train a ViT model to filter out sources with severe occlusion or excessive noise contamination. Subsequently, a ResNet-18 model is trained to accurately predict the optimal aperture sizes of an individual target. The pipeline of the proposed method is illustrated in Figure 1.

2.2. Data Set

To support the two-stage framework mentioned above, we collected 522 images from Muguang Observatory (coordinate: 31°10′07″N, 121°36′21″E; altitude: 30 m) on 2024 July 4. A QHY600M (Alarcon et al. 2023) camera was mounted on a CDK350 telescope.[5] The camera was set to use 1×1 binning, and no overscan was used. The exposure time for each image was set to 30 s. The Baader CMOS *L*-Filter (420–685 nm) was used in this observation.[6]

Input data: The input to our model comprises 128×128 pixel sub-images centered on the centroids of detected sources in the stacked astronomical images. These sub-images are generated with bias subtraction, flat-field calibration, image stacking, and alignment using a modified version of the Tianyu pipeline[7] (Rui et al. 2025). Among the 522 images, 506 high-quality images are utilized in the subsequent analysis. The source parameters, including centroid coordinates, full width at half maximum (FWHM), and other relevant metrics, are then extracted from the stacked image using SEP (Barbary 2016). For each detected source, a 128×128 pixel subimage centered on the source's centroid is extracted from the stacked image to serve as input to our data set.

Label construction: To construct the training labels, for each source in the input data, we employ brute-force enumeration to find its optimal aperture, and then assess its feasibility for aperture photometry based on apertures of surrounding sources. To optimize the aperture sizes for each object, we systematically vary the radii and select the combination that yields the light curve with the highest SNR. Specifically, for the ith source, the radius of the inner aperture, $r_{\mathrm{inn},i}$, is varied from 1 FWHM to 5 FWHM in increments of 0.05 FWHM. The radius of the middle aperture (the inner boundary of the sky background annulus), $r_{\mathrm{mid},i}$, is varied from $1.1r_{\mathrm{inn},i}$ to $2r_{\mathrm{inn},i}$, in steps of $0.1r_{\mathrm{inn},i}$. The outer aperture radius (the outer boundary for sky background estimation), $r_{\mathrm{out},i}$, is varied from $r_{\mathrm{mid},i} + 3$ to $r_{\mathrm{mid},i} + 15$, in steps of 3 pixels. For each set of radii ($r_{\mathrm{inn},i}$, $r_{\mathrm{mid},i}$, $r_{\mathrm{out},i}$), the corresponding light curve is extracted using SEP. The SNR of each light curve is defined as

$$\mathrm{SNR} = \frac{\mu_i}{\sigma_i}, \tag{1}$$

where μ_i and σ_i are the weighted mean and weighted standard deviation of the light curve for the ith object, respectively. The weight of each flux measurement in the light curve is assigned to be inversely proportional to the square of the photometric uncertainty for a single exposure, which are also provided by SEP. For each object in the training set, we enumerate ($r_{\mathrm{inn},i}$, $r_{\mathrm{mid},i}$, $r_{\mathrm{out},i}$) and obtain the corresponding SNR of the light curve.

To create the data set for aperture photometry feasibility classification, each image is labeled with a binary label: 0 indicates that the image is infeasible for aperture photometry due to overcrowding, while 1 indicates feasibility. Specifically, the image of the ith star is labeled as 1 if and only if there

[5] https://planewave.com/products/cdk350/
[6] https://www.baader-planetarium.com/en/baader-uv-ir-cut-l-filter-cmos-optimized.html
[7] https://github.com/ruiyicheng/Tianyu_pipeline

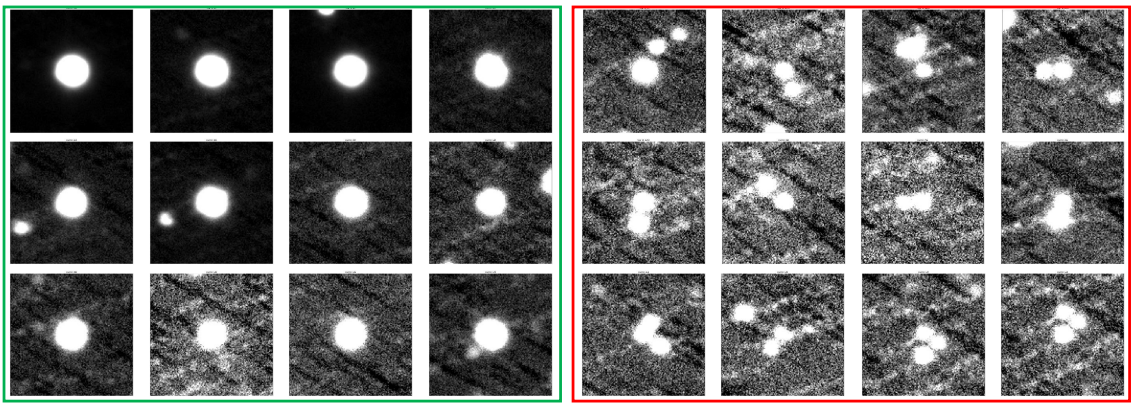


Figure 2. Visualization of samples from the aperture photometry feasibility determination data set. Images in the first four columns are labeled as 1 (feasible for aperture photometry), while the remaining columns are labeled as 0 (infeasible). The streak-like patterns in the background result from masks applied to hot pixels and do not affect the photometry. All resolved sources have been manually verified to ensure they represent genuine point sources rather than artifacts caused by these streak-like fluctuations.

exists a set of ($r_{\mathrm{inn},i}$, $r_{\mathrm{mid},i}$, $r_{\mathrm{out},i}$) such that for all $j \neq i$,

$$r_{ij} > r_{\mathrm{inn},i} + 3\,\mathrm{FWHM}_j; \quad \text{and} \quad (r_{ij} < r_{\mathrm{mid},i} - 3\,\mathrm{FWHM}_j \quad \text{or} \quad r_{ij} > r_{\mathrm{out},i} + 3\,\mathrm{FWHM}_j), \tag{2}$$

where r_{ij} denotes the distance between the ith and jth resolved sources, and FWHM_j is the FWHM of the jth source. The image is labeled as 0 otherwise. The 3 FWHM gap is required to ensure the existence of an uncontaminated aperture for photometry when a source is marked as feasible because the flux contamination of other sources is negligible at this distance (Howell 2000).

