Research in Astronomy and Astrophysics, 25:125014 (12pp), 2025 December
https://doi.org/10.1088/1674-4527/ae0c79
CSTR: 32081.14.RAA.ae0c79

STAR-YOLO: A Model for Detection of Complex Galaxy Morphology

Sheng-Qiang Zhu and Zhi-Jing Xu
School of Information Engineering, Shanghai Maritime University, Shanghai 201306, China; zjxu@shmtu.edu.cn
Received 2025 June 18; revised 2025 August 24; accepted 2025 September 23; published 2025 October 29

Abstract

Galaxy morphology detection is a pivotal task for unraveling cosmic evolutionary mechanisms, yet existing models exhibit insufficient detection accuracy for irregular and small-target galaxies. To address this, this paper proposes the STAR-YOLO galaxy morphology detection model. The backbone network incorporates the novel Multi-scale Attentive Context Aggregation module, which deeply integrates multi-scale dilated convolution with a progressive spatial-channel attention mechanism to enhance feature extraction for irregular and small galaxies. Meanwhile, we design the lightweight Lightweight Efficient Attention Network module that reduces parameters through channel compression. The proposed Adaptive Focal Spatial-IoU loss function further improves detection performance for small galaxies through dynamic focal mechanisms and scale-invariant optimization. Evaluated on Galaxy Zoo 2 data set, our STAR-YOLO achieves 96.3% mean average precision—a 2.5% improvement over baseline models, with irregular galaxy recognition accuracy notably increasing by 9.3%. Comparative experiments demonstrate superior detection capabilities for multi-target irregular galaxies compared to state-of-the-art models, providing an innovative solution for astronomical image analysis.

Key words: methods: data analysis – techniques: image processing – galaxies: irregular – galaxies: peculiar – galaxies: spiral – Galaxies

1. Introduction

Galaxy morphology detection is pivotal in astronomy, as it quantifies structural diversity (e.g., shape, substructure) to decode cosmic evolutionary processes. Beyond classification, it links dynamical histories (e.g., mergers, feedback) to observable features, offering insights into dark matter distribution and large-scale structure formation (Kormendy & Ho 2013). The morphological diversity of galaxies encodes direct evidence of their formation and dynamical history. For instance, tidal tails and filamentary structures in irregular galaxies are unambiguous signatures of recent merger events (Pfeffer et al. 2023). During galaxy interactions, gravitational torques drive gas inflows, triggering starburst activity that redistributes baryonic mass and alters the luminosity profile (Conselice 2006). These morphological imprints not only trace merger-induced dynamical perturbations but also correlate with feedback processes, such as supernova-driven winds, which eject gas and suppress further star formation, leading to asymmetric features in the interstellar medium (Hopkins et al. 2008).

A vast body of research in galaxy morphology has been built upon the paradigm of image classification. This approach, from the seminal Hubble sequence to modern computational methods, has provided the foundational framework for quantifying galactic structures (Abraham et al. 1996). The advent of machine learning (ML) (Ball & Brunner 2010) and, more recently, deep convolutional neural networks has significantly automated and enhanced the objectivity of this classification process. For example, Zhang et al. (2022) proposed a Self-Calibrated Convolutional Network employing few-shot learning, achieving high-precision morphology classification with limited training samples. Tarsitano et al. (2022) demonstrated ML algorithms can successfully distinguish early- and late-type galaxies in images with signal-to-noise ratios exceeding 300. These methods have been instrumental in pioneering the large-scale morphological analysis of galaxies, establishing a crucial baseline for the field.

However, this classification-centric paradigm reveals inherent and critical limitations when confronted with the data deluge from next-generation (Abolfathi et al. 2021), wide-field sky surveys such as LSST (Ivezić et al. 2019) and Euclid (Laureijs et al. 2011). The primary shortcoming is that classification models are designed to assign a single, global label to an entire image. This architecture is not suitable for massive multi-target galaxy images, as it is incapable of locating and identifying multiple discrete objects within a single field of view—a common scenario in crowded stellar and galactic fields. Furthermore, by collapsing an image into a single prediction, these models discard all crucial spatial information (Baron 2019), such as the precise celestial coordinates and structural extent of each object. This loss of spatial data is prohibitive for a vast array of astrophysical inquiries, including analyzing galaxy cluster dynamics,

identifying tidal interactions through the morphology and placement of debris, and mapping dark matter distribution via weak gravitational lensing techniques (Mandelbaum 2018).

To overcome these fundamental limitations of the classification paradigm, the astronomical community increasingly requires robust tools for object detection—a task that inherently involves both precise localization (with bounding boxes) and simultaneous classification of all objects of interest within an image (Krizhevsky et al. 2017). This capability transcends mere cataloging. A proficient detection framework enables the analysis of spatial correlations and the large-scale structure of the universe (Mandelbaum 2018), the identification of merger remnants through the precise segmentation of tidal tails and bridges, and the statistical study of dwarf galaxy populations in the outskirts of dark matter halos—all studies that are predicated on knowing not only what objects are, but also where they are (Gharat & Dandawate 2022).

Foundational frameworks from computer vision, such as the one-stage YOLO series (Redmon et al. 2016) and RetinaNet (Lin et al. 2017), or two-stage detectors like Faster R-CNN (Ren et al. 2015), have set the standard for object detection in natural images. Despite their success, directly applying these general-purpose detectors to astronomical images is non-trivial, as they face several domain-specific challenges (Ohnaka & Morales 2018):

(1) *Extreme Multi-scale Variability*. The vast difference in apparent size between large, nearby galaxies and distant, compact dwarfs or stars demands exceptional multi-scale feature extraction capabilities.

(2) *Crowded and Noisy Environments*. In dense galactic fields or low signal-to-noise regimes, models struggle with overlapping targets and faint morphological features, leading to missed detections (low recall) for irregular galaxies and merger-induced structures.

(3) *Sensitivity and Localization of Small Targets*. The progressive downsampling in standard CNNs aggressively discards spatial information vital for localizing subarcsecond targets (e.g., dwarf galaxies, distant compact systems), rendering them nearly invisible to the network.

(4) *Computational Inefficiency*. The high resolution of survey images and the sheer data volume render many sophisticated models too slow for real-time processing, creating a bottleneck.

Notably, Gu et al. achieved a high detection accuracy of 93% by applying a mask-based Mask R-CNN framework for four-type galaxy detection (Gu et al. 2023). Despite this impressive performance, their approach, along with others, continues to exhibit limited feature extraction capabilities for irregular galaxies and small-target sources—a performance gap that our work aims to fill.

To address these challenges, we propose STAR-YOLO, an optimized framework for feature extraction and localization precision, designed to enhance galaxy dynamics studies and dark matter distribution modeling. Our model incorporates innovative backbone network architectures and attention mechanisms to minimize information loss while improving recognition rates for irregular galaxies and small-target galaxies in wide-field observations. The principal contributions include:

(1) *Multi-scale Attentive Context Aggregation (MACA)*. Integrates progressive spatial-channel attention with multi-scale dilated convolutions, employing variable dilation rates to adapt to target scale variations while suppressing noise and emphasizing critical regions.

