
STAR: A Benchmark for Astronomical Star Fields Super-Resolution

Guocheng Wu^{1*} Guohang Zhuang^{1,2*}

Jinyang Huang² Xiang Zhang³ Wanli Ouyang^{1,4} Yan Lu^{1,4†}

¹ Shanghai Artificial Intelligence Laboratory ² Hefei University of Technology

³ University of Science and Technology of China ⁴ The Chinese University of Hong Kong

Abstract

Super-resolution (SR) advances astronomical imaging by enabling cost-effective high-resolution capture, crucial for detecting faraway celestial objects and precise structural analysis. However, existing datasets for astronomical SR (ASR) exhibit three critical limitations: flux inconsistency, object-crop setting, and insufficient data diversity, significantly impeding ASR development. We propose **STAR**, a large-scale astronomical SR dataset containing 54,738 flux-consistent star field image pairs covering wide celestial regions. These pairs combine Hubble Space Telescope high-resolution observations with physically faithful low-resolution counterparts generated through a flux-preserving data generation pipeline, enabling systematic development of field-level ASR models. To further empower the ASR community, **STAR** provides a novel Flux Error (FE) to evaluate SR models in physical view. Leveraging this benchmark, we propose a Flux-Invariant Super Resolution (FISR) model that could accurately infer the flux-consistent high-resolution images from input photometry, suppressing several SR state-of-the-art methods by 24.84% on a novel designed flux consistency metric, showing the priority of our method for astrophysics. Extensive experiments demonstrate the effectiveness of our proposed method and the value of our dataset. Code and models are available at <https://github.com/GuoCheng12/STAR>

1 Introduction

Image quality is critical to astronomical observation, while high quality means finer astrophysical structures and enables precise measurements [1, 2, 3]. This results in the astronomy community always establishing new telescopes to seek high-quality and high-resolution surveys, even facing high costs [4, 5]. Different from astronomy, in natural image processing, the software computer vision Super Resolution (SR) technique [6, 7, 8] has provided a series of successful methods to achieve high-quality and high-resolution observations in an economical way [9, 10]. So, there is obviously an opportunity to introduce the computer vision SR method to process high-quality astronomical images. However, there remains a great challenge – data.

Existing datasets [11, 12] in astronomical super resolution (ASR) have 3 drawbacks: physically trivial, object-centric, and limited-scale. 1). **Flux Inconsistency**: In the real world, telescopes under different observation resolutions have a flux consistency relation [13, 14]. Specifically, although a celestial object has different levels of distortions under low resolutions, it almost has the same total flux as in high resolutions because of the telescope imaging principle [13, 15]. However, existing datasets have significant drifts from this property because they directly use simple interpolation [16, 17], suitable for natural images but conflicts with astronomical observations. This catastrophic limitation

*Equal contribution.

†Corresponding authors. Email: luyan@pjlab.org.cn

makes existing datasets almost physically trivial, significantly affecting their scientific value. 2). **Object-Crop Configuration:** Each image in existing ASR datasets only contains a center-cropped and resized singular celestial object (e.g., stars or galaxies) [16, 18]. This ideal configuration neglects many valuable patterns beyond single object, important in astrophysics, such as large-scale structure [19], cross-object interaction [20], and weak lensing [21, 22], limiting the value of existing datasets. 3). **Insufficient Data Diversity:** The scale of existing ASR datasets ranges from 1,597 to 17,000 [11, 12, 16, 17, 18, 23]. The restricted scale limits the ability of the learned model and makes evaluation unreliable and unfair. To address the above-mentioned dataset limitations, we introduce a new dataset called **STAR**.

STAR is a **large-Scale** ASR dataset. It consists of 54,738 high-resolution star field images captured by the Hubble Space Telescope (HST) [24]. Each image is totally **field-level**, covering a large range of star fields and average containing 30 objects and complex scenarios including multiple celestial objects, cross-object interaction and weak lensing phenomenon, as Figure 1 shows. Compared with existing ASR datasets, STAR provides approximately at least 15 times more observation objects per image on average, while also offering 60% of cosmic information outside the object area (e.g, like diffuse interstellar medium (ISM) regions [25]), significantly showing the scale priority. We provide overall advantages of the STAR for other datasets in Tab. 1.

Except that, to tackle the ‘Physical trivial’ problem, STAR proposes a **flux-consistent** data generation pipeline, which processes cross-resolution image pairs fitting the aforementioned real telescope flux-consistent property, making the entire dataset physically faithful. Furthermore, STAR provides a novel Flux Error (FE) to evaluate SR models from a physical perspective, ensuring their outputs align with astrophysical principles critical for reliable scientific analysis.

With the STAR, we evaluate several state-of-the-art SR methods, including both natural [8, 26, 27, 28, 29, 30] and astronomical SR methods [11, 18] to quantify their generalization ability to the field-level ASR topic, noting that many astronomical SR methods directly adopt natural SR methods. Unfortunately, they cannot provide satisfactory results. We analyze that the main reason is the lack of specific optimization for the flux-consistency prior. Due to this, we propose a novel field-level ASR model, Flux-Invariant Super Resolution (FISR). It introduces the flux consistency property at both the model design and optimization views to fulfill the flux relationships neglected by previous ASR works. At the model view, FISR has a series of specific designs to extract flux information from low-resolution input as visual prompts following astrophysical ideas. These prompts are then injected into the model and give the ability to perceive input flux accurately, allowing the model to propagate consistent flux cues from low-resolution inputs to predicted high-resolution outputs. And at the optimization view, we provide a Flux consistency loss (FCL) which constrains the photometry gap for each celestial object between the ground-truths and predictions, highlighting the importance of flux during the model optimization process and leading to a more reliable trained model.

- **STAR Benchmark:** We introduce STAR, a large-scale, flux-consistent ASR benchmark with 54,738 cross-resolution image pairs from HST F814W star fields. Unlike prior datasets, STAR captures field-level complexity, offering 15 times more objects per image and 60% additional cosmic information, using a flux-consistent pipeline.
- **Flux Error (FE):** We present FE, a novel metric to evaluate SR models’ alignment with astrophysical flux conservation, ensuring reliable photometric analysis.
- **Flux-Invariant Super Resolution (FISR) Model:** We propose FISR, a field-level ASR model and a Flux Consistency Loss, outperforming existing methods by addressing flux relationships neglected in prior work.

2 Related Works

2.1 Super-Resolution Techniques

Super-resolution (SR) techniques, aimed at reconstructing high-resolution (HR) images from low-resolution (LR) inputs, have significantly advanced, offering critical tools for enhancing astronomical images. Traditional SR methods, including interpolation, deconvolution, and learning-based approaches like sparse representation [31], modeled image degradation or statistical relationships to recover details, laying the foundation for SR applications. The advent of deep learning revolutionized

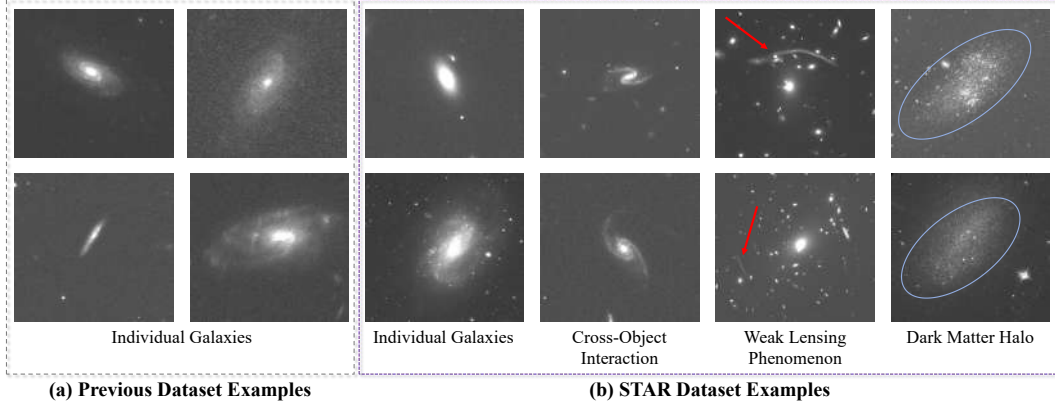


Figure 1: Comparison of previous datasets and ours, highlighting richer structures such as cross-object interaction, weak lensing, and dark matter halos.

Table 1: Comparison of existing astronomical SR datasets.

Dataset	Size	Type	Downsampling	Multiple Celestial	Flux Consistency
AstroSR [11]	2000	Galaxy	×	×	×
QQ Shan et al [12]	9383	Galaxy	×	×	×
W Song et al [16]	1597	Solar	✓	×	×
DiffLense [17]	2880	Galaxy	×	×	×
ZJ Luo et al [18]	17000	Galaxy	×	×	×
WJ Li et al [23]	14604	Galaxy	✓	×	×
STAR	54738	Star field	✓	✓	✓

SR, with convolutional neural networks (CNNs) enabling robust feature learning (e.g., [6, 8, 32]) and generative adversarial networks (GANs) enhancing perceptual quality through adversarial training (e.g., [33, 34, 35, 36]). Recent advancements introduced transformer-based models, leveraging global attention for superior detail recovery (e.g., [26, 29, 37, 29, 38]), and diffusion-based models, using iterative denoising for high-quality image generation (e.g., [36, 39, 40]). Unlike natural image SR, which prioritizes visual perception, astronomical SR must balance perceptual quality with the physical integrity of scientific data, as required in applications like stellar population analysis [10, 41].

2.2 Astronomical Image Super-Resolution

Super-resolution (SR) techniques tailored for astronomical images have evolved to address the unique challenges of celestial data, achieving notable success in enhancing specific targets like stars and galaxies. To improve image quality, the most direct method is to enhance the hardware capabilities of astronomical telescopes, leveraging advancements in optical and detection technologies. Common hardware improvements include increasing the telescope aperture, equipping telescopes with adaptive optics systems, advancing photodetector technology, and optimizing optical component design [42, 43, 44, 45, 46, 47]. These advancements complement software-based SR methods, which have significantly refined image resolution. Early software approaches, such as deconvolution [10, 41] and multi-frame stacking [48, 49], successfully improved the resolution of isolated stellar and galactic images by modeling point spread functions (PSFs) [47] or combining multiple exposures. These approaches enabled precise analyses of individual stars and galaxy morphologies [50]. More recently, computational advancements have explored SR for broader astronomical applications, primarily focusing on single-target scenarios like galaxies [51], Sun [52], X-ray sources (nebulae, active galactic nuclei, etc.) [53]. Our work extends this progress by developing a large-scale star field dataset that captures diverse astronomical conditions, enabling robust SR model training for complex star field scenes, an area previously underexplored. Additionally, we integrate flux consistency constraints to ensure reconstructed images preserve critical physical properties, enhancing their reliability for quantitative analyses such as photometry and stellar dynamics.