Applying this criterion to all 2677 sources, 2311 are labeled as feasible (label 1). The sources are then divided into training and testing subsets in a 7:3 ratio, resulting in 1983 sources in the training set and 804 in the test set for the aperture photometry feasibility classification task. Representative examples from the data set are presented in Figure 2. To augment the training set, we apply rotational transformations at angles of 36°, 72°, 108°, ..., 324° to full-frame stacked images using SCAMP (Bertin 2006) and SWARP (Bertin 2010). The corresponding 128 × 128 image cutouts are extracted around each source using the world coordinate system information and the R.A./ decl. positions of the 1983 training sources. This augmentation yields a total of 19,830 128 × 128 training images. The same labels are assigned to the augmented images, based on the assumption of rotational symmetry.

For aperture size optimization, we use the same 2311 sources labeled as feasible in the previous step. Each image corresponding to the ith source is labeled with the set of aperture radii ($r_{\mathrm{inn},i}$, $r_{\mathrm{mid},i}$, $r_{\mathrm{out},i}$) that maximize the SNR, as defined in Equation (1), under the constraint specified in Equation (2) for all $j \neq i$. The data set is then split into training and testing subsets using a 7:3 ratio, yielding 1617 training sources and 694 test sources. Representative examples from this data set are shown in Figure 3. As before, we apply 9-fold rotational augmentation to the training sources, resulting in 16,170 128 × 128 images. The aperture radius labels are inherited from the corresponding unrotated sources, again under the assumption of rotational symmetry.

2.3. Aperture Photometry Feasibility Assessment

The first stage of our method employs a ViT network to filter out celestial objects that are infeasible for aperture photometry due to severe occlusion or excessive noise contamination. The network architecture and data flow are illustrated in Figure 4 and are described below.

For an input clipped image $x \in \mathbb{R}^{H\times W\times C}$, where $H = W = 128$, $C = 1$ (image height, width, and channels), we first split it into N patches of size $P \times P$, with $N = 64$, $P = 16$. Each patch $x_p^i \in \mathbb{R}^{P\times P\times C}$ is then flattened into a 1D vector and linearly projected into a higher-dimensional space using a learnable patch embedding $E(\cdot)$. To preserve spatial information, a positional embedding $E_{\mathrm{pos}} \in \mathbb{R}^{N\times P^2}$ is added to the patch embeddings. Additionally, a classification token x_{class} is prepended to the sequence to aggregate global information for the final classification. The input to the Transformer Encoder is thus created as

$$Z_0 = [x_{\mathrm{class}}; E(x_p^1); E(x_p^2); \cdots; E(x_p^N)] + E_{\mathrm{pos}}. \tag{3}$$

The input sequence Z_0 is processed by L cascaded Transformer encoder layers. Each layer consists of two key components:

1. *Multi-head Self-Attention (MSA).* MSA captures long-range dependencies between image patches by enabling global interactions. Its multi-head design learns diverse patterns in parallel subspaces. In implementation, each head applies independent learnable projections to transform these shared embeddings into three distinct vector

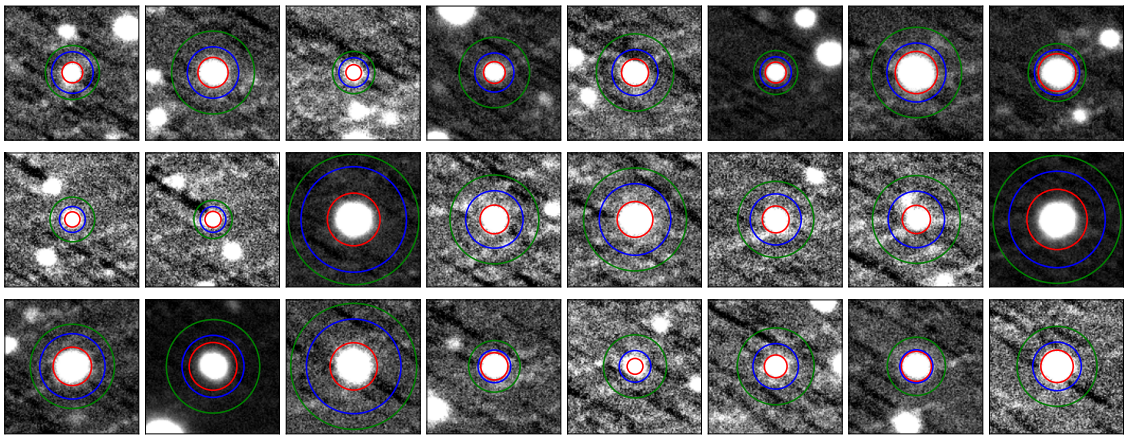


Figure 3. Visualization of samples from the aperture prediction data set (inner aperture: red, middle aperture: blue, outer aperture: green). The rings in each image represent the optimal apertures for photon counting (inner ring) and sky background estimation (outer two rings), determined through exhaustive enumeration.

sets: Queries (Q), Keys (K) and Values (V). Within each head, every query interacts with all keys to compute attention weights that dynamically blend values. This allows global context aggregation while maintaining head-specific perspective diversity. The combined head outputs are linearly fused, implemented as

$$Z_l' = \mathrm{MSA}(\mathrm{LN}(Z_{l-1})) + Z_{l-1}. \quad (4)$$

2. *Feed-Forward Network (MLP).* A two-layer perceptron with nonlinearity, applied to each token independently. Similar to MSA, layer normalization and residual connections are used

$$Z_l = \mathrm{MLP}(\mathrm{LN}(Z_l')) + Z_l'. \quad (5)$$

The final output token Z_L^0 (corresponding to the classification token) is passed through a Multi-Layer Perceptron (MLP) followed by a softmax activation, producing a probability distribution $y = [y_0, y_1]$ over the two classes

$$y = \mathrm{Softmax}(\mathrm{MLP}(Z_L^0)), \quad (6)$$

where y_0 and y_1 represent the probabilities of the input image being infeasible (label 0) or feasible (label 1) for aperture photometry, respectively.

The network is trained using Binary Cross-Entropy loss, defined as

$$\mathcal{L}_{\mathrm{fa}} = -p_0 \log(y_0) - p_1 \log(y_1), \quad (7)$$

where $p = [p_0, p_1]$ is the ground-truth label (one-hot encoded). This stage effectively filters out low-quality images, ensuring that only reliable candidates proceed to aperture prediction.