(2) *Lightweight Efficient Attention Network (LEANet)*. Replaces standard convolutions in C3 layers with partial convolutions (PConv), achieving 96.3% mAP@0.5 without significant parameter inflation through selective channel computation.

(3) *Adaptive Focal Spatial-IoU (AFS-IoU) Loss*. Supersedes traditional CIoU with angle-sensitive penalty terms and dynamic gradient adjustment for hard samples, enhancing localization accuracy for irregular and small-target galaxies.

(4) *Real-time Processing Solution*. Provides an efficient framework compatible with next-generation surveys like LSST, addressing critical throughput requirements for large-scale astronomical data analysis.

The remainder of this paper is organized as follows. Section 2 describes the data set and the morphology classes used in our study. Section 3 details the architecture of our proposed STAR-YOLO framework and its core components. Section 4 outlines the experimental setup, including implementation details and evaluation metrics. Section 5 presents and discusses the experimental results, including comparisons with state-of-the-art methods and ablation studies. Finally, Section 6 concludes the paper and suggests directions for future work.

2. Data Sets

In this study, experiments were conducted using the Galaxy Zoo 2 data set (Willett et al. 2013), a large-scale volunteer classification project derived from the Kaggle data set platform containing a sample of 245,609 galaxies from SDSS DR7. Users can download the classification table from the official Galaxy Zoo 2 release page, containing morphological labels for each galaxy. We first selected high-confidence samples by filtering galaxies with a debiased probability greater than 0.8. The bounding boxes for the central regions of these candidate galaxies were then manually annotated using LabelImg software to create the ground truth for our object detection task.

By collating the data set, a total of 3600 images of galaxies were processed and classified into six categories: barred spiral galaxies, elliptical galaxies, spiral galaxies, irregular galaxies, merging galaxies, and stars, as shown in Table 1. Barred spiral galaxies are characterized by their central bar structure and the spiral arms extending from the ends of the bar structure, while elliptical galaxies exhibit a smooth, symmetrical elliptical or circular appearance, lacking obvious structural

Table 1
Galaxy Morphology Data Set Characteristics with Sample Images

Type of Galaxy	Description	Sample Image
Barred Spiral	Barred spiral galaxies exhibit a prominent linear structure of stars	
Elliptical	Elliptical galaxies, more regular, elliptical-like	
Irregular	Irregular galaxies, with different shapes, often with fuzzy edges	
Merging	Describe the process by which galaxies are merging	
Spiral	Shaped like a spiral, it is easy to confuse elliptical galaxies	
Star	In a large field of view, it is easy to be confused with other galaxies	

features such as spiral arms or bar structure. Spiral galaxies are marked by their well-defined spiral arms and central nucleus, which usually has a spherical nucleus at the center. Irregular galaxies have no obvious axis of symmetry, neither rotational symmetry like spiral galaxies nor spherical symmetry like elliptical galaxies, and they usually do not have well-defined structures such as spiral arms, nuclei spheres, or dust bands. Unlike the classical Hubble classification system for galaxy morphology, the merging galaxies included in the data set are a dynamic process in which two or more galaxies gravitationally approach each other and eventually merge. This process can lead to significant changes in the morphology of galaxies, making them potentially irregular in appearance. Merging galaxies exhibit a complex process of galaxies colliding and merging with each other, often accompanied by tidal tails and bridge-like structures. Peculiar galaxies, on the other hand, do not fit the standard classification due to their unique appearance and formation history, but are distinguished from irregular galaxies by having very unusual morphological features, such as extreme shapes, complex structures, or unique features resulting from interactions with other galaxies or

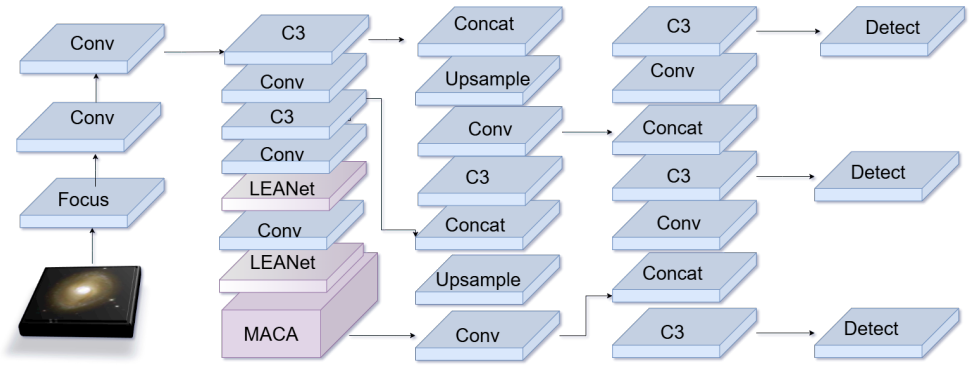

Figure 1. STAR-YOLO model structure.

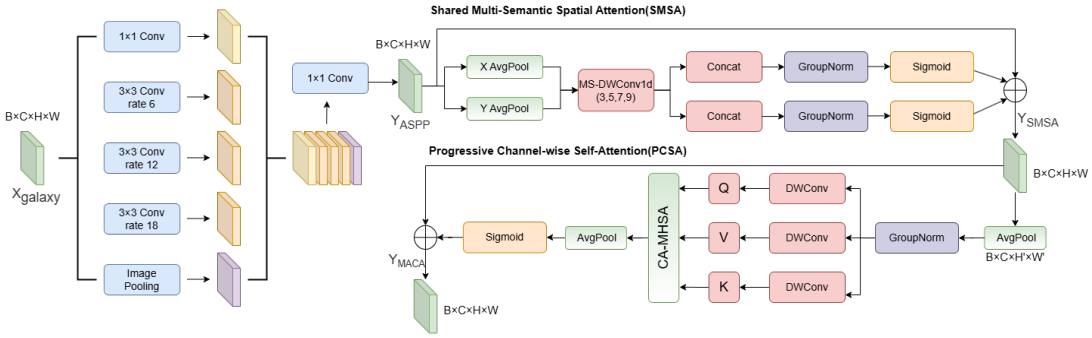

Figure 2. MACA model structure.

cosmic phenomena, and are galaxies that are difficult to categorize in the regular classification. In addition to galaxies, stars have been included to distinguish them from galaxies, since both galaxies and stars are small-target galaxies under large field-of-view conditions.

In the context of this study, "complex galaxy morphologies" refer specifically to those categories that present significant challenges for automated detection systems due to their non-canonical structural characteristics or physical properties. Our work primarily focuses on two types of complexity (Rodriguez-Gomez et al. 2015): (1) Irregular galaxies exhibiting asymmetric light distributions and fragmented morphological features, which lack defined spiral arms or elliptical symmetry. These systems often display chaotic stellar distributions and irregular gas dynamics, deviating from standard classification frameworks. (2) Small-scale targets such as distant dwarf galaxies or unresolved stellar populations occupying minimal pixel coverage in wide-field imaging. Their low surface brightness and compact angular size complicate automated recognition algorithms.