2.3 Flux Consistency in Astronomical Image Processing

Flux consistency, ensuring that the total light intensity (flux, or photons received per unit area) of celestial objects in processed images matches original observations, is a cornerstone of astronomical image analysis, underpinning reliable photometry and stellar population studies [54, 55]. Historical efforts prioritized flux consistency to preserve measurement accuracy in star clusters and galaxies. The modern space telescope SDSS also follows this principle [56]. However, the complexity of star fields, with diverse brightness and overlapping objects, poses ongoing challenges. Our work advances this field with STAR, a large-scale star field dataset ensuring flux-consistent image pairs, and novel flux consistency constraints, enhancing the scientific reliability of star field analyses.

3 STAR

Following natural SR works, we construct cross-resolution image pairs by downsampling high-resolution images. We choose Hubble Space Telescope (HST)³ survey data as our high-resolution images due to its widely recognized data quality and rich historical data accumulation. Given a high-resolution HST image, we first apply a point spread function (PSF) [47] kernel to simulate the optical blurring effects caused by low-quality telescopes and atmospheric turbulence.

Next, we downsample the image by a factor of s using a flux-conserving scheme. Finally, since both HR and LR images have large spatial dimensions, we divide them into smaller sub-images to facilitate model training. This pipeline generates physically consistent cross-resolution pairs of images, crucial for robust super-resolution model training.

3.1 High-resolution Data Collection

The HST is a space-based observatory designed to capture high-resolution astronomical images across a wide range of wavelengths, from ultraviolet to near-infrared. It provides two kinds of data, including calibration and science. Calibration data is used to correct instrumental effects while the science has a verified quality for scientific research. So we choose scientific data due to its high and reliable quality.

The science data consists of images captured by various imaging instruments onboard HST, such as the Advanced Camera for Surveys (ACS) [57], Wide Field Camera 3 (WFC3) [58], and Wide Field and Planetary Camera 2 (WFPC2) [59]. We selected the ACS Wide Field Channel (WFC/ACS) for its high sensitivity in optical wavelengths (350–1050 nm) and the widest field of view (202" \times 202"), which are critical for capturing high-resolution images of extended astronomical objects.

Astronomical data are captured under different filters, like natural images under Red, Green and Blue filters. Here, for the WFC/ACS science data, we keep the F814W filter (centered at 814 nm, also known as the I-band) data due to the band of the F814W is widely used in star field studies because its wavelength (centered at 814 nm) effectively resolves individual stars in crowded fields while maintaining high photometric accuracy [60]. So we choose it for its representative.

HST observes one location many times, resulting in a large number of overlapping images. To remove high-overlapping data but keep diversity as wide as possible, we use the farthest point sampling strategy [61] on the HST covered celestial regions and finally select 70 representative wide field images covering an extremely large range of celestial regions but without any overlapping.

3.2 PSF Blurring

Different resolution devices share different PSF blurring. To simulate this phenomenon, we adopt two representative PSF models. The first is the Gaussian PSF [62], which is widely used to approximate blur caused by atmospheric turbulence or instrumental imperfections. The second is the Airy PSF, which is a device-specific kernel widely used in astrophysics, describing the instrumental effect that occurs when a telescope resolution changes. Their detailed formulation could be seen in the supplementary. With these two PSFs, we define two blurring settings: one is a single Gaussian kernel G , and the other is a combination of Airy and Gaussian to simulate more complex scenarios.

³<https://www.stsci.edu/hst>

After operating PSF blurring, we perform a flux consistency downsampling scheme to obtain realistic low-resolution counterparts, which we detail in the next section.

3.3 Flux Consistency Downsampling

The flux consistency relation in real-world telescopic observations stems from the telescope imaging principle. The value of each pixel in observational data corresponds to the photon flux captured by its corresponding CCD pixel. Consequently, when imaging the same celestial region, a single CCD pixel in lower-resolution instruments covers a larger spatial area—equivalent to integrating photons from multiple high-resolution CCD pixels. This mechanism enables Flux-Consistent Downsampling by calculating the celestial receptive field ratio of each pixel across resolution scales. The details of this flux consistency downsampling could be seen in the supplementary.

3.4 HR Image Subdivision

The previous flux consistency downsampling scheme provides us a flux-consistent image pairs. To further optimize model training effectiveness, we divide the HR and LR wide field images into smaller sub-images. Because astronomical images contain some outlier regions (have NaN values) due to geometric calibration in DrizzlePac[63] processing, we retain only patches with >80% valid regions containing stellar features. As a result of these processing steps, we obtain a large set of high-quality HR-LR image pairs and construct the STAR dataset for training and evaluating astronomical super-resolution models.

3.5 Flux Error

The STAR provides a Flux Error (FE) to evaluate flux consistency between a ground truth and its corresponding prediction. The FE measures the flux value gap for each object, so its computation process is based on astrophysical photometry. The basic process of the photometry is deriving flux by detection. We follow this idea. For a given ground-truth and predicted image pair, we compute the FE as following two steps: **1**). For the ground-truth image, we use the Starfinder toolkit [64] to detect celestial objects and obtain their parameters. Then, we derive the flux of each object by a widely-used elliptical photometry method [64]. **2**). For predicted images, we do not operate object detection but directly use detection results from the ground-truth image because they provide reliable object catalogs covering both strong and weak sources. The following photometry is the same as the ground truth. After these two steps, we have a two set of flux values, denoted as $\{v_{pred}^1, v_{pred}^2, \dots, v_{pred}^N\}$ and $\{v_{gt}^1, v_{gt}^2, \dots, v_{gt}^N\}$ where N is the number of detected objects. The FE is computed by the following:

$$FE = \frac{1}{N} \sum_{i=1}^N |v_{gt}^i - v_{pred}^i|. \quad (1)$$

Lower FE means higher flux consistency. Since the flux value is related to the object shape and the pixel flux in object regions, this metric could reflect the geometric shape consistency for each object in the reconstruction image, physically informed flux consistency and weak source object reconstruction quality simultaneously.

4 Method

4.1 Overview

The entire pipeline of our Flux-Invariant Super-Resolution (FISR) is shown in Fig. 2. The low-resolution input image are sent into two paths. The first one is an encoder block consisting of a convolution operation and multiple transformer blocks, which represent the input image as multi-scale feature maps. The second is a novel Flux Guidance Generation (FGG) module that extracts flux information as multi-scale flux guidance representations. Each flux guidance is sent into a scale-wise Flux Guidance Controller (FGC) to enhance the encoded feature maps. This scheme highlights regions with significant flux to guide the network’s focus toward astrophysically relevant structures. Then, multiple decoder transformer blocks progressively process these enhanced features and finally upsample them as the output images.

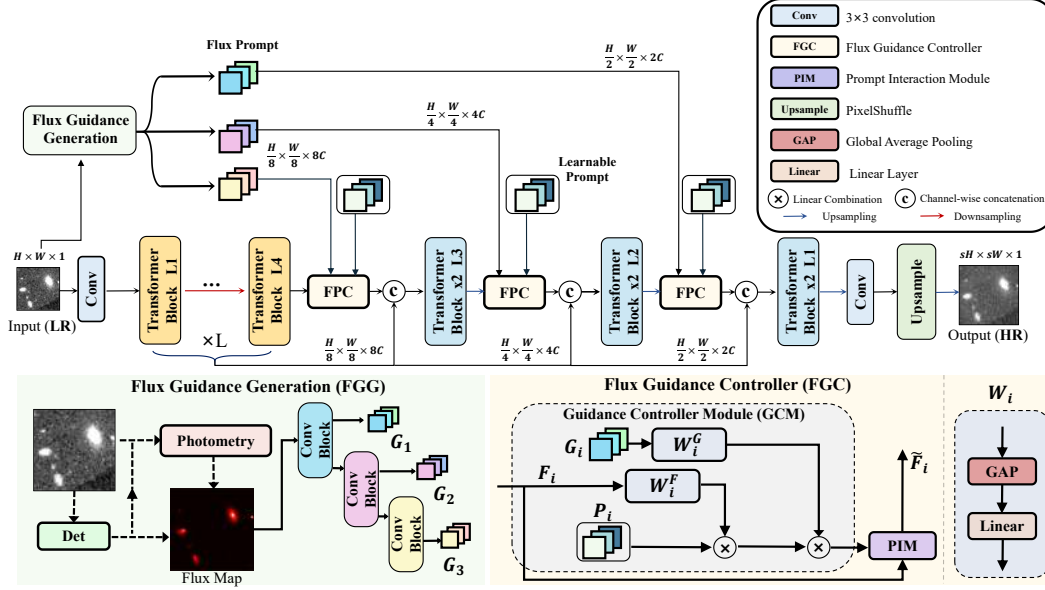


Figure 2: Overview of the FISR approach. The input image is processed through an encoder branch and a Flux Guidance Generation (FGG). Flux information is extracted via FGG and injected into the Flux Guidance Controllers (FGC) to enhance the encoded feature map. The enhanced features are then decoded and upsampled to produce the final high-resolution output..

4.2 Flux Guidance Generation

We propose the (Flux Guidance Generation) FGG module to introduce the flux information. Specifically, given an input image, FGG first represents flux information for every celestial object in the input image into a flux map, then transfers the flux map as multi-scale features as guidance which will then be used in the next FGC module to enhance multi-scale feature maps.

To generate the flux map, FGG also computes flux by detection. It detects celestial objects first and obtains a bounding box for each object. Then the photometry process is operated to derive object fluxes. The entire detection and photometry process is real-time. With these, FGG generates the map following the idea of drawing bounding boxes on a white background and corresponding flux value in each bounding box region.

In practice, because the 'bounding box' of a celestial object is essentially a rotatable ellipse rather than a rectangle in standard object detection, the drawing scheme FGG has some modifications. Specifically, for each ellipse, FGG puts a rotatable Gaussian Kernel at the center location and adjusts the Gaussian standard deviation based on the ellipse size. Then, FGG multiplies the object flux value directly with the rotated Gaussian kernel to modulate the kernel value. Finally, FGG draws all these kernels together as the final flux map.

Finally, a multi-scale convolutional block transforms the flux map into a corresponding flux guidance representation. In our architecture, we employ three such blocks at different feature levels to generate a hierarchy of flux guidance, denoted as $\{G_1, G_2, G_3\}$. The level 3 is set to align the block number in the decoder, following common setting in a popular SR baseline [27].