2.4. Optimal Aperture Size Estimation

In this stage, we adopt ResNet-18, a well-proven CNN in computer vision, as the backbone to predict the optimal apertures of a given source. The ResNet-18 network is illustrated in Figure 5, which takes a 128×128 chopped image as input and outputs three aperture values indicating the inner (r_1), middle (r_2) and outer (r_3) radii, respectively.

The ResNet-18 network consists of a series of convolutional layers, followed by a global average pooling (GAP) layer and a fully connected layer with three neurons. The convolutional layers are organized into eight "ResNet blocks," each containing a set of convolutional layers with skip connections that allow the network to learn residual functions. This design enables the network to achieve deeper architectures while mitigating the problem of vanishing gradients during training.

The data flow through the network can be described as follows:

1. *Input.* A 128×128 pixel single-channel chopped image is fed into the network.
2. *Initial Convolution and Pooling.* The input image is first processed by a 7×7 convolutional layer with 64 output channels and a stride of 2, followed by a 3×3 max pooling layer with a stride of 2. This initial block reduces the spatial resolution of the image while increasing the depth.
3. *ResNet Blocks.* The output of the initial block is then passed through seven Residual Neural Network (ResNet) blocks, each containing a series of 3×3 convolutional layers with 64, 128, 256, or 512 output channels depending on the block's position in the network. The skip connections within each block allow the network to learn the identity function, enabling it to learn more complex relationships between pixels in the input image.
4. *Global Average Pooling.* After the ResNet blocks, the network applies a GAP layer, which reduces the spatial dimensions of the feature maps to a single value per feature map.
5. *Fully Connected Layer.* The output of the GAP layer is then passed to a fully connected layer with three neurons.

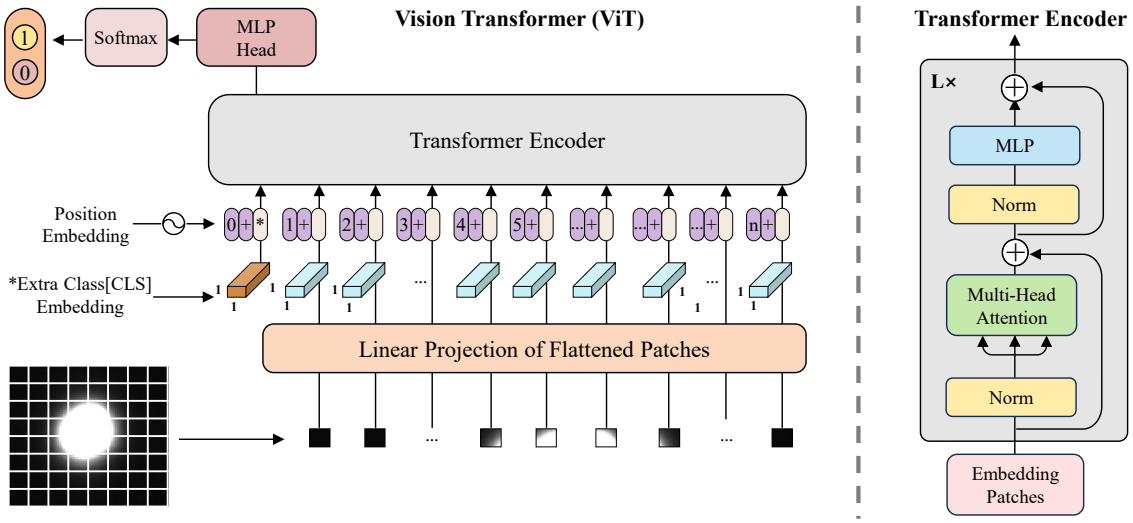


Figure 4. The ViT model for aperture photometry feasibility assessment. The ViT processes an input image by first dividing it into flattened patches. Then, each patch is linearly projected into an embedding vector, combined with positional embeddings and a learnable classification token [CLS]. Next, the sequence is fed into a multi-layer Transformer encoder, followed by an MLP block and the softmax activation function. Finally, it outputs a binary classification probability distribution.

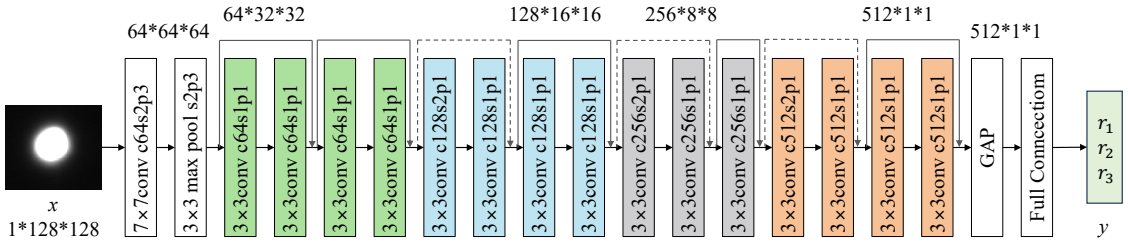


Figure 5. The ResNet-18 model for optimal aperture size estimation. The numbers following "c," "s," and "p" denote the channel, stride, and padding parameters of each layer, respectively. Solid arrows indicate direct forward connections, while dashed arrows represent down-sampling operations implemented with 1×1 convolution layers of stride 2.

This layer outputs the three real numbers representing the sizes of the three circular apertures.

We employ the Mean Squared Error (MSE) to measure the difference between the network's predicted aperture sizes (r_1, r_2, r_3) and the ground truth values $(r_{\rm inn}, r_{\rm mid}, r_{\rm out})$

$$\mathcal{L}_{\rm ap} = \frac{1}{3} \left\| (r_1, r_2, r_3) - (r_{\rm inn}, r_{\rm mid}, r_{\rm out}) \right\|^2 . \quad (8)$$

The ResNet-18 architecture offers a robust and efficient framework for learning features from a chopped image and regressing aperture sizes. Skip connections and deep network structure enable the network to capture complex patterns and relationships in the data, leading to accurate and reliable predictions.

3. Experiments

All of our experiments were conducted on a computer with the following configurations: An 13th Gen Intel(R) Core(TM) i7-13700 CPU (14 cores/20 threads), 2.10 GHz, coupled with 128 GB memory. In addition, the system is equipped with an NVIDIA GeForce RTX 4090 GPU that features 24 GB of GDDR6X VRAM, 16384 CUDA cores, and delivers 82.6 TFLOPS of single-precision floating-point performance.