3. Methodology

3.1. STAR-YOLO Model

Our STAR-YOLO framework is built upon the well-established architecture of yolov5s (Redmon & Farhadi 2018), which provides a robust balance between detection accuracy and inference speed. We selected this architecture as our foundation for several key reasons that align with the demands of astronomical image analysis: (a) Real-time capability: Its one-stage design enables efficient processing of large-volume survey data. (b) Multi-scale prediction: The Path Aggregation Network (PAN) in its neck effectively handles objects of vastly different scales, from large ellipticals to tiny stars. (c) Precise localization: It directly regresses bounding boxes, which is fundamental for astronomical applications requiring positional accuracy. (d) Multi-object handling: It naturally detects all objects in an image simultaneously, a necessity for studying crowded fields.

Facing the massive galaxy images, the STAR-YOLO model proposed in this paper is designed for the task of irregular galaxy and small-target galaxies detection in astronomical images. The overall architecture of our proposed STAR-YOLO framework is illustrated in Figure 1. It follows a mainstream one-stage detector design, consisting of a Backbone for multi-scale feature extraction, a Neck incorporating a Feature Pyramid Network (FPN) and a PAN for feature fusion, and a Detection Head for performing the final predictions.

The input image is first processed by the Backbone network. The data flow begins with a Focus module, followed by a

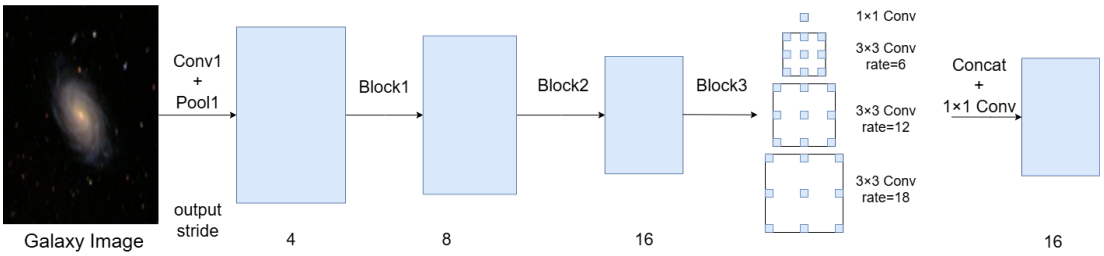

Figure 3. ASPP model structure.

series of convolutional (Conv) and C3 layers to extract hierarchical feature maps at different scales. To significantly enhance the backbone's capability for multi-scale contextual reasoning—which is crucial for capturing both large galaxies and tiny stellar objects—we integrate our novel Multi-scale Attentive Context Aggregation (MACA) module at the terminal stage of the backbone. This module is a cascade of an Atrous Spatial Pyramid Pooling (ASPP; Lian et al. 2021) component for multi-receptive-field feature extraction and a Spatial-Channel Synergistic Attention (SCSA; Slijepcevic et al. 2022) component for calibrating feature responses.

The extracted feature maps are then passed to the Neck. Here, the network employs Upsample and Concatenation operations to construct the top-down FPN path and the bottom-up PAN path, effectively aggregating and enhancing features from different semantic levels. A key innovation within the Neck is our LEANet, which replaces the standard convolutions in selected C3 modules with PConv. This design, clearly shown in the structure graph, strategically reduces computational redundancy and parameter count while effectively preserving the model's representational power.

Finally, the multi-scale fused features from the Neck are fed into the Detection Head. To address the specific challenges of astronomical detection, particularly the localization of small and irregular targets, we propose the AFS-IoU loss function for training. This loss supersedes traditional IoU variants by incorporating dynamic scale weights and a focal mechanism, which drastically improves localization accuracy for the most challenging categories in our data set.

3.2. Multi-scale Contextual Attention (MACA)

Because irregular galaxies have complex, asymmetric structures, conventional convolution is difficult to effectively capture their multi-scale features, small-target galaxies in the image with low resolution and poor signal-to-noise ratio, while the conventional attention module fails to effectively extract the multi-scale features of galaxies, and it is difficult to balance the local details as well as the global semantics of the irregular galaxies, and the original spatial pyramid pooling lacks the adaptive attention to the fine-grained structures. The original spatial pyramid pooling lacks adaptive attention to the fine-grained structure, and the pooling operation may lose the details of small-target galaxies. Therefore, in this paper, we design the MACA (Multi-scale Contextual Attention Mechanism), which is a deep fusion of the multi-scale cavity convolution and the progressive spatial channel attention mechanism, to cover the scale variations of the galaxy targets through the cavity convolution with different expansion rates, and at the same time, based on the spatial channel synergistic attention to dynamically calibrate the multi-scale feature weights, to suppress the noise and to focus on the critical regions of the galaxy. The MACA module captures galaxy core regions, spiral arms, and debris structures through multi-scale cavity convolution. This design directly targets the asymmetric morphology of irregular galaxies (e.g., the fibrous structure of the M82 starburst galaxy), and its multi-sense-field feature fusion effectively suppresses background noise (e.g., interference from stellar-dense regions), which improves the sensitivity to dynamical features, such as tidal tails, and provides more accurate morphology data for the study of the merger process of galaxies. The structure of the MACA model is shown in Figure 2. The left half is the multi-scale cavity convolution, and the right half is the spatial and channel cooperative attention module. Cavity convolution is a special convolution operation that expands the sensory field without increasing the number of parameters and computational complexity by introducing cavities between the convolution kernels. The network diagram of the ASPP module is shown in Figure 3, where the output of Block3 is input to the ASPP, which undergoes a pooling operation after sampling by multi-scale cavity convolution, and then the number of channels is reduced by the 1×1 convolution to the expected value. The input feature map XGalaxy dimension is $B \times C \times H \times W$, where B represents the batch size, C represents the number of channels, and H and W represent the height and width of the feature map, respectively. The MACA model first takes the input feature map of galaxies, XGalaxy, by convolving the voids with four different expansion rates in parallel, which can efficiently capture contextual information on different scales without changing the resolution of the galaxy feature maps. In order to capture the small-target galaxies more efficiently, in this paper, we use the 1×1, 3×3 (rate $= 6$), 3×3 (rate $= 12$), 3×3 (rate $= 18$) four expansion rates, and then,

the spatial dimension of the feature map is compressed to 1×1 by pooling operation to obtain the global information. The branch feature extraction by null convolution and global pooling with different expansion rates is shown in Equations (1) and (2):

$$\boldsymbol{F}_{\text{conv}}(\boldsymbol{X}_{\text{G}}, \gamma) = \text{Conv2D}(\boldsymbol{X}_{\text{G}}), \tag{1}$$

$$\boldsymbol{F}_{\text{gap}}(\boldsymbol{X}_{\text{G}}) = \text{Upsample}[\text{AvgPool}(\boldsymbol{X}_{\text{G}})], \tag{2}$$

where γ is the expansion rate.