4.3 Flux Guidance Controller

Based on the flux guidance produced by the FGG module, the (Flux Guidance Controller) FGC interacts such guidance with the encoder features, generating the enhanced features. The FGC is scale-wised and for the i -th scale FGC, its interaction pipeline is represented as follows:

$$\hat{F}_i = \text{PIM}_i(\text{GCM}_i(P_i, F_i, G_i), F_i). \quad (2)$$

It shows that the enhanced feature $\hat{F}_i \in \mathcal{R}^{H \times W \times C}$ is derived from a guidance-controlled feature and an original feature $F_i \in \mathcal{R}^{H \times W \times C}$ by a prompt interaction module function $\text{PIM}(\cdot, \cdot)$, which is set

same as combines image features with guidance-controlled feature and dynamically adjusts the input features through a transformer block. following PromptIR [27]. The guidance-controlled features are computed by a guidance controller module function $\text{GCM}(\cdot, \cdot, \cdot)$ that takes a learnable prompt \mathbf{P}_i , the flux guidance component $\mathbf{G}_i \in \mathcal{R}^{H \times W \times C}$ and the encoded feature \mathbf{F}_i as inputs. The learnable prompt $\mathbf{P}_i \in \mathcal{R}^{H \times W \times C \times K}$ contains K learnable patterns expected to represent blind property in the image restoration process [27]. The detail of GCM is as follows:

$$\text{GCM}_i(\mathbf{P}_i, \mathbf{F}_i, \mathbf{G}_i) = \sum_{k \in K} W_i^F(\mathbf{F}_i) \odot W_i^G(\mathbf{G}_i) \odot \mathbf{P}_i, \quad (3)$$

where W_i means learnable modules, consisting of a global average pooling and a linear layer, takes input corresponding features and derives a weight with the size of $1 \times 1 \times 1 \times K$ to indicate the importance of the learnable prompt patterns. \odot means Hadamard product with dimension broadcast while $\sum_{k \in K}$ means sum the last dimension of the multiplied features with the size of $H \times W \times C \times K$ to derive the final output features with the size of $H \times W \times C$.

4.4 Flux Consistency Loss

The idea of the Flux Consistency Loss is to train the model to generate flux consistent prediction. We propose a simple yet effective scheme to achieve this goal by highlighting flux-related regions. We first operate the flux map generation scheme described in Section 4.2 on the ground-truth image, denoted as M . Then we use this map to weighted the pixel wised supervision as follows:

$$\mathcal{L}_{\text{flux}}(I_{\text{pred}}, I_{\text{gt}}) = \sum_{x,y} M(x,y) \cdot |I_{\text{pred}}(x,y) - I_{\text{gt}}(x,y)|. \quad (4)$$

Combined it with a reconstruction loss (L1 or L2), the total loss is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}}(I_{\text{pred}}, I_{\text{gt}}) + \lambda \cdot \mathcal{L}_{\text{flux}}(I_{\text{pred}}, I_{\text{gt}}), \quad (5)$$

where λ balances terms. This loss takes into account both traditional regression supervision and flux consistency constraints, which brings more physically reliable ASR results.

5 Experiment

5.1 Experimental Setup

Dataset. We use the proposed STAR to process comparisons, evaluate our model and perform ablation studies. The downsampling ratio s is set as 2 and 4, respectively. As mentioned before, the blurring PSF has two settings: Gaussian only and Gaussian+Airy. The experimental results of the latter are placed in Appendix C.

Implementation details. Each model is trained for 100 training epochs with a batch size of 16 on 8 A100 GPUs. The initial learning rate is set to 2e-4 and decayed by a factor of 0.01 at the 50th epoch. To make the training more stable, we apply the linear warm-up strategy in the first epoch. More details about the environment setting can be found in Appendix F.

Evaluation protocols. Following prior work, we adopt Peak Signal-to-Noise Ratio (PSNR) [65] and Structural Similarity Index (SSIM) [66] as standard evaluation metrics to measure reconstruction quality. Further, we use the proposed FE to evaluate physical fidelity to quantify the alignment between the output of the model and astrophysical principles. Note that we do not use SR commonly-used perceptual metrics, such as LPIPS [67], which leverages ImageNet [68] pre-trained deep models [28, 69] to measure semantic similarity. Because in our ASR task, they are not suitable due to the significant domain gap between the pre-trained domain and astronomical images.

5.2 Quantitative and Qualitative Results

We compare our method with several methods in Tab. 2, including state-of-the-art natural SR methods, such as HAT [29] and classical SR methods, for example, SwinIR [26]. SwinIR and RealESRGAN [34] are also popular in the ASR topic [11, 17], which is also an important factor that we choose them to compare.

Table 2: Performance comparison of different methods under $\times 2$ and $\times 4$ super-resolution. Evaluation metrics include PSNR, SSIM, and Flux Error.

Scale	Metric	Bicubic	EDSR [8]	RealESRGAN [34]	RCAN [32]	SwinIR [26]	HAT [29]	FISR (ours)
$\times 2$	PSNR \uparrow	28.6021	35.3816	36.8363	36.3703	37.2205	37.2501	37.8779
	SSIM \uparrow	0.6842	0.8054	0.8225	0.8240	0.8286	0.8295	0.8311
	FE \downarrow	4.9418	1.4623	7.3632	0.9237	0.813	0.7636	0.5739
$\times 4$	PSNR \uparrow	26.0518	33.8736	34.3725	34.9823	34.6655	34.9142	35.1788
	SSIM \uparrow	0.6005	0.7201	0.7223	0.7263	0.7258	0.7276	0.7266
	FE \downarrow	7.8733	1.3841	4.0782	1.0550	1.0657	1.0256	1.0125

From Tab. 2, it could be shown that our approach outperforms existing methods across different evaluation criteria. Compared with HAT [29], our method achieves complete priority under the $\times 2$ case while keeping most advantages under the harder $\times 4$ case, showing the effectiveness of our approach under different scenarios. The better PSNR and SSIM demonstrate higher reconstruction quality of our FISR model. What’s more, a significant FE priority shows the satisfactory physical reliability of our method, proving its value in physically faithful high-precision astrophysics image processing. The comparison of the second-best SwinIR [26] who is the current state-of-the-art ASR method, demonstrates that our method achieves a new state of the art in the ASR topic, further proving our effectiveness.

5.3 Ablation Study

To evaluate the contribution of each proposed component in our FISR framework, we conduct comprehensive ablation studies on the STAR dataset under the $\times 2$ super-resolution setting. Tab. 3 presents the performance changes when incorporating or removing the Flux Guidance Generation (FGG), Flux Guidance Controller (FGC) and Flux Consistency Loss (FCL).

The ablation of FCL is straightforward, but it is not easy and also trivial for us to ablate FGC and FGG solely. So here, we modify the ablation study of FGC and FGG as the ablation of FG-Modules and Flux cues. FG-Modules means keeping all learnable modules of FGG+FGC, but replacing the flux map of FGG as the original input image. Only using FG-Modules means ablating the input flux cues. The idea behind this design is that the key motivation of FGC+FGG is to introduce input flux information. So, the ablation of the flux map is effective because it could demonstrate that the gains introduced by the FGC and FGG are actually from flux information rather than more parameters.

Effect of FCL: In the 1st line of Tab. 3, we show the performance of our baseline, PromptIR [27]. In the 2nd line, we introduce the FCL, significantly decreasing the physical faithful metric FE about 0.06+, showing its function to achieve flux consistency. Similar gains can be found by comparing the 5th and the 6th lines. Except that, FCL also introduces little PSNR and SSIM gains, showing its compatibility and potential benefits for image reconstruction. To further demonstrate the generalization of the FCL, we combine FCL with several state-of-the-art SR methods, as shown in Tab. 4. It could be seen that the FCL widely increases their performances, solidly demonstrating its generalization.

Effect of FGG+FGC: In the 3rd line, we introduce the FG-Modules. Compared with the 1st line, we could find that directly introducing the FG-Modules is trivial, even bringing large FE downgradation. Comparing 2nd and 4th lines could derive similar conclusions. It shows that more parameters are trivial. However, when we introduce our flux cues in the 5th line, it brings a significant increase across all metrics. The 0.1+ FE decrease makes the model achieve high flux consistency without the FCL. Finally, when we introduce FCL, the performance has further gains, showing the compatibility of our different proposed modules.

5.4 Quality Analysis

Fig. 3 visually compares various super-resolution models on star field regions. The FISR model demonstrates superior reconstruction quality over traditional CNN-based methods such as EDSR and RCAN. Although these baselines can recover structures, they often fail to preserve flux consistency, particularly in regions with bright or overlapping celestial sources.

Table 3: Ablation study on the effectiveness of the Flux Guidance Generation (FGG) and Flux Error (FE). Metrics are reported on $\times 2$ SR task.

	FG-Modules	Flux cues	FCL	PSNR \uparrow	SSIM \uparrow	FE \downarrow
1 st				37.7715	0.8288	0.7022
2 nd				37.8570	0.8283	0.6389
3 rd	✓		✓	37.7101	0.8286	0.7572
4 th	✓		✓	37.8681	0.8301	0.7467
5 th	✓	✓		37.8454	0.8302	0.6527
6 th	✓	✓	✓	37.8779	0.8311	0.5739

Table 4: Comparison of different methods with and without FCL under $\times 2$ and $\times 4$ ASR.

Scale	Flux Loss	EDSR [8]	RealESRGAN [34]	RCAN [32]	SwinIR [26]	PromptIR [27]	HAT [29]
$\times 2$	PSNR \uparrow (w/o)	35.3816	36.8363	36.3703	37.2205	37.7715	37.2501
	SSIM \uparrow (w/o)	0.8054	0.8225	0.8240	0.8286	0.8288	0.8295
	Flux Error \downarrow (w/o)	1.4623	7.3632	0.9237	0.8130	0.7022	0.7636
	PSNR \uparrow (w/)	35.6259	36.8647	37.6441	37.5098	37.8570	38.0880
	SSIM \uparrow (w/)	0.8064	0.8222	0.8291	0.8280	0.8283	0.8320
	Flux Error \downarrow (w/)	1.3334	6.4402	0.6631	0.7809	0.6389	0.6042
$\times 4$	PSNR \uparrow (w/o)	33.8736	34.3725	34.9823	34.6655	34.6726	34.9142
	SSIM \uparrow (w/o)	0.7201	0.7223	0.7263	0.7258	0.7230	0.7276
	Flux Error \downarrow (w/o)	1.3841	4.0782	1.0550	1.0657	1.0800	1.0256
	PSNR \uparrow (w/)	34.2381	34.9243	34.2024	35.1610	35.0936	35.3156
	SSIM \uparrow (w/)	0.7205	0.7234	0.7234	0.7265	0.7253	0.7280
	Flux Error \downarrow (w/)	1.3659	1.1219	1.0755	1.0634	1.0227	1.0199

To further emphasize the impact of our flux consistency loss, we compute the Kullback-Leibler (KL) [70] and Jensen-Shannon (JS) [71] divergence between the predicted and ground-truth intensity distributions within selected regions. Notably, FISR—especially when trained with the flux loss—achieves significantly lower divergence scores, reflecting improved flux preservation and more physically accurate reconstructions.

