STAR: A Benchmark for Astronomical Star Fields Super-Resolution

Appendix

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

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 IR 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. 1. 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:

$$F_i^{LR} = \sum_{j \in S_i^o} w_{i,j} \cdot f_j^{HR}, \tag{1}$$

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:

$$w_{i,j} = \frac{A_{i,j}}{A_j^{HR}}, \tag{2}$$

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 (1), 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:

$$w_{i,j} \cdot f_j^{HR} = \left(\frac{A_{i,j}}{A_j^{HR}} \right) \cdot f_j^{HR} = A_{i,j} \cdot \frac{f_j^{HR}}{A_j^{HR}}. \tag{3}$$

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. 2, we compare flux consistency downsampling with traditional bilinear interpolation. It can be found that the result of Fig. 2 (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. 2 (c). Noticeable

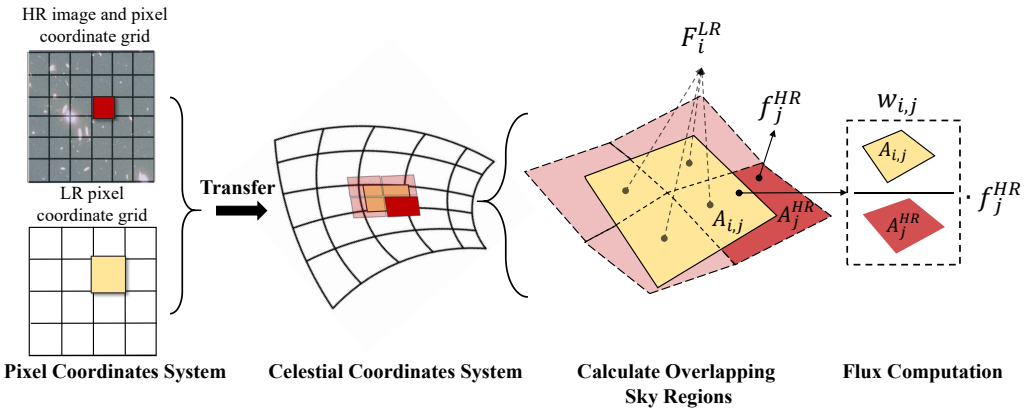


Figure 1: Schematic diagram of the flux-consistent downsampling process. The workflow illustrates the transformation of HR and LR image pixels into the celestial coordinate system, the computation of overlapping sky regions between HR and LR receptive fields, and the flux calculation for LR pixels using 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 [**?**], aiming to increase training data diversity. The Gaussian PSF is a simple model often used to approximate blur in astronomical observations [1, 2]. In contrast, the Airy PSF captures diffraction effects from a telescope's circular aperture, making it suitable for space-based instruments like HST [3].

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.

$$\mathrm{PSF}_{\mathrm{Gaussian}}(x, y) = \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right), \tag{4}$$

and

$$\mathrm{PSF}_{\mathrm{Airy}}(r) = \left[\frac{2J_1(kr)}{kr}\right]^2. \tag{5}$$

We define these parameters based on the telescope's observed blur characteristics, following Schawinski et al. [4], 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 $\sigma \in [\mathbf{0.8}, \mathbf{1.2}]$ and the Airy PSF radius $\mathbf{r} \in [\mathbf{1.9}, \mathbf{2.2}]$ pixels based on the FWHM [5], 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