3.1. Training

In our two-stage framework, the ViT model is trained for 50 epochs with a batch size of 256, while the ResNet-18 model is trained for 100 epochs with a reduced batch size of 12. Both models use the same Adam optimizer and learning rate (0.0001).

For the ViT model, the training loss and testing loss show a fast decline. Eventually, after 50 epochs, the training loss stabilizes around 0.10, and the testing loss stabilizes around 0.21. The consistent downward trend and convergence of both training and testing losses indicate effective optimization without overfitting. This alignment suggests that the model successfully captures meaningful patterns in the data, achieving relatively accurate classification.

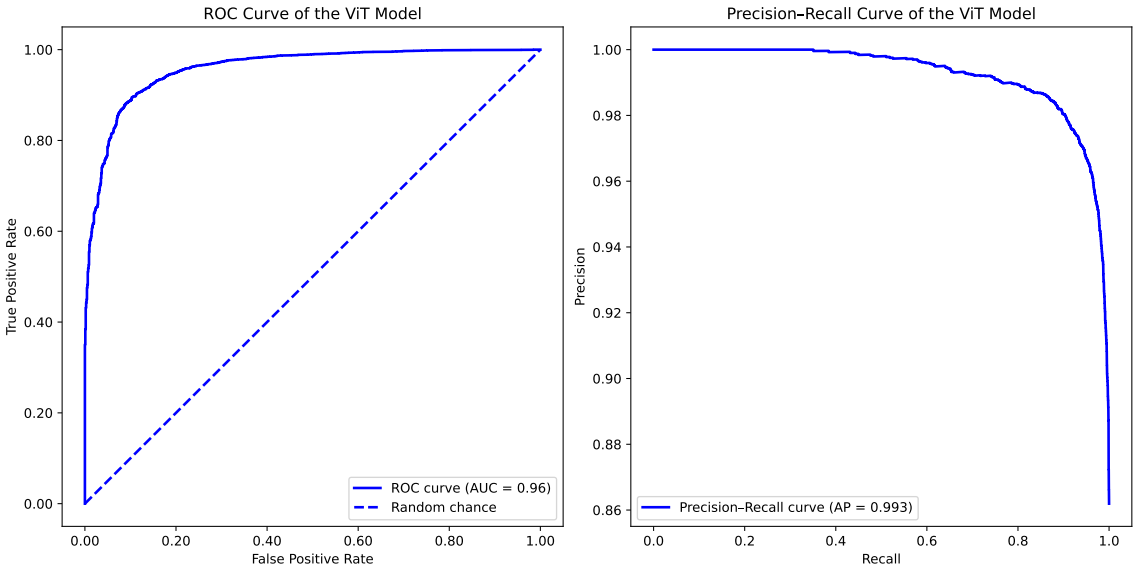


Figure 6. The ROC and precision–recall curves of the ViT Model.

The ResNet-18 model exhibits rapid convergence in the training set, with a loss rapidly decreasing from 0.382 to 0.074 within the first 20 epochs. After 100 epochs, the training loss stabilizes at a low value of approximately 0.007. The test loss remains consistent, but higher than the training loss throughout the training process. The test loss rapidly converges after the first epoch, and after 100 epochs, the testing loss is approximately 0.240. As the loss of the ResNet-18 model as defined in Equation (8), a loss of 0.240 corresponds to an average aperture prediction error of $\sqrt{0.240} \approx 0.5$ pixels. Given that the step size for enumerating the outer radius is 3 pixels, the theoretical loss for a model possessing infinite granularity, which accounts for the discretization limit, is approximately: $\frac{1}{3}\int_0^{3/2} x^2 dx = 0.375$. This value reflects the expected loss when the true optimal radius lies between the enumerated values. The test loss of 0.240 is well below this bound, indicating strong generalization capabilities despite the apparent discrepancy. The observed loss is primarily attributable to the granularity of the radius enumeration, rather than to substantial overfitting.

3.2. Performance on Test Set

For the aperture photometry feasibility assessment, the confusion matrix of the ViT model is presented in Table 1. It achieves a precision of 0.974, a recall of 0.930, and an F1 score of 0.952 (support: 6,930). The performance of the ViT model on the test set is illustrated in Figure 6. The left panel shows the ROC curve (AUC = 0.96), while the right panel presents the precision–recall curve (AP = 0.993). The ViT model exhibits a median inference time of 4 ms for a batch size of 10 images, corresponding to an average of 0.4 ms per image.

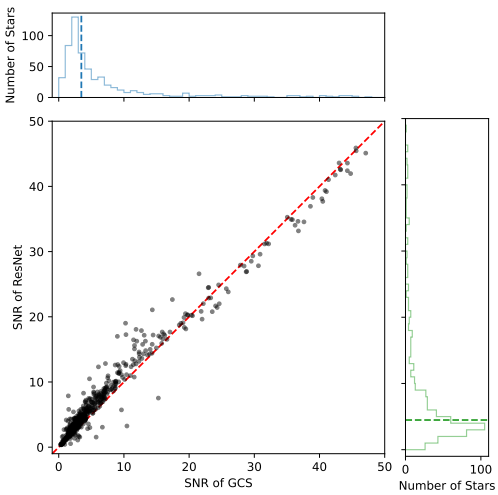
Table 1
The Confusion Matrix of the ViT Model (Threshold: 0.5)

	Predicted Feasible	Predicted Infeasible
Actual Feasible	TP: 6446	FN: 484
Actual Infeasible	FP: 173	TN: 937

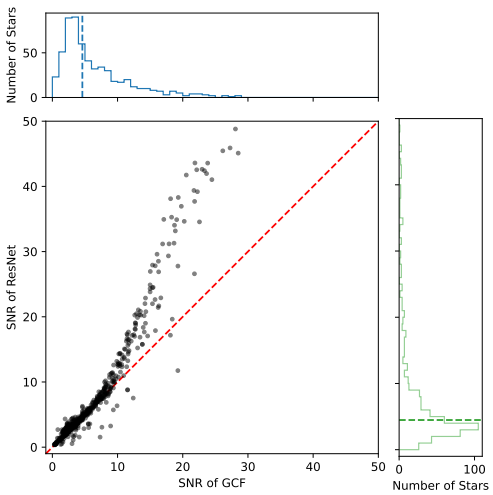
For the aperture size determination task, we compare the performance of the ResNet-18 model with the following baseline models:

1. *GCS (growth curve on the stacked image).* This method applies the "growth curve" approach to the stacked image, with background subtraction performed using SEP.
2. *GCF (growth curve on the first image).* This method applies the "growth curve" approach to the first calibrated image, with background subtraction performed using SEP.
3. *Exhaustive Enumeration.* The optimal aperture size obtained through an exhaustive search. This method was also used to generate the ground truth labels for the training and testing data sets, as described in Section 2.2.