5.5 Evaluation on Downstream Scientific Tasks

To address concerns about error propagation to scientific outputs, we evaluated our method on two representative downstream tasks: **stellar mass estimation** and **weak lensing shear measurement**.

For **stellar mass estimation**, we applied a simplified photometric pipeline to the STAR test set, converting predicted fluxes to magnitudes ($\text{mag} = -2.5 \times \log_{10}(\text{flux}) + 25.0$) and inferring stellar masses using a constant mass-to-light ratio ($M/L \approx 3.0$). As shown in Table 5, our FISR method achieves the lowest predicted magnitude error (1.66×10^{-7}) alongside RealESRGAN (1.67×10^{-7}),

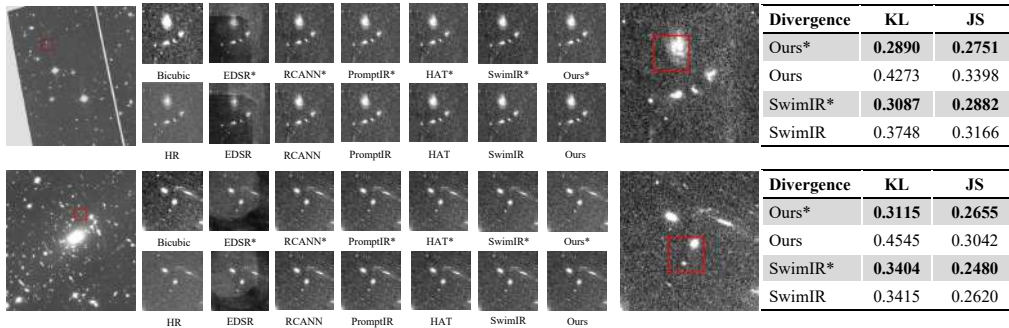


Figure 3: Visual comparison on two star field regions. Red boxes mark areas for computing KL and JS divergence between predictions and ground truth. Models with (*) are trained using FCL.

Table 5: Evaluation on downstream scientific tasks. We report predicted magnitude error, stellar mass MAE, and mean shear error for different methods. Lower values indicate better performance.

Method	Pred. Mag. Error	Mass MAE	Mean Shear Error
Bicubic	3.23×10^{-7}	1.79×10^{-7}	2.10×10^{-1}
SwinIR	2.02×10^{-7}	5.19×10^{-8}	1.98×10^{-1}
EDSR	2.96×10^{-7}	1.37×10^{-7}	2.14×10^{-1}
RCAN	2.01×10^{-7}	5.36×10^{-8}	2.06×10^{-1}
HAT	1.67×10^{-7}	3.06×10^{-8}	1.88×10^{-1}
RealESRGAN	3.37×10^{-7}	3.99×10^{-7}	1.95×10^{-1}
FISR (Ours)	1.66×10^{-7}	2.81×10^{-8}	1.87×10^{-1}

and demonstrates the best stellar mass MAE (2.81×10^{-8}), significantly outperforming other methods including EDSR (1.37×10^{-7}) and Bicubic (1.79×10^{-7}).

For **weak lensing shear measurement**, a critical task for cosmological studies, we computed shear components $\gamma = \gamma_1 + i\gamma_2$ from SEP-detected galaxy ellipses, where:

$$\gamma_1 = \frac{a^2 - b^2}{a^2 + b^2} \times \cos(2\theta) \quad (6)$$

$$\gamma_2 = \frac{a^2 - b^2}{a^2 + b^2} \times \sin(2\theta) \quad (7)$$

with semi-major axis a , semi-minor axis b , and position angle θ . We measured the mean shear difference $|\gamma_{\text{pred}} - \gamma_{\text{gt}}|$ between predictions and ground truth. The results in Table 5 show that FISR achieves competitive shear preservation (1.87×10^{-1}), comparable to the best-performing HAT (1.88×10^{-1}) and superior to methods like RCAN (2.14×10^{-1}). These comprehensive evaluations demonstrate that our flux-preserving approach maintains both photometric accuracy and morphological fidelity essential for downstream scientific inference, validating its practical utility in astronomical applications.

6 Conclusion

In this work, we present STAR, a large-scale, flux-consistent benchmark specifically designed for astronomical super-resolution (ASR). Addressing critical limitations in prior datasets—such as flux inconsistency, object-centric bias, and limited diversity—STAR captures complex star fields with physically faithful flux distributions, cross-object interactions, and weak lensing effects. Alongside, we introduce a novel evaluation metric, Flux Error (FE), and propose the Flux-Invariant Super-Resolution (FISR) model, which incorporates flux-aware prompts and consistency loss. Extensive experiments show that FISR not only achieves state-of-the-art reconstruction quality but also significantly improves flux consistency.

7 Limitations and future work

While our study offers promising insights, it has a few limitations that merit further exploration. First, our experiments are based on observations from a single telescope, the HST WFC/ACS with the F814W filter, which may limit the generalizability of our findings to other instruments or observational contexts. Additionally, although our network design performs well, it could benefit from incorporating more domain-specific optimizations rooted in astronomical knowledge, such as leveraging physical principles or astronomical priors to enhance performance in complex scenarios. These areas present opportunities for future refinement. Looking forward, we aim to broaden the applicability of our method by extending it to a wider array of advanced telescopes, such as the James Webb Space Telescope (JWST) [72] or the upcoming Large Synoptic Survey Telescope (LSST) [73], to explore its potential across diverse astronomical contexts. Through these efforts, we hope to make modest contributions to the field of astronomical image processing, fostering the development of more robust and adaptable tools for future discoveries.

8 Acknowledgments

This work was supported by the Shanghai Artificial Intelligence Laboratory. This work was also supported by the JC STEM Lab of AI for Science and Engineering, funded by The Hong Kong Jockey Club Charities Trust, and the Research Grants Council of Hong Kong (Project No. CUHK14213224). We sincerely thank Hao Du, Yating Liu, Jiaze Li, Yingfan Hua, and Jun Yao for their invaluable guidance and insightful feedback throughout this research.

References

- [1] A Labeyrie. Ii high-resolution techniques in optical astronomy. *Progress in optics*, 14:47–87, 1977.
- [2] Aaron Bryant and Alfred Krabbe. The episodic and multiscale galactic centre. *New Astronomy Reviews*, 93:101630, 2021.
- [3] Muhammad Faaique. Overview of big data analytics in modern astronomy. *International Journal of Mathematics, Statistics, and Computer Science*, 2:96–113, 2024.
- [4] W Patrick McCray. *Giant telescopes: Astronomical ambition and the promise of technology*. Harvard University Press, 2006.
- [5] W Patrick McCray. Large telescopes and the moral economy of recent astronomy. *Social studies of science*, 30(5):685–711, 2000.
- [6] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.
- [7] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5197–5206, 2015.
- [8] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.
- [9] R Marsh, TR Young, T Johnson, and D Smith. Enhancement of small telescope images using super-resolution techniques. *Publications of the Astronomical Society of the Pacific*, 116(819):477, 2004.
- [10] Klaus G Puschmann and Franz Kneer. On super-resolution in astronomical imaging. *Astronomy & Astrophysics*, 436(1):373–378, 2005.
- [11] Jiawei Miao, Liangping Tu, Bin Jiang, Xiangru Li, and Bo Qiu. Astrosr: A data set of galaxy images for astronomical superresolution research. *The Astrophysical Journal Supplement Series*, 274(1):7, 2024.
- [12] Qian-Qian Shan, Cui-Xiang Liu, Bo Qiu, A-Li Luo, Fu-ji Ren, Zhi-Ren Pan, and Yi-Cong Chen. Galaxy image super-resolution reconstruction using diffusion network. *Engineering Applications of Artificial Intelligence*, 142:109836, 2025.
- [13] Karl D Gordon, Ralph Bohlin, GC Sloan, George Rieke, Kevin Volk, Martha Boyer, James Muzerolle, Everett Schlafin, Susana E Deustua, Dean C Hines, et al. The james webb space telescope absolute flux calibration. i. program design and calibrator stars. *The Astronomical Journal*, 163(6):267, 2022.
- [14] Allan W Smith, John T Woodward, Colleen A Jenkins, Steven W Brown, and Keith R Lykke. Absolute flux calibration of stars: calibration of the reference telescope. *Metrologia*, 46(4):S219, 2009.
- [15] Ralph C Bohlin, Susana E Deustua, and Gisella de Rosa. Hubble space telescope flux calibration. i. stis and calspec. *The Astronomical Journal*, 158(5):211, 2019.
- [16] Wei Song, Ying Ma, Haoying Sun, Xiaobing Zhao, and Ganghua Lin. Improving the spatial resolution of solar images using super-resolution diffusion generative adversarial networks. *Astronomy & Astrophysics*, 686:A272, 2024.
- [17] Pranath Reddy, Michael W Toomey, Hanna Parul, and Sergei Gleyzer. Diffense: a conditional diffusion model for super-resolution of gravitational lensing data. *Machine Learning: Science and Technology*, 5(3):035076, 2024.