Then, the null convolution outputs with different expansion rates and the global average pooled output are fused thereby obtaining the output feature map as shown in Equation (3):

$$\boldsymbol{Y}_{\text{ASPP}} = W_{\text{f}} \cdot \text{Concat}(\boldsymbol{F}_{\gamma_i}, \boldsymbol{F}_{\text{gap}}), \tag{3}$$

where W_{f} denotes the fusion convolution weights, and $\boldsymbol{F}_{\gamma_i}$ represents the convolutional features extracted with the ith dilation rate.

Since the output feature map is also a four-dimensional tensor, it can be used as an input to the subsequent module. The galaxy feature map is input to the SMSA module (Shared Multi-semantic Spatial Attention). SMSA first decomposes the galaxy input feature map along the height and width directions and further divides the feature map in each direction into multiple independent sub-features to efficiently extract the spatial information at different semantic levels. Then, the capturing the spatial structure of each sub-feature using depth-separable 1D convolutions at different scales, the article uses four convolution kernel sizes of 3, 5, 7, and 9, respectively, and uses shared convolutions to align the feature maps in different directions, as shown in Equations (4) and (5):

$$\hat{\boldsymbol{Y}}_{\text{h}}^{i} = \text{DWConv1D}[\text{R}(\boldsymbol{Y}_{\text{ASPPh}})^{i}, k_{\text{i}}], \tag{4}$$

$$\hat{\boldsymbol{Y}}_{\text{w}}^{i} = \text{DWConv1D}[\text{R}(\boldsymbol{Y}_{\text{ASPPw}})^{i}, k_{\text{i}}], \tag{5}$$

where k_{i} denotes the kernel size of the ith sub-feature, and R represents the reshaping function, ensuring the tensor dimensions: $\boldsymbol{Y}_{\text{ASPPh}} \in \mathbb{R}^{\text{B}\times\text{C}\times\text{W}}$, $\boldsymbol{Y}_{\text{ASPPw}} \in \mathbb{R}^{\text{B}\times\text{C}\times\text{H}}$.

SMSA also performs group normalization on each sub-feature to avoid the effect of batch noise and effectively reduce the semantic interference between sub-features. Finally, it splices the sub-features with different semantics and generates the spatial attention graph using Sigmoid activation function as shown in Equations (6), (7) and the mathematical formulation of SMSA is shown in Equation (8):

$$\boldsymbol{A}_{\text{w}} = \sigma\{\text{GN}(\text{Concat}(\hat{\boldsymbol{Y}}_{\text{w}}^{i}))\}, \tag{6}$$

$$\boldsymbol{A}_{\text{h}} = \sigma\{\text{GN}(\text{Concat}(\hat{\boldsymbol{Y}}_{\text{h}}^{i}))\}, \tag{7}$$

$$\boldsymbol{Y}_{\text{SMSA}} = \boldsymbol{A}_{\text{h}} \times \boldsymbol{A}_{\text{w}} \times \boldsymbol{Y}_{\text{ASPP}}, \tag{8}$$

where $\boldsymbol{A}_{\text{h}}$ and $\boldsymbol{A}_{\text{w}}$ denote the spatial attention maps along the height and width, respectively, and σ denotes the Sigmoid normalization, and $K = 4$ is the number of feature groups.

Progressive Channel Self-Attention (PCSA): The feature maps output from SMSA are first compressed using an average pooling operation to reduce the computational effort while preserving the spatial a priori information as shown in Equation (9):

$$\boldsymbol{Y}_{\text{p}} = \text{AvgPool}(\boldsymbol{Y}_{\text{SMSA}}), \tag{9}$$

It then utilizes the compressed feature map for single-head self-attention computation to generate Q, K, and V to explore the similarity between channels and mitigate the semantic differences between different sub-features in SMSA, as shown in Equation (10):

$$\boldsymbol{Q} = \boldsymbol{W}_{\text{Q}} \cdot \boldsymbol{Y}_{\text{p}}, \ \boldsymbol{K} = \boldsymbol{W}_{\text{K}} \cdot \boldsymbol{Y}_{\text{p}}, \ \boldsymbol{V} = \boldsymbol{W}_{\text{V}} \cdot \boldsymbol{Y}_{\text{p}}. \tag{10}$$

Attention weights are calculated weights as shown in Equation (11):

$$A_{\text{channel}} = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\top}}{\sqrt{C}}\right)\boldsymbol{V}, \tag{11}$$

Finally, the constructed MACA is shown in Equation (12):

$$\boldsymbol{Y}_{\text{MACA}} = \boldsymbol{Y}_{\text{SMSA}} \times \sigma(\text{AvgPool}(A_{\text{channel}})), \tag{12}$$

The advantages of the MACA module lie in its multi-semantic guidance and semantic discrepancy mitigation capabilities, as well as its computational efficiency. The multi-scale cavity convolution of ASPP provides contextual information of different sensing fields, and the attention mechanism of SCSA dynamically focuses on the key regions. The two synergistically enhance the feature expression of irregular galaxies and small-targeted galaxies, and then the features are progressively optimized through the gradual compression strategy of SMSA, the spatial structure information is gradually injected into the channel attention to avoid information loss, and the output of SMSA is multiplied with the output of PCSA at the element level to obtain the final MACA output. Then the output of MACA is used as the input of the subsequent layers. Through these operations, the MACA module can help STAR-YOLO learn features better and improve the performance of target detection, especially in complex scenes, such as irregular galaxies and small-target galaxies detection under a large field of view.

3.3. LEANet

The introduction of the MACA module enhances feature extraction but also increases the model's parameter count and computational complexity. To maintain a balance between performance and efficiency, we propose the LEANet module, as shown in Figure 4, whose core is the PConv operation (Chen et al. 2023).

The design of PConv is motivated by a key observation in convolutional neural networks: the feature maps of consecutive channels often exhibit strong redundancy and high correlation . This implies that a significant portion of the

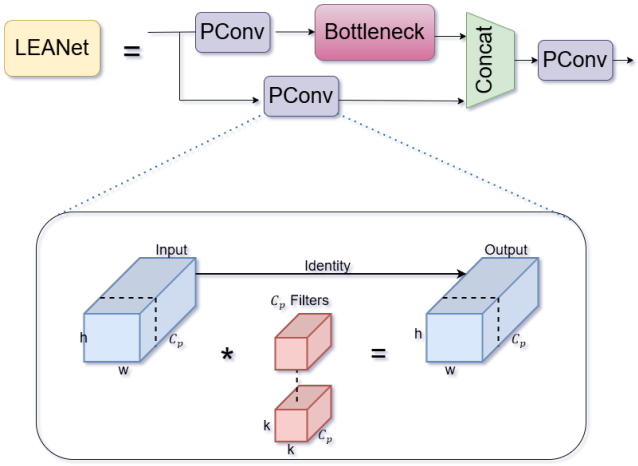

Figure 4. LEANet model structure diagram.

computation in standard convolution, which processes all channels, is repetitive and wasteful. PConv addresses this by strategically processing only a fraction of the input channels while leaving the remaining channels untouched. This approach is theoretically justified because the information from the learned features in the processed channels can be effectively propagated through subsequent pointwise convolutions or other operations, minimizing the loss of representational capacity.