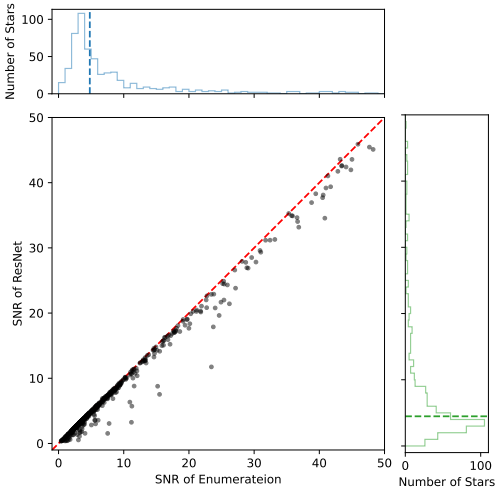
Optimal aperture sizes using the growth curve method are determined by varying the trial aperture radius from 1 FWHM to 5 FWHM in 0.05 FWHM steps. In the unrotated test set, comprising 694 sources, we assess the feasibility of obtaining an optimal aperture for the GCS and GCF methods. Feasibility as



(a) ResNet-18 vs. GCS (median SNR: 3.47)



(b) ResNet-18 vs. GCF (median SNR: 4.60)



(c) ResNet-18 vs. enumeration (median SNR: 4.79)

Figure 7. Comparison of the SNRs of light curves obtained with our ResNet-18 (Median SNR: 4.48), two growth curve methods, and the aperture enumeration method. The dashed lines in the histograms indicate the median SNR for each method.

defined by the first condition of Equation (2) is achieved for 658 sources with GCS and 692 sources with GCF. The ResNet-18 model is also applied to estimate aperture sizes on the same test set, yielding 600 sources satisfying the feasibility condition in Equation (2). To ensure a consistent comparison across GCS, GCF, and ResNet-18, we considered only the sources where a feasible aperture is obtained by all three methods. This intersection results in a common sample of 587 sources, which serve as the basis for all subsequent comparisons.

Figure 7 presents a comparison of SNRs generated by different methods, as defined in Equation (1). The SNR obtained using estimated apertures from ResNet-18 is higher than that of the growth curve method applied to the stacked image (GCS) for 83.5% of the sources. Meanwhile, ResNet-18 outperforms the growth curve method applied to the first calibrated image (GCF) for 66.4% of the targets. Specifically, ResNet-18 demonstrates superior performance compared to GCS for faint targets with low SNR, and better performance than GCF for bright targets with high SNR. Figure 8 shows a comparison of the inner aperture sizes selected by ResNet-18, aperture size enumeration and the growth curve methods. We observe that GCS typically selects larger apertures than GCF. This difference can be attributed to the characteristics of the images: the stacked image benefits from averaged-out sky background fluctuations, allowing larger apertures to capture more signal photons. However, this strategy is less effective for faint sources, as it also incorporates more background noise. In contrast, GCF, applied to a single, potentially noisier image, selects a smaller aperture to better match the source profile and minimize background noise. While this minimizes noise, it leads to poorer performance for bright sources by excluding some source photons. The aperture size selected by ResNet-18 adaptively behaves similarly to GCS for

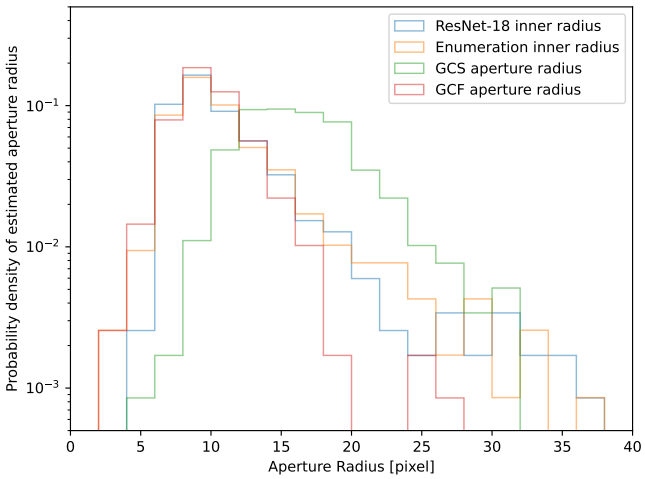


Figure 8. Distribution of the estimated inner aperture size of ResNet-18 compared with the results of aperture enumeration and growth curve.

bright sources and GCF for faint sources. Meanwhile, Figure 7(c) compares the performance of ResNet-18 with that of the exhaustive enumeration method, which serves as our training labels. Surprisingly, for 13 targets, our method outperforms the original training labels. Examples of improved SNR using ResNet-18 are shown in Figure 9. For instance, in the case of Gaia DR3 2188595766451568768 (Gaia G-band magnitude: $G = 12.2255$), a representative example of a bright star with a large FWHM, the ResNet-18 output exceeds the resolution of the training labels. This improvement is attributed to the finer output granularity of ResNet-18 compared to the step size used in the enumeration trials. Meanwhile, for Gaia DR3 2188679565554441856 ($G = 17.0869$), the performance gain arises because the optimal outer aperture radius lies beyond the enumeration range ($r_{\rm out}$–$r_{\rm mid}$ = 16.45), a situation frequently encountered with faint stars. The median inference time for the ResNet-18 model is 14 ms when processing a batch of 10 images. Meanwhile, we observe that the inner radius estimated by ResNet-18 closely matches that obtained by enumeration. The Kolmogorov–Smirnov test yields a p-value of 0.012, indicating no statistically significant difference between the distributions at the 3σ level. These results indicate that ResNet-18 successfully mitigates the inherent biases of classical growth curve methods across different source brightnesses and image types.

Integrating the ResNet-18 with the ViT network, the DeepAP framework achieves a total processing time of 18 ms for 10 images, representing a speed-up of approximately 5.9×10^4 times compared to exhaustive aperture size enumeration. These results highlight the strong generalizability and high speed of DeepAP.