- [18] Zhijian Luo, Shaohua Zhang, Jianzhen Chen, Zhu Chen, Liping Fu, Hubing Xiao, Wei Du, and Chenggang Shu. Cross-survey image transformation: Enhancing sdss and decals images to near-hsc quality for advanced astronomical analysis. *The Astrophysical Journal Supplement Series*, 277(1):22, 2025.
- [19] Yookyung Noh. *The Large-scale Structure of the Universe: Probes of Cosmology and Structure Formation*. University of California, Berkeley, 2013.
- [20] Christopher J Conselice. Galaxy mergers and interactions at high redshift. *Proceedings of the International Astronomical Union*, 2(S235):381–384, 2006.
- [21] Henk Hoekstra and Bhuvnesh Jain. Weak gravitational lensing and its cosmological applications. *Annual Review of Nuclear and Particle Science*, 58(1):99–123, 2008.
- [22] Dark Energy Survey Collaboration et al. The dark energy survey. *arXiv preprint astro-ph/0510346*, 2005.
- [23] Wanjun Li, Zhe Liu, and Hongtao Deng. A self-attention based srgan for super-resolution of astronomical image. In *2022 IEEE 8th International Conference on Computer and Communications (ICCC)*, pages 1977–1981. IEEE, 2022.
- [24] Nick Scoville, RG Abraham, H Aussel, JE Barnes, A Benson, AW Blain, D Calzetti, A Comastri, P Capak, C Carilli, et al. Cosmos: Hubble space telescope observations. *The Astrophysical Journal Supplement Series*, 172(1):38, 2007.
- [25] Karl D Gordon, Adolf N Witt, and Brian C Friedmann. Detection of extended red emission in the diffuse interstellar medium. *The astrophysical journal*, 498(2):522, 1998.
- [26] Jingyun Liang, Jiezhong Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021.
- [27] Vaishnav Potlapalli, Syed Waqas Zamir, Salman H Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one image restoration. *Advances in Neural Information Processing Systems*, 36:71275–71293, 2023.
- [28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [29] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22367–22377, 2023.
- [30] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018.
- [31] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pages 711–730. Springer, 2010.
- [32] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018.
- [33] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.
- [34] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021.

- [35] Kelvin CK Chan, Xiangyu Xu, Xintao Wang, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for image super-resolution and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3154–3168, 2022.
- [36] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022.
- [37] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution. In *European conference on computer vision*, pages 649–667. Springer, 2022.
- [38] Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tiejiong Zeng. Transformer for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 457–466, 2022.
- [39] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 132(12):5929–5949, 2024.
- [40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [41] Jean-Luc Starck, Eric Pantin, and Fionn Murtagh. Deconvolution in astronomy: A review. *Publications of the Astronomical Society of the Pacific*, 114(800):1051, 2002.
- [42] Michael C Roggemann, Byron M Welsh, and Robert Q Fugate. Improving the resolution of ground-based telescopes. *Reviews of Modern Physics*, 69(2):437, 1997.
- [43] James T Early, Roderick Hyde, and Richard L Baron. Twenty-meter space telescope based on diffractive fresnel lens. In *UV/Optical/IR Space Telescopes: Innovative Technologies and Concepts*, volume 5166, pages 148–156. SPIE, 2004.
- [44] Andreas Glindemann, Stefan Hippler, Thomas Berkefeld, and Wolfgang Hackenberg. Adaptive optics on large telescopes. *Experimental Astronomy*, 10:5–47, 2000.
- [45] Paul Hickson. Atmospheric and adaptive optics. *The Astronomy and Astrophysics Review*, 22:1–38, 2014.
- [46] Christopher R Kitchin. *Astrophysical techniques*. CRC press, 2020.
- [47] James B Breckinridge, Wai Sze T Lam, and Russell A Chipman. Polarization aberrations in astronomical telescopes: the point spread function. *Publications of the Astronomical Society of the Pacific*, 127(951):445, 2015.
- [48] Michael Hirsch, S Harmeling, S Sra, and B Schölkopf. Online multi-frame blind deconvolution with super-resolution and saturation correction. *Astronomy & Astrophysics*, 531:A9, 2011.
- [49] Stefano Zibetti, Brice Ménard, Daniel B Nestor, Anna M Quider, Sandhya M Rao, and David A Turnshek. Optical properties and spatial distribution of mg ii absorbers from sdss image stacking. *The Astrophysical Journal*, 658(1):161, 2007.
- [50] Rainer Schödel, David Merritt, and Andreas Eckart. The nuclear star cluster of the milky way: proper motions and mass. *Astronomy & Astrophysics*, 502(1):91–111, 2009.
- [51] Kevin Schawinski, Ce Zhang, Hantian Zhang, Lucas Fowler, and Gokula Krishnan Santhanam. Generative adversarial networks recover features in astrophysical images of galaxies beyond the deconvolution limit. *Monthly Notices of the Royal Astronomical Society: Letters*, 467(1):L110–L114, 2017.
- [52] CJ Díaz Baso and A Asensio Ramos. Enhancing sdo/hmi images using deep learning. *Astronomy & Astrophysics*, 614:A5, 2018.

- [53] Sam F Sweere, Ivan Valtchanov, Maggie Lieu, Antonia Vojtekova, Eva Verdugo, Maria Santos-Lleo, Florian Pacaud, Alexia Briassouli, and Daniel Cámpora Pérez. Deep learning-based super-resolution and de-noising for xmm-newton images. *Monthly Notices of the Royal Astronomical Society*, 517(3):4054–4069, 2022.
- [54] Sidney Van den Bergh. Ubv photometry of globular clusters. *Astronomical Journal*, Vol. 72, p. 70-81 (1967), 72:70–81, 1967.
- [55] Michael S Bessell. Standard photometric systems. *Annu. Rev. Astron. Astrophys.*, 43(1):293–336, 2005.
- [56] Nikhil Padmanabhan, David J Schlegel, Douglas P Finkbeiner, JC Barentine, Michael R Blanton, Howard J Brewington, James E Gunn, Michael Harvanek, David W Hogg, Željko Ivezić, et al. An improved photometric calibration of the sloan digital sky survey imaging data. *The Astrophysical Journal*, 674(2):1217, 2008.
- [57] Holland C Ford, Mark Clampin, George F Hartig, Garth D Illingworth, Marco Sirianni, Andre R Martel, Gerhardt R Meurer, William Jon McCann, Pamela C Sullivan, Frank Bartko, et al. Overview of the advanced camera for surveys on-orbit performance. In *Future EUV/UV and Visible Space Astrophysics Missions and Instrumentation*, volume 4854, pages 81–94. SPIE, 2003.
- [58] L Dressel, M Wong, C Pavlovsky, K Long, et al. Wide field camera 3 instrument handbook, 2010.
- [59] Jon A Holtzman, J Jeff Hester, Stefano Casertano, John T Trauger, Alan M Watson, Gilda E Ballester, Christopher J Burrows, John T Clarke, David Crisp, Robin W Evans, et al. The performance and calibration of wfpc2 on the hubble space telescope. *Publications of the Astronomical Society of the Pacific*, 107(708):156, 1995.
- [60] M Sirianni, MJ Jee, N Benítez, JP Blakeslee, AR Martel, Gerhardt Meurer, M Clampin, G De Marchi, HC Ford, R Gilliland, et al. The photometric performance and calibration of the hubble space telescope advanced camera for surveys. *Publications of the Astronomical Society of the Pacific*, 117(836):1049, 2005.
- [61] Yuval Eldar, Michael Lindenbaum, Moshe Porat, and Yehoshua Y Zeevi. The farthest point strategy for progressive image sampling. *IEEE transactions on image processing*, 6(9):1305–1315, 1997.
- [62] Jan J Koenderink and Andrea J van Doorn. Representation of local geometry in the visual system. *Biological cybernetics*, 55(6):367–375, 1987.
- [63] S Gonzaga, W Hack, A Fruchter, and J e Mack. The drizzlepac handbook. *The DrizzlePac Handbook*, 2012.
- [64] Kyle Barbary. Sep: Source extraction and photometry. *Astrophysics Source Code Library*, pages ascl–1811, 2018.
- [65] Jochen Antkowiak, T Jamal Baina, France Vittorio Baroncini, Noel Chateau, France FranceTelecom, Antonio Claudio França Pessoa, F Stephanie Colonnese, Italy Laura Contin, Jorge Caviedes, and France Philips. Final report from the video quality experts group on the validation of objective models of video quality assessment march 2000. *Final report from the video quality experts group on the validation of objective models of video quality assessment march*, 2000, 2000.
- [66] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
- [67] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

- [68] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [69] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [70] Solomon Kullback. Kullback-leibler divergence, 1951.
- [71] María Luisa Menéndez, Julio Angel Pardo, Leandro Pardo, and María del C Pardo. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318, 1997.
- [72] Klaus M Pontoppidan, Jaclyn Barrientes, Claire Blome, Hannah Braun, Matthew Brown, Margaret Carruthers, Dan Coe, Joseph DePasquale, Néstor Espinoza, Macarena García Marin, et al. The jwst early release observations. *The Astrophysical Journal Letters*, 936(1):L14, 2022.
- [73] Željko Ivezić, Steven M Kahn, J Anthony Tyson, Bob Abel, Emily Acosta, Robyn Allsman, David Alonso, Yusra AlSayyad, Scott F Anderson, John Andrew, et al. Lsst: from science drivers to reference design and anticipated data products. *The Astrophysical Journal*, 873(2):111, 2019.
- [74] Wikipedia contributors. Point spread function — Wikipedia, the free encyclopedia, 2025. [Online; accessed 6-October-2025].
- [75] John W Hardy. *Adaptive optics for astronomical telescopes*, volume 16. Oxford university press, 1998.
- [76] François Roddier. V the effects of atmospheric turbulence in optical astronomy. In *Progress in optics*, volume 19, pages 281–376. Elsevier, 1981.

A Flux Consistency Downsampling Details

Computing image plane coordinates: For a given high-resolution (HR) image with the resolution of $H \times W$ and a downsampling rate s , we generate the size of the downsampled low-resolution (LR) image, referred to as $\frac{H}{s} \times \frac{W}{s}$. With the two sizes, we have specific coordinates of pixels in both LR and HR images.

Transfer pixels to sky: For HR and LR pixel coordinates, we transfer them into the celestial coordinate system as: $(u, v) \rightarrow (ra, dec)$, where (u, v) is a coordinate in the image plane while (ra, dec) is the longitude and latitude coordinates of the Earth. Note that, each pixel is not an ideal point and actually a rectangle on the image plane. After the mapping, it becomes a quadrilateral surface of the celestial coordinate system. The physical meaning of this quadrilateral surface is the sky area covered by a pixel, denoted as the receptive field here. For the i -th pixel of the LR image and the j -th pixel of the HR image, we calculate and denote the area value of their receptive field as A_i^{LR} and A_j^{HR} .

The transformation process in the aforementioned process is implemented by the telescope calibration information carried by the high-resolution (HR) image, which could be interpreted as camera intrinsic and extrinsic parameters of the telescope.