The choice of processing 25% of the channels ($cp = c/4$) is not arbitrary. It follows the established practice *i* and represents a sweet spot empirically identified to achieve maximal computational savings while preserving accuracy. A lower ratio might risk losing critical information, while a higher ratio yields diminishing returns in efficiency gains.

As illustrated in Figure 5, for an input feature map, PConv applies the standard convolution on only a portion of the input channels while leaving the rest of the channels unchanged. The formula for calculating the FLOPs for PConv is shown in Equation (13):

$$F_{\text{PConv}} = h \times w \times k^2 \times c_p^2, \tag{13}$$

where h, w denote the height and width of the input feature map, respectively, k denotes the size of the convolution kernel, and cp denotes the number of channels for the convolution operation.

Since cp is usually much smaller than the number of channels c of the input feature map, the FLOPs of PConv are significantly lower than that of regular convolution, and if we choose to process only $1/4$ channels, the FLOPs are only $1/16$ of that of the regular convolution. Similarly, the computation of one window is negligible, so the memory access will be $1/4$ of the original one when processing $1/4$ channel. The number

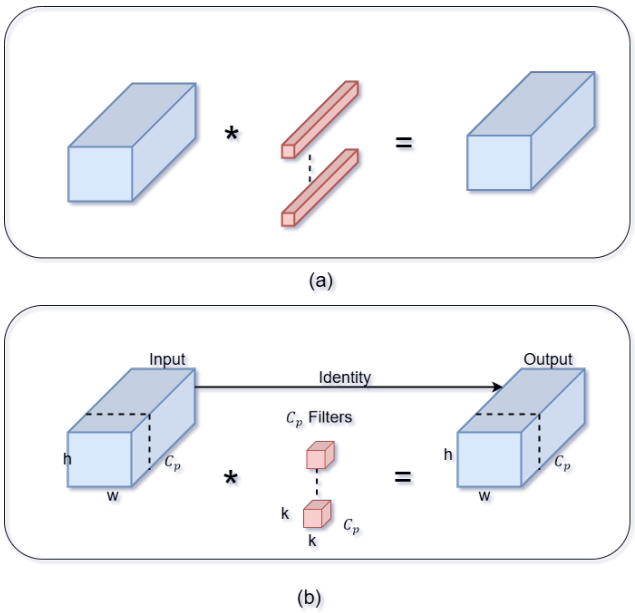

Figure 5. Illustration of different convolutions. (a) Convolution; (b) Partial convolution.

of memory accesses is calculated as shown in Equation (14):

$$M_{\text{PConv}} = h \times w \times 2c_p + k^2 \times c_p^2. \tag{14}$$

In the context of galaxy detection, where features across channels are highly similar due to the nature of astronomical images, LEANet (by integrating PConv) significantly reduces parameters and accelerates inference without compromising the model's accuracy. This makes STAR-YOLO exceptionally suitable for processing large-field images containing complex irregular and small-target galaxies.

3.4. AFS-IoU Loss Function

In the detection of galaxies, the gradient signals of difficult samples, such as fuzzy small-target galaxies and irregular galaxies, are easily flooded by the gradients of a large number of simple samples (high IoU targets) during the training process. This leads to the difficulty for the model to fully learn the features of difficult samples, which in turn affects the detection accuracy. Therefore, a loss function named Adaptive Focal Scale-IoU (AFS-IoU) is designed. The SIoU loss function is a loss function designed specifically for target detection tasks (Gevorgyan 2022). The advantage of the SIoU loss function is that it comprehensively takes into account the bounding box regression and the category prediction, which makes the evaluation of the model performance more comprehensive. The total loss function is expressed as shown in Equation (15).

$$L_{\text{SIoU}} = 1 - \text{IoU} + \frac{\Delta + \Omega}{2}, \tag{15}$$

Here Δ is the distance loss, which measures the distance error between the predicted frame and the real frame in the x, y directions, and Ω is the shape loss, which penalizes the difference between the predicted frame and the real frame in terms of aspect ratio.

However, the SIoU lacks a dedicated mechanism to address class imbalance and the dominance of easy samples during training, which is particularly detrimental for detecting irregular galaxies and hard-to-localize small targets.

To mitigate this, the Focal Loss idea was introduced to IoU variants (Lin et al. 2017), leading to functions like FSIoU. These functions apply a modulating factor to down-weight the loss contribution of easy examples (high IoU) and focus training on hard examples (low IoU).

Our proposed AFS-IoU loss is designed to synergistically combine the strengths of both SIoU and Focal Loss. It achieves this through two key improvements over its predecessors:

(1) *Angle-Sensitive Penalty from SIoU*. AFS-IoU retains the angle cost term from SIoU. This term directly penalizes orientation deviations between the predicted and ground-truth bounding boxes, which is crucial for accurately capturing the elongated and asymmetric shapes of tidal features in merging galaxies and irregular galaxies.

(2) *Enhanced Dynamic Focus Mechanism*. While FSIoU introduces a scale-aware focal term, AFS-IoU enhances this mechanism to be more adaptive and aggressive. The focal factor is calibrated to drastically amplify the gradient signals for extremely hard samples (e.g., low-IoU small targets). Concurrently, the scale-aware weight dynamically adjusts the loss based on the target's size, providing a much stronger learning signal for smaller objects compared to larger ones.

The improved loss function is shown in Equation (16):

$$L_{\mathrm{FSIoU}} = (1 - \mathrm{SIoU})^{\gamma} \cdot \left(1 + \frac{S_{\max}}{S_{\mathrm{target}}}\right) L_{\mathrm{SIoU}}, \tag{16}$$

Here γ is the focusing factor, which controls the weight intensity of difficult samples. Starget is the current target pixel area and Smax is the maximum target area in the image.

In this study, the FSIoU loss function is used instead of the CIoU. AFS-IoU supersedes FSIoU by integrating essential spatial reasoning (angle penalty) and supersedes SIoU by incorporating a dynamic focus on hard samples. This dual advantage leads to superior localization accuracy, especially for the challenging cases of irregular and small-target galaxies that are most affected by orientation bias and class imbalance.

4. Experiments

4.1. Experimental Environment

All the experiments in this paper are run on the same server with Xeon(R) Platinim 8225 C CPU, RTX 4090 GPU. In this paper, the proposed STAR-YOLO is implemented on the framework of Pytorch2.0 and CUDA 12.5. The images in the model are adaptively scaled to 640×640 pixels. Initial learning rate was empirically set to 0.0005 through iterative experiments, the number of epochs is set to 300, the weight decay is set to 0.0005, the batch size is set to 16, and the optimizer uses Adam. To ensure the fairness of the model comparisons, the parameters used in this study are consistent.

4.2. Model Evaluation

In order to comprehensively evaluate the model's detection performance for galaxy images, a set of evaluation metrics were chosen, with precision, recall, FLOPs, and parameters being the key metrics for assessing the model's performance.