3.3. Case Study: Gaia DR3 2188697333838401536

Gaia DR3 2188697333838401536 (R.A.: 305.°537403919; decl.: 59.°563370877) is a relatively faint star ($G = 15.9809$)

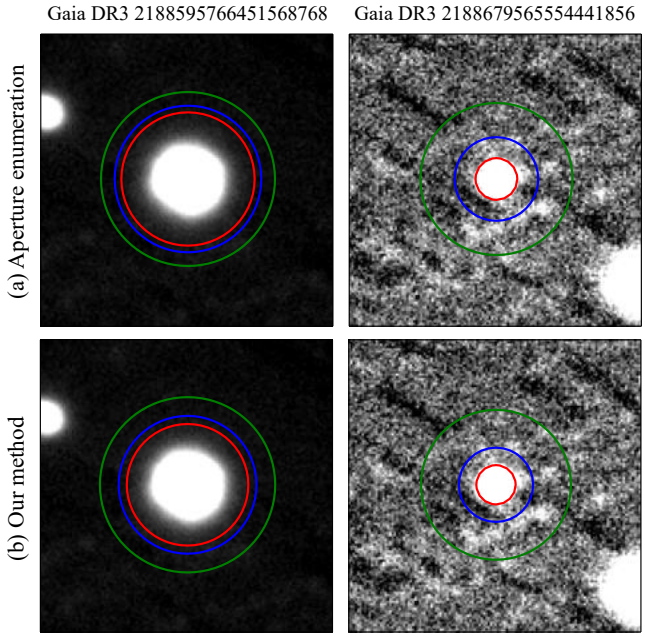


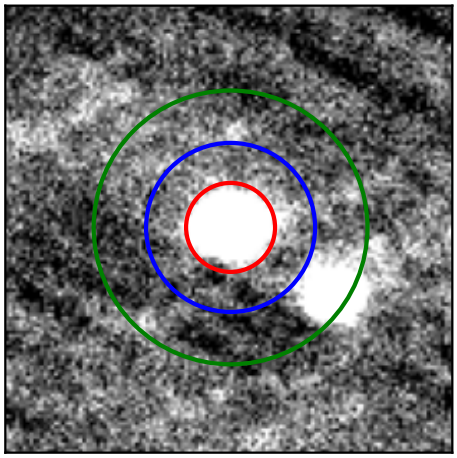
Figure 9. Comparison of the predicted aperture sizes from our method (ResNet-18) with those obtained through aperture enumeration. In both cases, the apertures predicted by our method achieve a higher SNR than those from aperture enumeration.

within our field of view. This object serves as a false positive example for the ViT model: while its true label is 0 (infeasible), the model incorrectly classifies it as 1 (feasible), with a softmax output exceeding 0.5. Figure 10 presents the apertures selected by both the enumeration baseline and the ResNet-18 model, overlaid on the stacked image and the first calibrated image of the source.

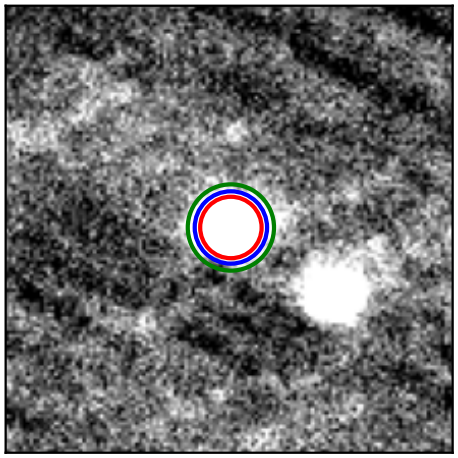
Interestingly, the ResNet-18 model identifies a feasible aperture (as defined in Equation (2)) that successfully avoids background contamination, whereas the enumeration method fails to do so. The inner aperture radius predicted by ResNet-18 is 0.97 FWHM—slightly smaller than the lower bound used in enumeration—highlighting the limitation of the latter's fixed search space. Although the ResNet-18 aperture may appear too small in the stacked image (Figure 10(b)), it aligns well with the source profile in the first calibrated image (Figure 10(c)), consistent with the analysis in Section 3.2. Furthermore, as shown in Appendix, the ResNet-18 aperture provides photometric robustness against contamination from a nearby source located at the lower right. This example underscores the strong generalization capability of DeepAP.

4. Conclusion and Discussion

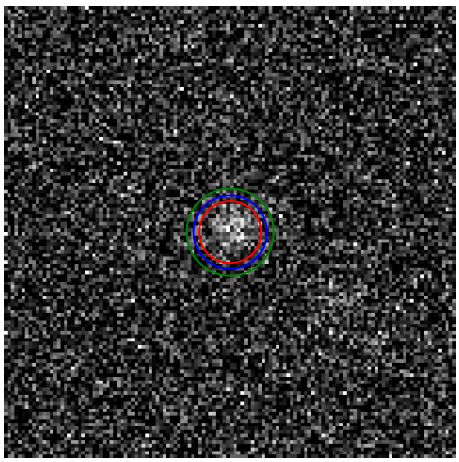
In this study, we introduced DeepAP, an efficient and accurate two-stage deep learning framework for aperture photometry, specifically designed to meet the demands of modern and future large-scale astronomical surveys. By



(a) Aperture from enumeration overlaid on stacked image



(b) Aperture from ResNet-18 overlaid on stacked image



(c) Aperture from ResNet-18 overlaid on first calibrated image

Figure 10. Qualitative comparison of aperture selection between enumeration and ResNet-18 for an infeasible case misclassified by the ViT model.

combining a ViT for filtering infeasible sources for aperture photometry and a ResNet-18 network for predicting optimal aperture sizes, our method addresses the corresponding two major challenges traditionally encountered in aperture photometry: determining target suitability and selecting optimal aperture parameters.

Experimental results demonstrate that DeepAP achieves strong performance in both feasibility assessment and aperture size determination tasks. The ViT model achieves an area under the ROC curve (AUC) of 0.96, and achieves a precision of 0.974, a recall of 0.930, and an F1 score of 0.952 on the test set. For aperture size determination, the ResNet model effectively mitigates the biases inherent in classical growth curve methods by adaptively selecting apertures appropriate for sources with different brightnesses, leading to enhanced SNR across a wide range of targets. By integrating ResNet with the ViT network, the DeepAP framework achieves a median total processing time of 18 ms for a batch of 10 images, representing a speed-up of approximately 5.9×10^4 times compared to exhaustive aperture size enumeration.