Low-resolution image Flux Computation: The previous steps essentially transferred HR and LR image plane grids into two surface meshes in the celestial coordinate system, as shown in Fig. 4. Obviously, the average receptive field of the LR image is larger than the HR one because the LR pixel corresponds to larger regions, leading to an LR pixel covers multiple HR pixels in the sky. To compute the flux of the i -th LR pixel, we first identify the set of HR pixels S_i^o whose receptive fields overlap with that of the i -th LR pixel, i.e., $S_i^o = \{j \mid A_j^{HR} \cap A_i^{LR} \neq \emptyset\}$. This set represents all HR pixels whose sky areas contribute to the i -th LR pixel's flux. The flux of the i -th LR pixel, F_i^{LR} , is then computed by summing the weighted contributions from all overlapping HR pixels:

$$F_i^{LR} = \sum_{j \in S_i^o} w_{i,j} \cdot f_j^{HR}, \quad (8)$$

where f_j^{HR} is the flux of the j -th HR pixel, and $w_{i,j}$ is the weight representing the fractional contribution of the j -th HR pixel to the i -th LR pixel.

The weight $w_{i,j}$ is calculated as:

$$w_{i,j} = \frac{A_{i,j}}{A_j^{HR}}, \quad (9)$$

where $A_{i,j}$ is the overlapping sky area between the i -th LR pixel and the j -th HR pixel, representing their shared quadrilateral patch in the celestial coordinate system, and A_j^{HR} is the total sky area covered by the j -th HR pixel's receptive field. To better understand the role of this weight in flux computation, we substitute $w_{i,j}$ into Equation (8), transforming the contribution term as follows. The flux contribution from the j -th HR pixel to the i -th LR pixel is $w_{i,j} \cdot f_j^{HR}$, where f_j^{HR} is the flux of the j -th HR pixel. Substituting $w_{i,j} = \frac{A_{i,j}}{A_j^{HR}}$ into this term, we obtain:

$$w_{i,j} \cdot f_j^{HR} = \left(\frac{A_{i,j}}{A_j^{HR}} \right) \cdot f_j^{HR} = A_{i,j} \cdot \frac{f_j^{HR}}{A_j^{HR}}. \quad (10)$$

Here, $\frac{f_j^{HR}}{A_j^{HR}}$ represents the flux density of the j -th HR pixel, i.e., the photon count per unit sky area, as recorded by the telescope's CCD sensor over the receptive field area A_j^{HR} . Thus, $A_{i,j} \cdot \frac{f_j^{HR}}{A_j^{HR}}$ is the flux contributed by the j -th HR pixel over the overlapping area $A_{i,j}$, ensuring that the contribution is proportional to the shared sky area between the LR and HR pixels. This approach preserves the total photon flux during downsampling, maintaining flux consistency across resolutions.

As shown in Fig. 5, we compare flux consistency downsampling with traditional bilinear interpolation. It can be found that the result of Fig. 5 (a) is closer to the average flux of HR star sources, indicating that flux consistency downsampling can better keep the original HR flux information. To further highlight the differences between the two methods, we visualize their residuals in Fig. 5 (c). Noticeable

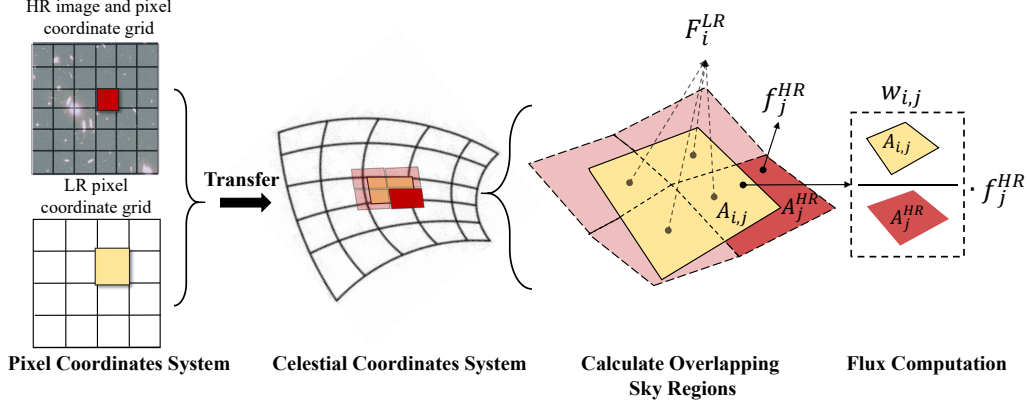


Figure 4: Schematic diagram of the flux-consistent downsampling process. The workflow illustrates the transformation of HR and LR image pixels into the celestial coordinate system, the computation of overlapping sky regions between HR and LR receptive fields, and the flux calculation for LR pixels using weighted contributions from overlapping HR pixels.

differences can be observed at the locations of stellar sources. The bilinear interpolation method tends to cause flux reduction when handling bright targets such as stars, making it less suitable for flux consistency astronomical applications.

B PSF Details

We simulate the imaging blur in the STAR dataset using two PSF models: the Gaussian PSF and the Airy PSF [74], aiming to increase training data diversity. The Gaussian PSF is a simple model often used to approximate blur in astronomical observations [75, 76]. In contrast, the Airy PSF captures diffraction effects from a telescope’s circular aperture, making it suitable for space-based instruments like HST [24].

In the Gaussian PSF and Airy PSF models, σ and r serve as adjustable parameters to control the spread of the blur by modulating the energy dispersion of the filter. For instance, in the Gaussian PSF, a larger σ leads to a less concentrated signal with greater energy spread across the filter, while in the Airy PSF, r governs the radial extent of energy distribution due to diffraction, as defined below.

$$\text{PSF}_{\text{Gaussian}}(x, y) = \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right), \quad (11)$$

and

$$\text{PSF}_{\text{Airy}}(r) = \left[\frac{2J_1(kr)}{kr}\right]^2. \quad (12)$$

We define these parameters based on the telescope’s observed blur characteristics, following Schawinski et al. [51], who used the observed blur to set the PSF parameters for a realistic simulation of hardware-specific degradation effects. Accordingly, we set the Gaussian PSF parameter $\sigma \in [0.8, 1.2]$ and the Airy PSF radius $r \in [1.9, 2.2]$ pixels based on the FWHM [60], which measures the blur width at half its peak intensity, to approximate the actual HST WFC/ACS F814W filter observations where the blur is characterized by its FWHM. This enables effective super-resolution training.

C Additional Experiments with Gaussian + Airy PSFs

The original submission focuses on experiments using Gaussian PSF data. Here, we further evaluate the combination of Gaussian PSF and Airy PSF (Gaussian + Airy PSFs) and validate the effectiveness

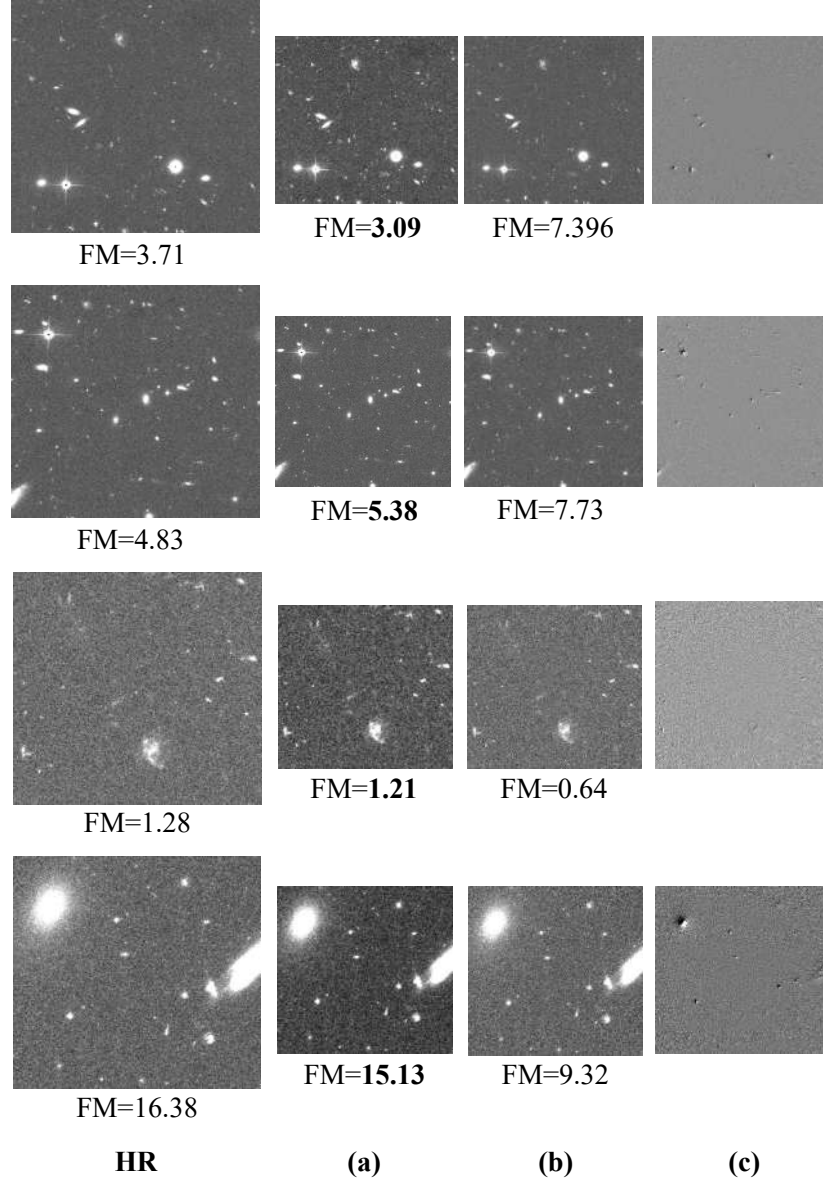


Figure 5: Comparison between downsampling methods. Each HR patch is shown alongside three columns: (a) flux-consistent downsampling, (b) bilinear interpolation, and (c) their pixel-wise difference. FM means flux mean.

Table 6: Performance of different methods under $\times 2$ super-resolution with Gaussian PSF and Airy PSF on the STAR dataset. Metrics: PSNR \uparrow , SSIM \uparrow , Flux Error (FE) \downarrow .

Metric	Bicubic	EDSR	RCAN	SwinIR	FIRS
PSNR	29.4434	35.7398	37.4639	37.1347	38.2678
SSIM	0.7125	0.8086	0.8277	0.8279	0.8334
FE	4.286	1.3249	0.7451	0.7593	0.5585

Table 7: Performance of different methods under $\times 2$ super-resolution with Gaussian PSF and Airy PSF on the STAR dataset (with and without FCL). Metrics: PSNR \uparrow , SSIM \uparrow , Flux Error (FE) \downarrow .