Precision is the ratio of the number of actual positive samples to the number of positive samples tested. The formula is shown in Equation (17):

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{17}$$

where P denotes the accuracy, TP denotes the number of positive samples predicted to be in the positive category, and FP denotes the number of negative samples predicted to be in the positive category.

Recall is the proportion of samples that are detected as positive out of all actual positive samples. The formula is shown in Equation (18):

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{18}$$

where R denotes the recall rate and FN denotes the number of positive samples predicted to be in the negative category.

The multi-category average precision (mAP) is the average of the average precision of all categories. It is one of the most important evaluation metrics in target detection algorithms and can be used to indicate the detection accuracy of the target detection model. mAP is calculated as shown in Equation (19):

$$\mathrm{mAP} = \frac{\sum \mathrm{AP}}{N}, \tag{19}$$

where N is the number of target categories.

In addition, the loss value is an important indicator of the difference between the model prediction and the actual value, including the training loss and the validation loss, which reflect the model's ability to adapt and generalize over the data set, respectively. Floating point operation (FLOP) is a widely used metric in resource efficient modules. In this study, the number of FLOPs refers to the number of floating-point operations and is used to measure the complexity of the algorithm or model.

5. Results

As clearly demonstrated in Table 2, STAR-YOLO achieves state-of-the-art performance specifically on the most challenging categories: irregular galaxies and small-target galaxies (stars). For

Table 2
Comparison Experiment of Different Models

No	Model	Precision	Map@0.5	Bar	Ell	Mer	Spi	Irr	Star	Par/M
1	Yolov5	0.903	0.938	0.952	0.96	0.705	0.925	0.782	0.853	7.04
2	Yolov6	0.889	0.864	0.899	0.904	0.699	0.892	0.722	0.775	4.23
3	Yolov8	0.912	0.905	0.931	0.927	0.828	0.944	0.763	0.832	6.46
4	Yolov9	0.691	0.858	0.887	0.863	0.642	0.865	0.732	0.757	2.01
5	Yolov10	0.866	0.93	0.931	0.945	0.839	0.955	0.77	0.842	2.71
6	EfficientNetV2	0.892	…	0.935	0.948	0.783	0.996	0.821	0.873	…
7	ResNet-26	0.92	…	0.991	0.993	0.805	0.987	0.81	0.792	…
8	RTDETR	0.854	0.804	0.932	0.918	0.704	0.904	0.68	0.702	…
9	Mask-R-CNN	0.918	0.932	0.977	0.968	0.782	0.919	0.81	0.852	…
10	EfficientDet-D2	0.879	0.901	0.938	0.945	0.795	0.932	0.805	0.86	8.3
11	Swin Transformer	0.898	0.957	0.992	0.99	0.965	0.983	0.84	0.882	20.2
12	CenterNet++	0.872	0.923	0.915	0.928	0.745	0.911	0.758	0.79	…
13	STAR-YOLO	**0.914**	**0.963**	0.989	**0.994**	**0.879**	0.988	**0.875**	**0.897**	7.06

Table 3
Comparative Tests of Different Attention Mechanisms

No	Model	Precision	Recall	Map@0.5	Irregular	Star	Map@0.5:0.95	FLOP/G
1	Yolov5-RFEM	0.937	0.913	0.947	0.802	0.855	0.717	15.8
2	Yolov5-CBAM	0.920	0.900	0.929	0.814	0.843	0.703	15.8
3	Yolov5-SE	0.908	0.878	0.935	0.760	0.859	0.688	15.8
4	Yolov5-CA	0.927	0.917	0.933	0.793	0.838	0.713	15.8
5	Yolov5-SCSA	**0.914**	**0.929**	**0.963**	**0.875**	**0.897**	**0.713**	16.0

irregular galaxies, STAR-YOLO attains an mAP@0.5 of 87.5%, which represents a significant improvement of 9.3% over the yolov5 baseline and also outperforms other strong competitors like EfficientNetV2 (82.1%), Mask R-CNN (81.0%), RTDETR (68.0%) (Zhao et al. 2024) and notably, the heavyweight Swin Transformer (84.0%) (Liu et al. 2021). The bold values indicate the best performance for each evaluation metric. This superior performance underscores the effectiveness of the MACA module, particularly its spatial-channel attention mechanism, in capturing the faint and fragmented morphological features inherent to irregular systems.

For small-target detection, crucial for large-field surveys, STAR-YOLO achieves a leading mAP@0.5 of 89.7% on stars, surpassing all other YOLO (such as yolov7 (Wang et al. 2023) and yolov10 (Wang et al. 2024)) variants and EfficientDet-D2 (86.0%) (Tan et al. 2020). It also holds a slight edge over the computationally expensive Swin Transformer (88.2%). Most importantly, these advancements are achieved without a substantial increase in model complexity. With merely 7.06M parameters, STAR-YOLO is significantly lighter than Swin Transformer (20.2M) and maintains high efficiency comparable to yolov5 (7.04M). This exceptional balance between accuracy and efficiency makes STAR-YOLO uniquely suited for processing high-volume data from next-generation sky surveys like LSST, enabling large-scale statistical studies of irregular galaxies and faint, distant objects.

In the selection of attention mechanisms, this study also compared the experimental results of integrating different attention mechanisms into yolov5, as shown in Table 3. The tested mechanisms included RFEM, CBAM, SE, CA, and SCSA. Notably, the incorporation of the SCSA mechanism achieved an mAP@0.5 of 0.953, representing a 1.5% improvement over the baseline model. Additionally, the detection accuracy for irregular galaxies and stars was significantly enhanced. Specifically, the accuracy for irregular galaxies increased from 78.2% to 86.8%, a gain of 8.6%. The bold values indicate the best performance for each evaluation metric. These results demonstrate that SCSA effectively mitigates information diffusion, enhances feature extraction capabilities, and strengthens the model's ability to focus on critical regions of irregular targets.

In order to verify the effectiveness of each improved module in this experiment, several ablation experiments were carried out using this galaxy data set, and the ablation experiments are shown in Table 4. The bold values indicate the best performance for each evaluation metric. In the ablation experiments, it can be seen that the average model accuracy and the irregular galaxy detection accuracy were improved after the introduction of the SCSA attentional mechanism, which is due to the spatial channel synergistic attentional