These results suggest several important implications:

1. *Scalability and Automation.* DeepAP is capable of processing large volumes of sources rapidly and automatically, making it ideal for surveys such as LSST and Tianyu, where manual photometric tuning is infeasible.
2. *Generalization for more sources.* By successfully applying DeepAP to targets beyond the limits of exhaustive enumeration, aperture photometry can be extended to sources traditionally considered too crowded for such techniques, thereby broadening the range of scenarios in which aperture photometry can be effectively applied.

We need to notice that the example as shown in Figure 10 and Appendix does not guarantee that the ResNet-18 model can always find a feasible aperture. Meanwhile, we cannot simply assume that all the false positive cases of the ViT model represent the error of labels. The ResNet-18 model does not find feasible apertures for other ViT false positive cases, which should be considered as the failure cases of this method. In real utilization, the logit threshold should be carefully chosen to control the false positive rate.

Despite its advantages, DeepAP still has opportunities for future improvements. Our training data set was based on images from a single telescope-camera setup under specific observing conditions. It is worth noting that DeepAP's performance depends on the training data set, and retraining or fine-tuning will be necessary when applying it to different telescopes or observational conditions. Transfer learning or domain adaptation techniques could also be considered. Furthermore, when performing relative flux calibration, the aperture size should be consistent within a group of stars to ensure that a comparable proportion of photons is measured for each source. In future work, we can apply more advanced numerical techniques like GPU acceleration to expand the range of aperture enumeration and improve the training set's data quality. Meanwhile, an ensemble of images should be used as input, with the model simultaneously determining both the groupings for relative photometry and the optimal aperture size for each group, making it more effective in practice.

In conclusion, DeepAP offers a practical, scalable, and high-precision method to determine the feasibility and the aperture size of aperture photometry, aligning well with the needs of next-generation time-domain astronomy projects like LSST and Tianyu. Its deployment in future surveys has the potential to substantially enhance the scientific return by enabling accurate and efficient light curve extraction at scale.

Acknowledgments

This work is supported by the Youth Program of the Natural Science Foundation of Qinghai Province (2023-ZJ-951Q), and Qinghai University Research Ability Enhancement Project (2025KTSQ26).

Appendix
Effects of Contamination of Nearby Variable Sources

In this appendix, we conduct an injection test to illustrate the effects of contamination caused by nearby variable sources and the importance of choosing an aperture to avoid it. The target shown in Figure 10 is used as an example.

In this test, we use two sets of aperture radii for the injection test:

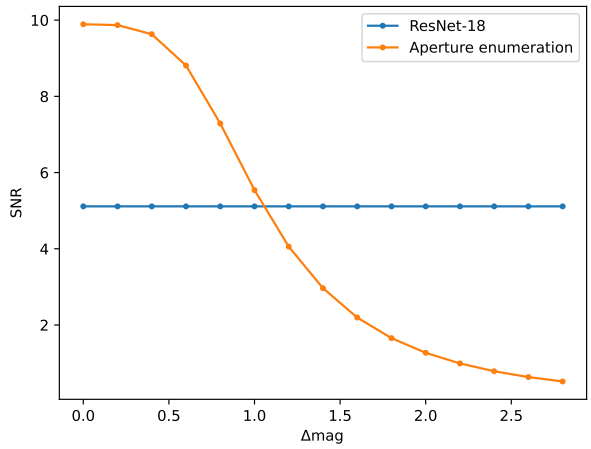


Figure A1. SNR comparison between ResNet-18 and the enumeration aperture across different variations of the contaminating source.

1. The aperture obtained through enumeration yields the optimal SNR for the light curve, as shown in Figure 10(a). The SNR of the light curve on the image before injection, defined in Equation (1), is 9.89. However, this aperture is contaminated by a nearby source (Gaia DR3 2188697333838401152) located at the lower right. In this case, the aperture parameters are $(r_{\rm inn}, r_{\rm mid}, r_{\rm out}) = (12.74, 24.20, 39.20)$.
2. The aperture obtained by the ResNet-18 model yields an SNR of 5.11, which is lower than that of the enumerated aperture. Here, the aperture parameters are $(r_{\rm inn}, r_{\rm mid}, r_{\rm out}) = (8.80, 10.36, 12.35)$. This aperture is smaller, and therefore collects fewer photons than the enumerated one. However, it avoids contamination from the star in the lower right.

The injection δ upon the pixel (x, y) of the ith chopped image is

$$\delta(\Delta\mathrm{mag}) = \sin(2\pi\frac{i}{506}) \cdot (100^{\Delta\mathrm{mag}/5} - 1) \cdot A_c \exp\left[\frac{(x - x_c)^2 + (y - y_c)^2}{2w_c^2}\right], \quad \text{(A1)}$$

where $A_c = 11$, $x_c = 93$, $y_c = 44$, and $w_c = 5$ represent the profile of the contamination star (Gaia DR3 2188697333838401152) in the lower right of Figure 10; Δmag is the variational magnitude of the contamination star . The original flux of the contamination source ($G = 16.7870$) is about half that of the target source ($G = 15.9808$).

We calculate the SNR of a light curve obtained by these two apertures at different Δmag. The results are shown in Figure A1. The SNR of the light curve obtained from the enumeration aperture decreases with respect to larger variation of the contamination, while the aperture obtained by ResNet-18 is not affected by the contamination. Meanwhile, the light curves at