Flux Loss	Metric	EDSR	RCAN	SwinIR
w/o	PSNR	35.7398	37.4639	37.1347
	SSIM	0.8086	0.8277	0.8279
	FE	1.3249	0.7451	0.7593
w/	PSNR	35.8921	37.8914	37.6049
	SSIM	0.8092	0.8286	0.8281
	FE	1.242	0.5914	0.6767

of Flux-Consistent Loss (FCL) in this setting. In this setting, each image is degraded by randomly selecting either the Gaussian or Airy PSF with equal probability.

We compare the performance of different methods under $\times 2$ super-resolution with Gaussian PSF and Airy PSF on the STAR dataset, analyzing the results model-wise and loss-wise. Tab. 6 compares the performance of all methods in this setting. FIRS surpasses baselines like SwinIR and RCAN, achieving a 3.05% higher PSNR and 26.45% lower FE than SwinIR, demonstrating its superior ability to recover fine stellar details and preserve flux accuracy in astronomical image super-resolution. Additionally, Tab. 7 compares EDSR, RCAN, and SwinIR with and without FCL to focus on the impact of FCL across baseline methods. For instance, SwinIR with FCL improves PSNR by 1.27% and reduces FE by 10.88% compared to the version without FCL, while RCAN with FCL improves PSNR by 1.14% and reduces FE by 20.63%, highlighting FCL’s role in enhancing image quality and flux preservation.

D Additional visualizations

We present additional visualizations to demonstrate the effectiveness of our approach in star-field super-resolution (ASR) tasks. Fig. 6 displays the $\times 2$ super-resolution results for the Gaussian PSF experiment, comparing baselines (EDSR, RCAN, PromptIR, SwinIR, HAT) against our FIRS model. The visualizations reveal that FIRS consistently outperforms all baselines, achieving superior visual quality with finer stellar details and sharper structures. To further quantify these improvements, we compute the KL divergence and JS divergence between the intensity distributions of the predicted and ground truth values in selected regions, following the experimental settings in the original submission. The results show that FIRS significantly reduces distribution discrepancies compared to SwinIR and HAT, confirming its superior capability in preserving stellar details and flux accuracy in ASR tasks.

E Hyperparameters Tuning

We tune the parameter λ to balance the Flux-Consistent Loss (FCL) and reconstruction loss in our star-field super-resolution (ASR) model. We evaluate different λ values (0.1, 0.05, and 0.01) under the $\times 2$ Gaussian PSF + Airy PSF setting, with results shown in Tab. 17. The performance metrics show that $\lambda = 0.01$ yields the best results, improving PSNR by 1.40% and reducing FE by 15.88% compared to $\lambda = 0.1$. These results indicate that a proper λ matters in the balance between reconstruction loss and the Flux-Consistent Loss. Fortunately, 0.01 seems to perform well in most cases.

F Experimental setting/details

We ensure reproducibility by providing the experimental environment and computational resources. Tab. 8 shows the environment configuration, including hardware and software details. Tab. 9 summarizes the computational resources used for training. For detailed training settings and parameters of each model, please see the code.

Table 8: Experimental Environment Setup.

Component	Version
OS	Ubuntu 20.04.5 LTS
Python	3.10.15
PyTorch	2.0.0
CUDA	11.8

Table 9: Computational Resources for Different Methods (Training Time in Hours).

Method	Training Time (Hours)
EDSR	52
RCAN	40
Hat	70
SwinIR(light weight)	14
PromptIR	15
GAN	27
FISR (ours)	15

G Additional Experiments

To further validate the robustness and scientific utility of our proposed dataset and model, we conducted a series of additional experiments in response to reviewer feedback. These experiments evaluate the model’s generalization capabilities across different domains, its performance on downstream scientific tasks, its robustness to noise, and its computational efficiency.

G.1 Generalization and Robustness Analysis

Cross-Filter Generalization: To test the model’s performance on data from different instrumental filters, we evaluated our F814W-trained model on test sets from the F606w and F435w filters of the Hubble Space Telescope (HST). As shown in Tab. 10, while there is a performance drop as the filter domain shifts further from the training domain (F814W), the model maintains strong performance, demonstrating satisfactory generalization capabilities. The F606w filter, being spectrally closer to F814W, yields better results than the more distant F435w filter, confirming that domain similarity influences performance.

Table 10: Cross-filter generalization performance of the FISR model trained on the F814W filter.

Metric	F435w	F606w	F814w (In-Domain)
PSNR	35.9192	36.3522	37.8779
SSIM	0.7305	0.7667	0.8311
Flux Error	0.9193	0.8242	0.5739

Robustness to Noise: We evaluated FISR’s robustness by introducing random Poisson noise to each image during inference, simulating realistic observational noise. The results in Tab. 11 show that FISR maintains its state-of-the-art performance, achieving the best results across all metrics compared to other methods under noisy conditions.

Cross-Dataset Evaluation: Although direct evaluation is challenging due to differences in data units (STAR uses scientific counts, while AstroSR uses RGB), we re-trained our FISR model on the

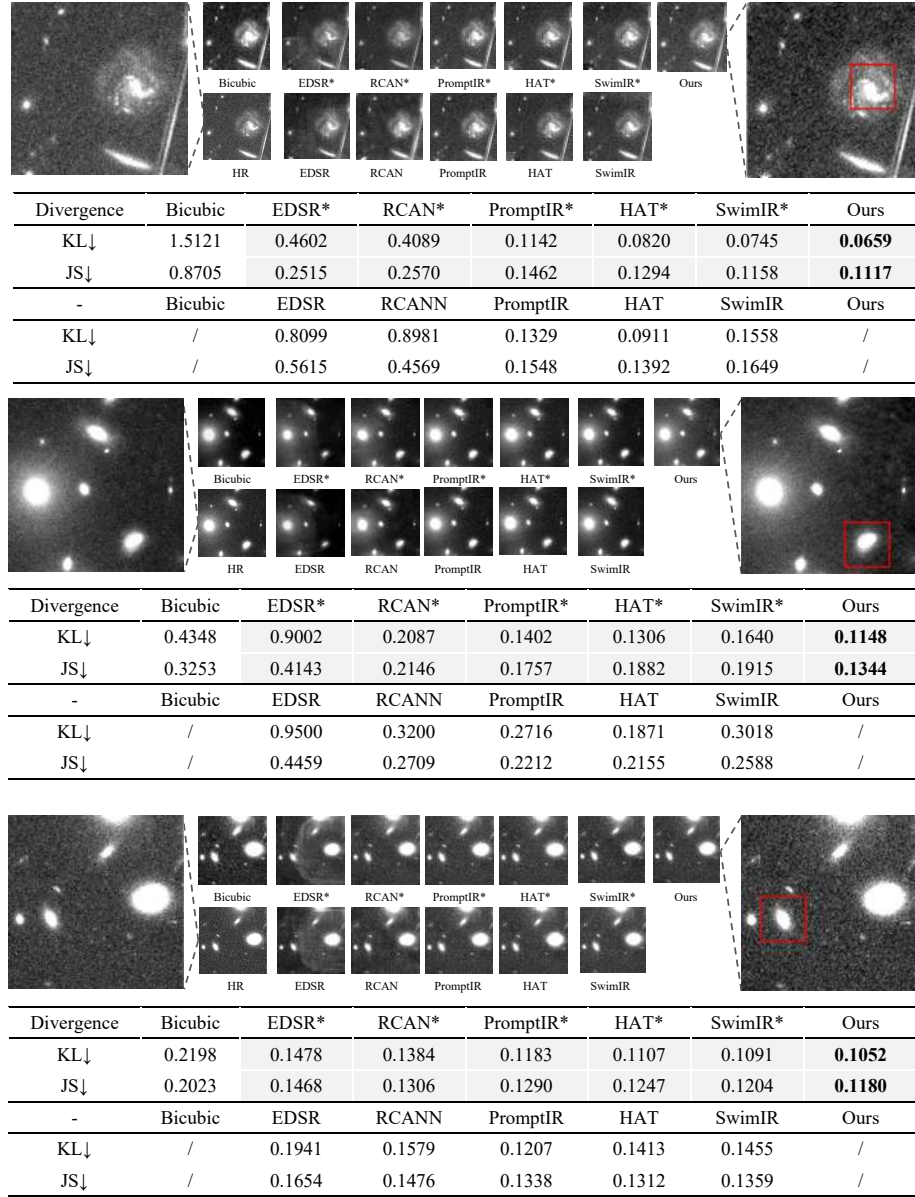


Figure 6: We further demonstrate several sets of visualization results on the $\times 2$ Gaussian PSF experiment. Models with (*) are trained using FCL.

Table 11: Performance comparison under Poisson noise injection during inference. Best results are in **bold**.

Method	Bicubic	EDSR	SwinIR	RCAN	HAT	RealESRGAN	FISR (Ours)
PSNR	28.9823	34.8191	36.3945	35.2918	36.8342	35.8098	36.7803
SSIM	0.6825	0.7684	0.7883	0.7848	0.7743	0.7852	0.7888
Flux Error	4.7889	1.5682	1.1433	1.1993	1.1943	6.9292	1.1025

AstroSR dataset. Tab. 12 demonstrates that our method outperforms the original baseline models reported in the AstroSR paper, showcasing its architectural effectiveness on different data types.

Table 12: Performance comparison on the AstroSR dataset after re-training. Best results are in **bold**.

Method	Bicubic	EDSR	RCAN	ENLCA	SRGAN	FISR (Ours)
PSNR	17.7714	23.2168	23.6082	23.4267	23.0039	24.0211
SSIM	0.1686	0.3910	0.3966	0.3963	0.3854	0.4025
Flux Error	233.2564	50.5872	61.3863	59.1659	42.3078	33.2331

G.2 Evaluation on Downstream Scientific Tasks

To quantify the practical impact of our super-resolution model on real-world scientific analysis, we evaluated its performance on four representative downstream astronomical tasks. These experiments are designed to demonstrate that improvements in standard metrics like PSNR, SSIM, and our proposed Flux Error (FE) directly translate to higher fidelity in scientific measurements. The methodologies and results for these tasks are detailed below, with a final comparative summary in Table 13.

G.3 Evaluation on Downstream Scientific Tasks

To quantify the practical impact of our super-resolution model on real-world scientific analysis, we evaluated its performance on two representative downstream astronomical tasks. These experiments are designed to demonstrate that improvements in standard metrics and our proposed Flux Error (FE) directly translate to higher fidelity in scientific measurements. The methodologies and results for these tasks are detailed below.