Table 4
Table of Ablation Experiments

SCSA	AIFI	LEANet	SIoU	*P*	Recall	mAP@0.5	Irr	Star	Params/M	FLOPs/G
…	…	…	…	0.903	0.914	0.938	0.782	0.853	7.04	15.8
✓	…	…	…	0.916	0.944	0.953	0.868	0.862	8.32	17.2
…	✓	…	…	0.916	0.936	0.961	0.835	0.89	7.23	16.3
…	…	✓	…	0.889	0.926	0.945	0.802	0.859	6.39	14.8
…	…	…	✓	0.922	0.938	0.952	0.801	0.872	7.05	16.0
…	✓	✓	…	0.943	0.919	0.945	0.802	0.859	7.05	16.0
✓	…	✓	…	0.942	0.936	0.953	0.868	0.862	7.82	16.8
✓	✓	…	…	0.916	0.94	0.959	0.868	0.891	8.85	16.3
✓	✓	✓	…	0.943	0.919	0.959	0.87	0.897	7.05	16.0
✓	✓	✓	✓	**0.914**	**0.929**	**0.963**	**0.875**	**0.897**	7.06	16.0

mechanism focusing on the fragmented regions and highlighting the irregular galaxy. The average accuracy of the model is improved by 2.3% and the detection accuracy of stars is improved by 3.7% after the ASPP module is added separately, which is due to the fact that the ASPP module extracts multiple sensory field features in parallel through the convolution of voids with different expansion rates and global average pooling, solves the problem of the target scale change, and avoids the loss of the spatial information caused by downsampling, which in turn optimizes the small-target galaxies. In order to make the model meet the high accuracy of irregular galaxies while focusing on the features of small-target galaxies, the above modules are combined, and it can be seen that the combined MACA module significantly improves the model's ability to recognize irregular and small-target galaxies by 8.6% and 3.8%, respectively, but at the same time, the number of parameters also rises to a certain extent. In order to make the model satisfy the accuracy while making the model keep a small number of parameters, LEANet is continued to be introduced, and it can be seen that the number of parameters of the model is significantly improved after the introduction of LEANet, which is due to the replacement of several C3 modules by PConv in this module, which makes the model parameters decrease from 7.04M to 5.22M. In addition, the experimental data of SCSA and ASPP respectively in combination with the experimental data of the combination of LEANet also both demonstrate its efficient design in reducing the number of parameters. In order to further improve the performance of the model and describe the regression of the target frame more efficiently, as well as to solve the problem of sample imbalance, the AFS-IoU loss function is used to replace the original CIoU loss function, for the asymmetric structure of irregular galaxies, the angular penalty term reduces the bounding box offset due to the orientation bias, and for the small-targeted galaxies, the shape-matching mechanism mitigates the traditional IoU due to the scale sensitivity of the False detection. Meanwhile, AFS-IoU does not significantly increase the computational volume, which

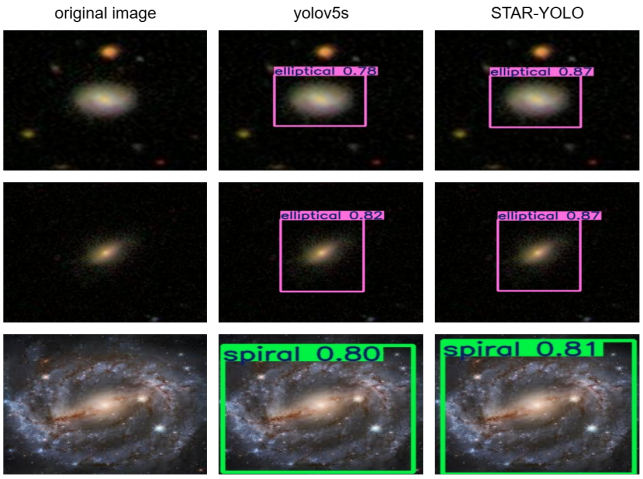

Figure 6. Comparison of the detection results of regular galaxies under the STAR-YOLO and yolov5 baseline modeling algorithms.

complements the lightweight design of LEANet. Compared to the baseline model, the final STAR-YOLO achieved an improvement of 1.1% in overall accuracy, 2.5% in mAP@0.5, 9.3% in detection accuracy for irregular galaxies, and 4.4% in detection accuracy for stars, while the parameter count increased by only 0.2M.

To address the characteristics of the data set, in the qualitative analysis this paper divides the galaxy images into three scenarios, namely, regular galaxies, irregular galaxies, and small-target galaxies and stars under the large view, and conducts prediction experiments on the three scenarios using the proposed STAR-YOLO model and the yolov5 baseline model, respectively, and then performs a qualitative analysis on the obtained visualization results. The first column in the figure is the original image, the second column is the detection effect of the yolov5 baseline model, and the third column is the detection effect of STAR-YOLO. It can be seen that the detection of regular galaxy images, as shown in Figure 6, is due to their more obvious features, but there is the problem of noise. Through the spatial channel synergetic attention to suppress the background noise and

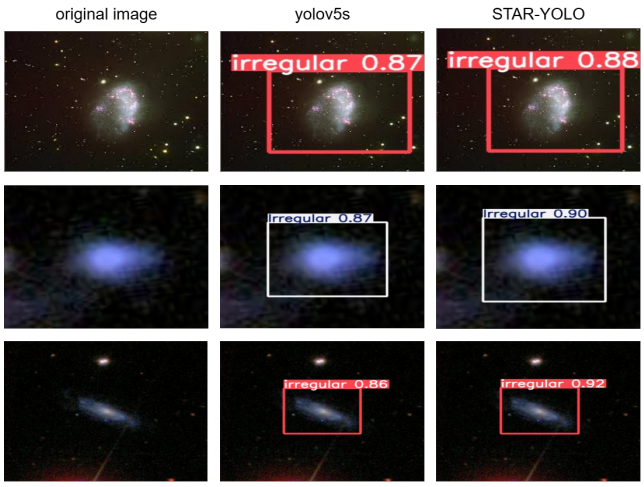

Figure 7. Comparison of irregular galaxy detection results under STAR-YOLO and yolov5 baseline model algorithms.

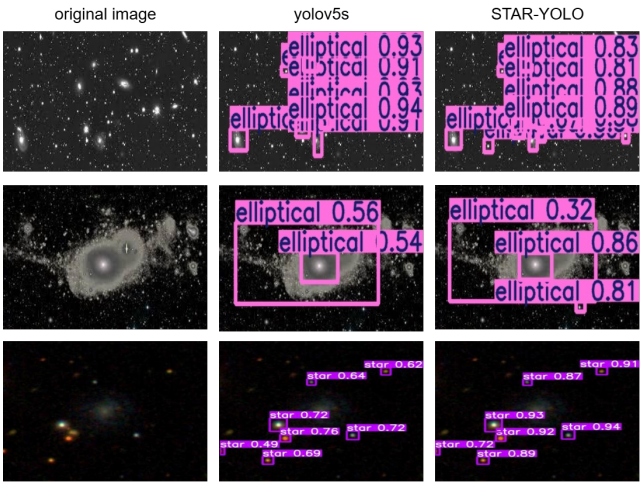

Figure 8. Comparison of detection results of small-target galaxies under the STAR-YOLO and yolov5 baseline model algorithms.

enhance the core region of galaxies accordingly, it can be seen that the detection effect of STAR-YOLO compared to the yolov5 algorithm has a certain degree of improvement.