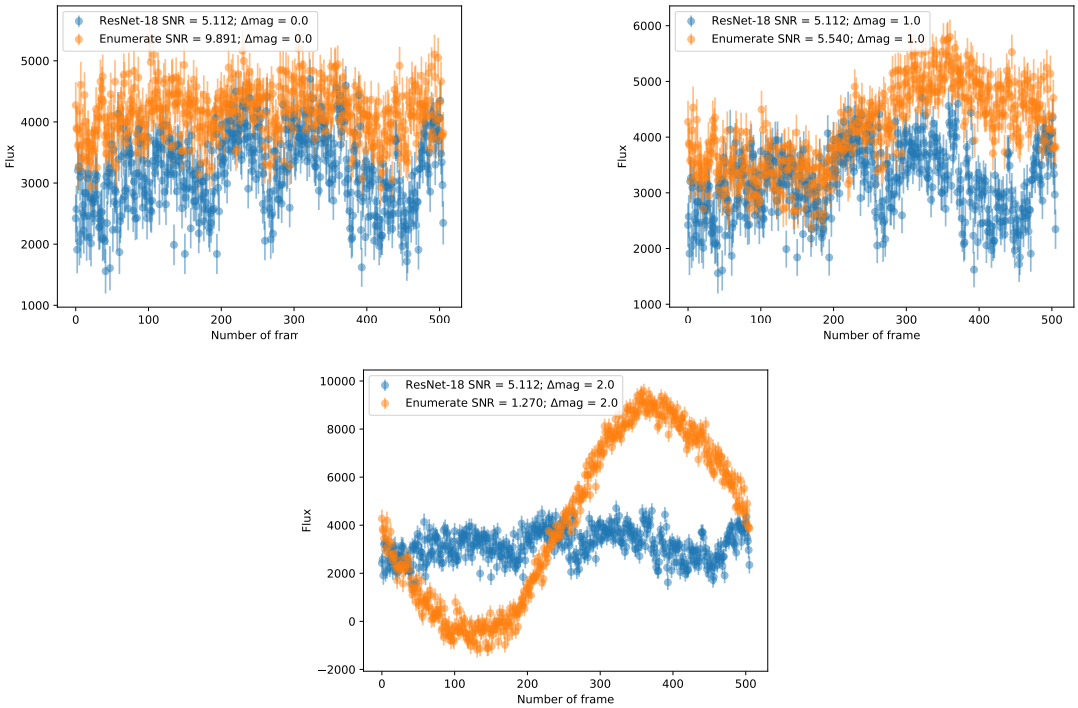


Figure A2. Light curve obtained using aperture of enumeration and ResNet-18 with different variations of contamination source.

Δmag = 0 and Δmag = 3 are shown in Figure A2. As the variability of the contaminating star increases, the flux extracted by the contaminated aperture increasingly reflects the contaminant's signal rather than that of the target source. In cases of high contaminant variability (Δmag = 2.0), the extracted light curve from the contaminated aperture even exhibits behavior opposite to the expected signal, driven primarily by fluctuations in the contaminant's flux rather than the target's. Therefore, the aperture extracted by ResNet-18 is more robust to the contamination star variation in this case.

References

Alarcon, M. R., Licandro, J., Serra-Ricart, M., et al. 2023, PASP, 135, 055001
Allam, T., & McEwen, J. D. 2024, RASTI, 3, 209
Barbary, K. 2016, JOSS, 1, 58
Bellm, E. C., Kulkarni, S. R., Graham, M. J., et al. 2019, PASP, 131, 018002
Bertin, E. 2006, in ASP Conf. Ser. 351, Astronomical Data Analysis Software and Systems XV, ed. C. Gabriel et al. (San Francisco, CA: ASP), 112
Bertin, E., 2010 SWarp: Resampling and Co-adding FITS Images Together, Astrophysics Source Code Library, ascl:1010.068
Bhavanam, S. R., Channappayya, S. S., Srijith, P. K., & Desai, S. 2024, Ap&SS, 369, 92
Borucki, W. J., Koch, D., Basri, G., et al. 2010, Sci, 327, 977
Burke, C. J., Aleo, P. D., Chen, Y.-C., et al. 2019, MNRAS, 490, 3952
Cao, J., Xu, T., Deng, Y., et al. 2024, A&A, 683, A42
Davies, A., Serjeant, S., & Bromley, J. M. 2019, MNRAS, 487, 5263
Dieleman, S., Willett, K. W., & Dambre, J. 2015, MNRAS, 450, 1441
Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., et al. 2018, MNRAS, 476, 3661
Donoso-Oliva, C., Becker, I., Protopapas, P., et al. 2023, A&A, 670, A54
Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. 2020, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, in Int. Conf. Learning Representations *(Virtual Event, Austria)* (OpenReview.net),
Elhakiem, A. A., Ghoniemy, T. E., & Salama, G. I. 2021, Astronomical image denoising based on Convolutional Neural Network, in 2021 10th Int. Conf. Intelligent Computing and Information Systems (ICICIS) *(Cairo, Egypt)* (Piscataway, NJ: IEEE), 51
Feng, F., Rui, Y., Du, Z., et al. 2024, AcASn, 65, 34
Gheller, C., & Vazza, F. 2022, MNRAS, 509, 990
Gravet, R., Cabrera-Vives, G., Pérez-González, P. G., et al. 2015, ApJS, 221, 8
He, K., Zhang, X., Ren, S., & Sun, J. 2016, Deep Residual Learning for Image Recognition, in Proc. IEEE Conf. Computer Vision and Pattern Recognition (Las Vegas, NV: IEEE), 770
Howell, S. B. 2000, Handbook of CCD Astronomy, Vol. 2 (Cambridge: Cambridge Univ. Press)
Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111
Jia, P., Liu, Q., & Sun, Y. 2020, AJ, 159, 212
Liu, T., Quan, Y., Su, Y., et al. 2025, NatAs, 9, 608
Luo, Z., Chen, J., Chen, Z., et al. 2025a, ApJS, 279, 17
Luo, Z., Zhang, S., Chen, J., et al. 2025b, ApJS, 277, 22
Masci, F. J., Laher, R. R., Rusholme, B., et al. 2019, PASP, 131, 018003
Morris, R. L., Twicken, J. D., & Smith, J. C. 2020, in Kepler Data Processing Handbook: Photometric Analysis, ed. J. M. Jenkins, 6, Kepler Science Document KSCI-19081-003
Pollacco, D. L., Skillen, I., Collier Cameron, A., et al. 2006, PASP, 118, 1407
Primack, J., Dekel, A., Koo, D., et al. 2018, ApJ, 858, 114
Ricker, G. R., Winn, J. N., Vanderspek, R., et al. 2014, Proc. SPIE, 9143, 914320
Rui, Y., Xuan, Y., Zheng, S., et al. 2025, PASP, 137, 064501
Sandeep, V., Sen, S., & Santosh, K. 2021, Analyzing and Processing of Astronomical Images using Deep Learning Techniques, in 2021 IEEE Int. Conf. on Electronics, Computing and Communication Technologies (CONECCT) *(Bangalore, India)* (IEEE), 1
Stetson, P. B. 1990, PASP, 102, 932
Vojtekova, A., Lieu, M., Valtchanov, I., et al. 2021, MNRAS, 503, 3204
Wheatley, P. J., West, R. G., Goad, M. R., et al. 2018, MNRAS, 475, 4476
Yang, Y., & Li, X. 2024, Univ, 10, 214
Yeganehmehr, A., & Ebrahimnezhad, H. 2025, AUT Journal of Electrical Engineering, 57, 385