Object Detection Sensitivity: The ability to detect faint objects is fundamental to astronomical surveys, determining the depth and completeness of celestial catalogs. An effective SR model should enhance faint sources, thereby improving detection sensitivity. In our experiment, we performed bipartite matching between sources detected in the predicted images and the ground-truth catalog, with a match considered successful if the spatial distance was within 2 pixels. The sensitivity was quantified using the **Recall** metric. Our FISR model achieves a high recall of **81.47%**, indicating strong performance in identifying celestial objects.

Distance Estimation: Accurately measuring the distances to celestial objects is a cornerstone of cosmology, combining both object detection and precise photometry. To evaluate this, we used the successfully matched object pairs from the detection task. We converted each object’s flux to an apparent magnitude (m) and then applied the distance modulus formula, $d = 10^{(m-M+5)/5}$, to estimate the distance (d) in megaparsecs (MPC), assuming a constant absolute magnitude (M) of 4.83 (typical for Sun-like stars). The accuracy was evaluated by the Mean Absolute Error (MAE) between the predicted and ground-truth distances, with the results shown in Table 13.

Table 13: Evaluation on the downstream task of distance estimation. Lower values indicate better performance. Best results are in **bold**.

Metric	Bicubic	SwinIR	EDSR	RCAN	HAT	R-ESRGAN	FISR
Distance MAE (MPC)	6.82E+03	5.37E+03	6.44E+03	5.61E+03	4.44E+03	4.89E+03	4.12E+03

H More Ablation Studies on FGG Module

We performed ablation studies to analyze the sensitivity of the Flux Guidance Generation (FGG) module.

Kernel Choice in FGG: We tested alternative kernels (Airy, and a random mix of Gaussian/Airy) for rendering the flux map. Tab. 14 shows that performance remains stable across different kernel choices, suggesting that the module’s primary function is to provide a spatial prior for flux information, rather than depending on a specific kernel formulation.

Table 14: Ablation study on the kernel choice within the FGG module.

Kernel Type	PSNR	SSIM	Flux Error
Gaussian	37.8779	0.8311	0.5739
Airy	37.6988	0.8305	0.5664
Gaussian/Airy (Random)	37.8186	0.8311	0.5726

Sensitivity to Detection Errors: To assess FGG’s robustness, we introduced noisy detections by lowering the source detection threshold, resulting in twice the number of sources, including many false positives. As seen in Tab. 15, while performance degrades slightly, FISIR remains robust and achieves results comparable to the model trained with clean detections. This indicates that the model’s performance does not solely depend on the precision of the FGG’s input.

Table 15: Performance of FISIR with clean versus noisy source detections in the FGG module.

Detection Quality	PSNR	SSIM	Flux Error
Clean Detections	37.8779	0.8311	0.5739
Noisy Detections	37.3176	0.8275	0.6872

I Computational Efficiency

We measured the single-image inference time for all compared methods. The results in Tab. 16 show that FISIR is computationally efficient, with an inference time comparable to other high-performing transformer-based models like SwinIR.

Table 16: Inference time per image (in seconds) for various SR methods.

Method	Bicubic	EDSR	SwinIR	RCAN	HAT	RealESRGAN	FISIR (Ours)
Time (s)	0.0014	0.1908	0.1088	0.1237	0.6747	0.0995	0.1698

J Limitations and future work

While our study offers promising insights, it has a few limitations that merit further exploration. First, our experiments are based on observations from a single telescope, the HST WFC/ACS with the F814W filter, which may limit the generalizability of our findings to other instruments or observational contexts. Additionally, although our network design performs well, it could benefit from incorporating more domain-specific optimizations rooted in astronomical knowledge, such as leveraging physical principles or astronomical priors to enhance performance in complex scenarios. These areas present opportunities for future refinement. Looking forward, we aim to broaden the applicability of our method by extending it to a wider array of advanced telescopes, such as the James Webb Space Telescope (JWST) [72] or the upcoming Large Synoptic Survey Telescope (LSST) [73], to explore its potential across diverse astronomical contexts. Furthermore, we plan to enhance our network design by integrating more astronomy-driven optimizations, incorporating physical knowledge and astronomical priors to better address challenges like crowded stellar regions or variable noise conditions. Through these efforts, we hope to make modest contributions to the

field of astronomical image processing, fostering the development of more robust and adaptable tools for future discoveries.

Table 17: Ablation study on the penalty factor λ ($\times 2$ on Gaussian PSF + Airy PSF).

FCL Weight λ	PSNR\uparrow	SSIM\uparrow	FE\downarrow
0.1	37.0843	0.8198	0.7842
0.05	37.2672	0.8252	0.7064
0.01	37.6049	0.8281	0.6767

K More Visualizations of the STAR Dataset

To further illustrate the unique characteristics and scale of the STAR benchmark, this section provides additional visualizations of the source data. We present examples of the original, full-frame observational images from the Hubble Space Telescope (HST) WFC/ACS instrument, which constitute the raw data prior to the patch subdivision process for model training 7.

These wide-field views underscore a core advantage of STAR over previous object-centric datasets. Instead of focusing on isolated, cropped targets, our dataset provides a holistic view of extensive celestial regions, preserving the crucial spatial context and inter-object relationships (e.g., cross-object interaction, weak lensing). Furthermore, we showcase a gallery of selected image patches to highlight the rich diversity within STAR 8. These examples span a wide range of astronomical environments, from dense, crowded stellar fields and sparsely populated regions to complex nebulae and fields containing multiple galaxies. Collectively, these visualizations reinforce the value of STAR as a comprehensive and physically faithful benchmark for advancing astronomical super-resolution research.

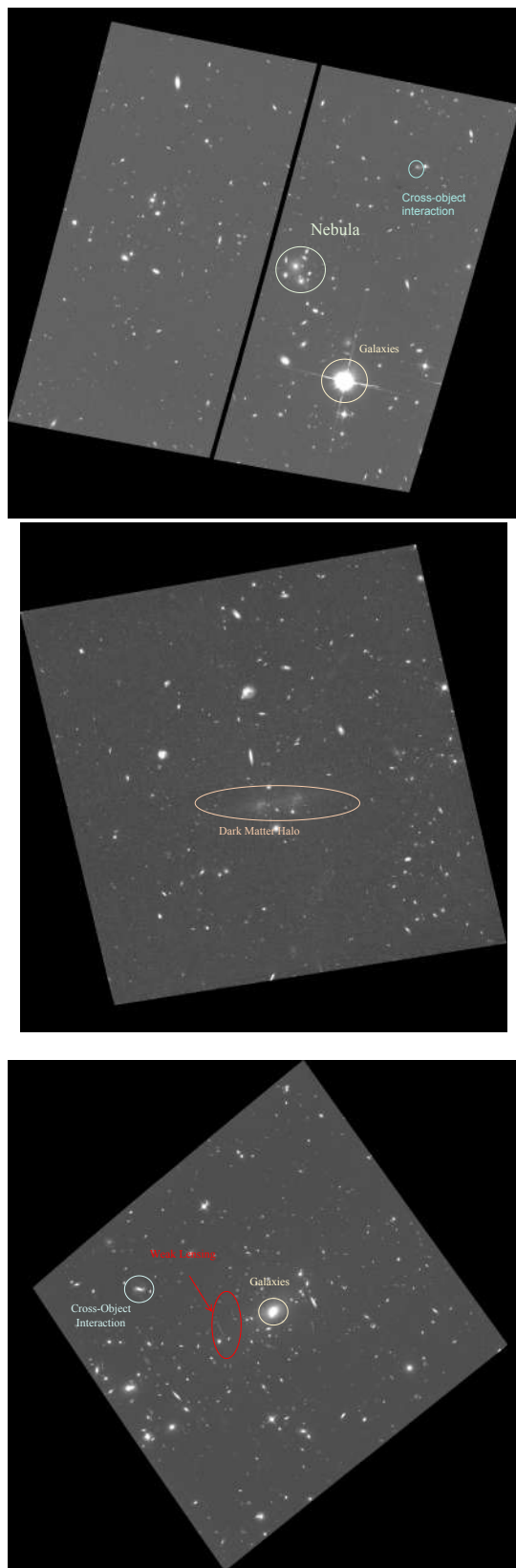


Figure 7: Examples of the original wide-field raw data from the HST WFC/ACS survey, which form the basis of the STAR dataset.

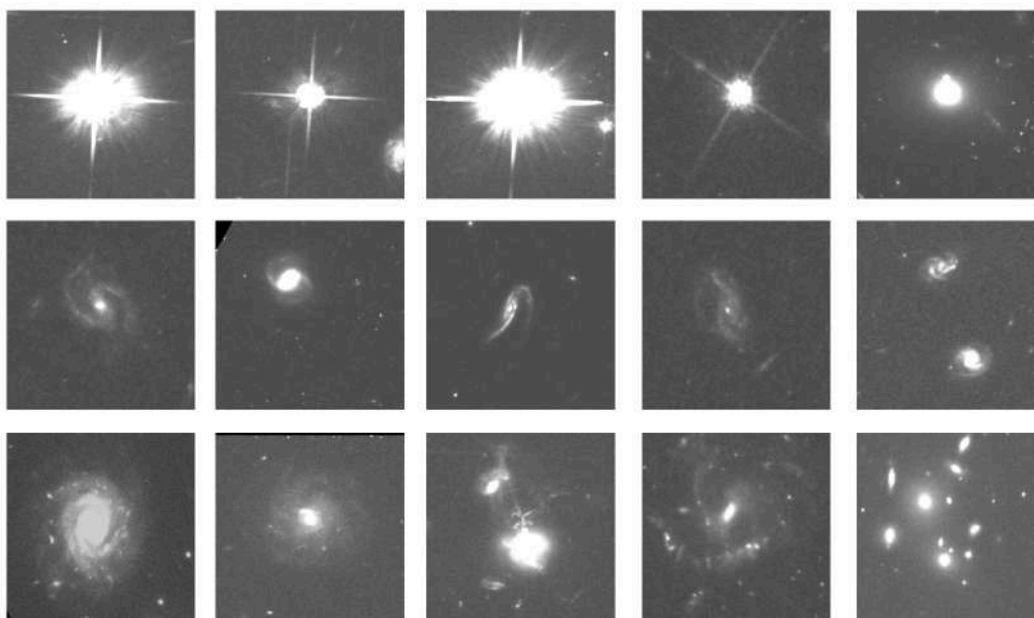


Figure 8: A selection of patches from the STAR dataset, showcasing its diversity. The examples include crowded stellar fields, regions with interacting galaxies, and complex nebulae, demonstrating the variety of astronomical scenes available for training robust models.