In irregular galaxies, due to the complexity of their features, they often have irregular features such as fuzzy edges or prominent corners, as shown in Figure 7. It can be seen that the first set of maps has no symmetric structure and the overall brightness distribution is uneven, and there are breaks on the edges, and the second set of maps has fuzzy edge corners due to the low signal-to-noise ratio, which is easy to confuse with elliptical galaxies, and there will be false detections as well as a low accuracy rate in the detection, and it can be seen that the STAR—YOLO model has a better ability to focus on the key regions of irregular targets in the detection of irregular galaxies and can obtain a higher confidence score.

In the background of the large field of view, various galaxies become small-target galaxies, as shown in Figure 8, in the quantitative analysis, due to the lack of classification of small-target galaxies, stars are used as their analogs, here, in order to verify the model's ability to detect small targets, galaxies are detected at the same time as stars to be compared. In the large field of view, the galaxy morphology is small, and due to the low pixel percentage and poor signal-to-noise ratio, the shallow features lose details and the deep features have insufficient semantic information. From the detection effect, it can be seen that yolov5 also has the phenomenon of missed detection, as shown in the second set of maps in Figure 8. In the detection effect on stars, its confidence level is generally improved, and the detection effect of STAR-YOLO on small targets is significantly better than the yolov5 baseline model.

6. Discussion and Conclusions

In this study, we propose the STAR-YOLO model for object detection in galaxy images. Compared to classification models, STAR-YOLO enables real-time monitoring of multiple targets, making it applicable to a broader range of scenarios. To improve detection accuracy for galaxy morphologies—particularly for irregular galaxies and small-target galaxies under large-field conditions—we enhanced the feature extraction capability by introducing the MACA module into the backbone network. The ASPP component within MACA captures multi-receptive-field features, addressing scale variation issues and mitigating spatial information loss caused by downsampling, thereby improving precision in detecting small-target galaxies. Concurrently, the SCSA mechanism focuses on fragmented regions, highlighting key areas of irregular galaxies, which elevates both the overall mean average precision (mAP@0.5) and detection accuracy for irregular galaxies. However, accuracy improvements often come at the cost of increased parameter complexity. To address this, we designed the LEANet module, replacing conventional Convolution-BatchNorm-Silu layers with lightweight PConv, effectively reducing the parameter count while maintaining accuracy. Finally, we adopted the AFS-IoU loss function to further refine localization precision for irregular and small-target galaxies.

The final STAR-YOLO achieved an accuracy of 91.4%, recall of 92.9%, and mAP@0.5 of 96.3%, representing improvements of 1.1%, 1.5%, and 2.5% over the baseline, respectively. Notably, the mAP@0.5 for irregular galaxies and stars increased by 9.3% and 4.4%, respectively, while the parameter count remained nearly unchanged. Experimental results demonstrate that STAR-YOLO outperforms the yolov5 baseline in both accuracy and parameter efficiency. With its lightweight design (7.06M parameters) and real-time inference capability (16 GFLOPs), STAR-YOLO can be deployed in real-time data processing pipelines for large-scale sky surveys such as LSST, processing over 50 frames per second for 2048×2048 pixel images. LSST is expected to generate 20 TB of data per day, and STAR-YOLO's lightweight design

(7.06M parameters) and real-time inference capability (50 FPS) can efficiently process massive images to help the scientific goal of the Dark Energy Survey. By significantly improving detection efficiency for irregular galaxies and merger remnants, STAR-YOLO enables large-scale statistical analyses of galaxy evolution and dark matter halo properties. STAR-YOLO provides an efficient tool for studying the history of galactic mergers, dark matter distribution and early cosmic galaxy formation, filling the technological gap in the detection of complex morphology by traditional methods. In future work, we plan to integrate SDSS spectroscopic data to explore correlations between morphological features and physical parameters (e.g., stellar mass, metallicity).

References

Abolfathi, B., Alonso, D., Armstrong, R., et al. 2021, ApJS, 253, 31
Abraham, R. G., Tanvir, N. R., Santiago, B. X., et al. 1996, MNRAS, 279, L47
Ball, N. M., & Brunner, R. J. 2010, IJMPD, 19, 1049
Baron, D. 2019, arXiv:1904.07248
Chen, J. R., Kao, S. H., He, H., et al. 2023, in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) (Piscataway, NJ: IEEE), 12021
Conselice, C. J. 2006, MNRAS, 373, 1389
Gevorgyan, Z. 2022, arXiv:2205.12740
Gharat, S., & Dandawate, Y. 2022, MNRAS, 511, 5120
Gu, M., Wang, F., Hu, T., & Yu, S. 2023, in 4th Int. Conf. Computer Engineering and Application (ICCEA) (Piscataway, NJ: IEEE), 512
Hopkins, P. F., Cox, T. J., Kereš, D., & Hernquist, L. 2008, ApJS, 175, 390
Ivezić, Z., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111
Kormendy, J., & Ho, L. C. 2013, ARA&A, 51, 511
Krizhevsky, A., Sutskever, I., & Hinton, G. E. 2017, CACM, 60, 84
Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv:1110.3193
Lian, X. H., Pang, Y. W., Han, J. G., & Pan, J. 2021, PatRe, 110, 107622
Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. 2017, in Proc. IEEE Int. Conf. on Computer Vision (Piscataway, NJ: IEEE), 2980
Liu, Z., Lin, Y., Cao, Y., et al. 2021, in Proc. IEEE/CVF Int. Conf. on Computer Vision (Piscataway, NJ: IEEE), 10012
Mandelbaum, R. 2018, ARA&A, 56, 393
Ohnaka, K., & Morales, C. A. L. 2018, A&A, 620, A23
Pfeffer, J., Cavanagh, M. K., Bekki, K., et al. 2023, MNRAS, 518, 5260
Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. 2016, in Proc. IEEE Conf. on Computer Vision and Pattern Recognition, ed. Eds. IEEE (Piscataway, NJ: IEEE), 779
Redmon, J., & Farhadi, A. 2018, arXiv:1804.02767
Ren, S., He, K., Girshick, R., & Sun, J. 2015, in Advances in Neural Information Processing Systems, 28, (Curran Associates, Inc.), 91
Rodriguez-Gomez, V., Genel, S., Vogelsberger, M., et al. 2015, MNRAS, 449, 49
Slijepcevic, I. V., Scaife, A. M. M., Walmsley, M., et al. 2022, MNRAS, 514, 2599
Tan, M., Pang, R., & Le, Q. V. 2020, in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (Piscataway, NJ: IEEE), 10781
Tarsitano, F., Bruderer, C., Schawinski, K., & Hartley, W. G. 2022, MNRAS, 511, 3330
Wang, A., Chen, H., Liu, L., et al. 2024, in Advances in Neural Information Processing Systems, 37, ed. A. Globerson, L. Mackey, D. Belgrave et al. (Curran Associates, Inc.), 107984
Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. 2023, in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (Piscataway, NJ: IEEE), 7464
Willett, K. W., Lintott, C. J., Bamford, S. P., et al. 2013, MNRAS, 435, 2835
Zhang, Z. R., Zou, Z. Q., Li, N., & Chen, Y. L. 2022, RAA, 22, 055002
Zhao, Y., Lv, W., Xu, S., et al. 2024, in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (Piscataway, NJ: IEEE), 16965