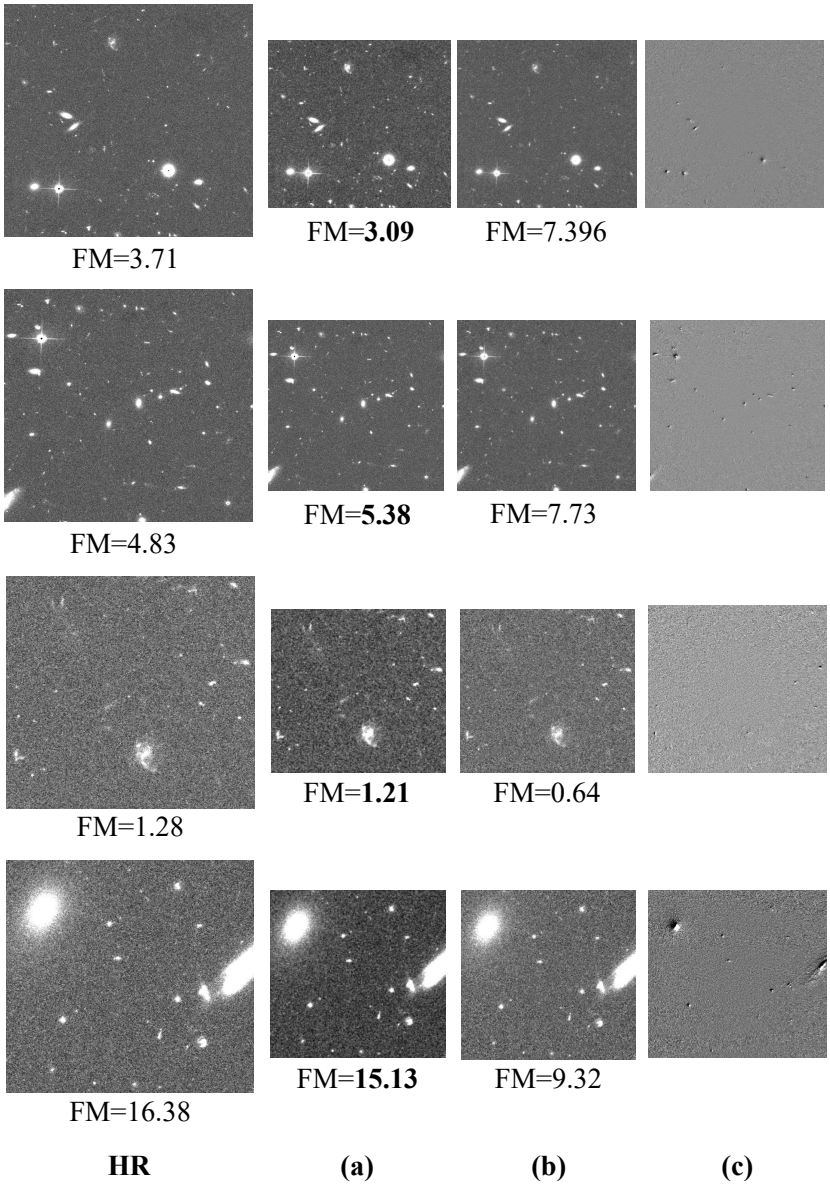


Figure 2: Comparison between downsampling methods. Each HR patch is shown alongside three columns: (a) flux-consistent downsampling, (b) bilinear interpolation, and (c) their pixel-wise difference. FM means flux mean.

Table 1: Performance of different methods under $\times 2$ super-resolution with Gaussian PSF and Airy PSF on the STAR dataset. Metrics: PSNR↑, SSIM↑, Flux Error (FE)↓.

Metric	Bicubic	EDSR	RCAN	SwinIR	FISR
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 2: 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↑, SSIM↑, Flux Error (FE)↓.

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. 1 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. 2 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. 3 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. 12. 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. 3 shows the environment configuration, including hardware and software details. Tab. 4 summarizes the computational resources used for training. For detailed training settings and parameters of each model, please see the code.

Table 3: Experimental Environment Setup.

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

Table 4: 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

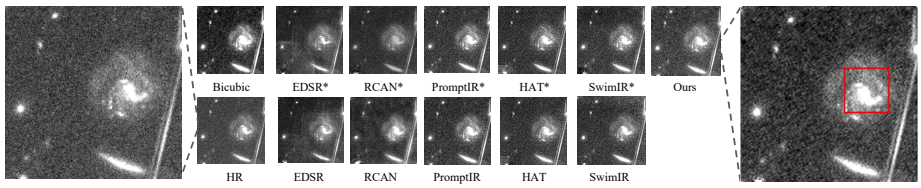
Cross-Filter Generalization: To test the model's performance on data from different instrumental filters, we evaluated our F814W-trained model on test sets from the F606w and F435w filters of the Hubble Space Telescope (HST). As shown in Tab. 5, while there is a performance drop as the filter domain 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 5: Cross-filter generalization performance of the FISR model trained on the F814W filter.

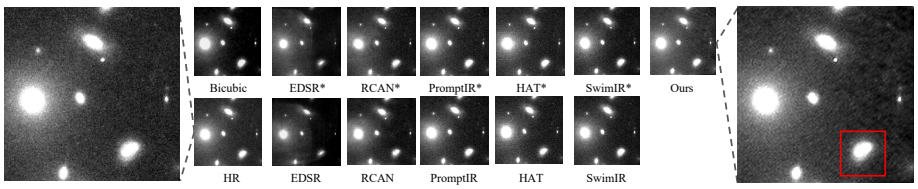
Metric	F435w	F606w	F814w (In-Domain)
PSNR	35.9192	36.3522	37.8779
SSIM	0.7305	0.7667	0.8311
Flux Error	0.9193	0.8242	0.5739

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. 6 show that FISR maintains its state-of-the-art performance, achieving the best results across all metrics compared to other methods under noisy conditions.

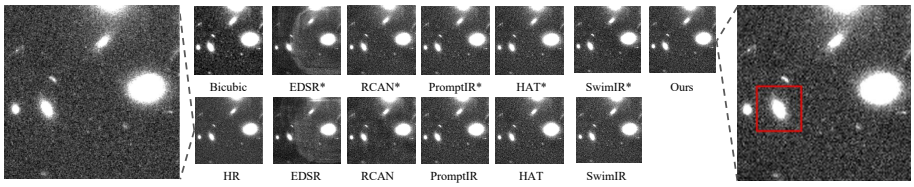
Cross-Dataset Evaluation: Although direct evaluation is challenging due to differences in data units (STAR uses scientific counts, while AstroSR uses RGB), we re-trained our FISR model on



Divergence	Bicubic	EDSR*	RCAN*	PromptIR*	HAT*	SwimIR*	Ours
KL↓	1.5121	0.4602	0.4089	0.1142	0.0820	0.0745	**0.0659**
JS↓	0.8705	0.2515	0.2570	0.1462	0.1294	0.1158	**0.1117**
-	Bicubic	EDSR	RCANN	PromptIR	HAT	SwimIR	Ours
KL↓	/	0.8099	0.8981	0.1329	0.0911	0.1558	/
JS↓	/	0.5615	0.4569	0.1548	0.1392	0.1649	/



Divergence	Bicubic	EDSR*	RCAN*	PromptIR*	HAT*	SwimIR*	Ours
KL↓	0.4348	0.9002	0.2087	0.1402	0.1306	0.1640	**0.1148**
JS↓	0.3253	0.4143	0.2146	0.1757	0.1882	0.1915	**0.1344**
-	Bicubic	EDSR	RCANN	PromptIR	HAT	SwimIR	Ours
KL↓	/	0.9500	0.3200	0.2716	0.1871	0.3018	/
JS↓	/	0.4459	0.2709	0.2212	0.2155	0.2588	/



Divergence	Bicubic	EDSR*	RCAN*	PromptIR*	HAT*	SwimIR*	Ours
KL↓	0.2198	0.1478	0.1384	0.1183	0.1107	0.1091	**0.1052**
JS↓	0.2023	0.1468	0.1306	0.1290	0.1247	0.1204	**0.1180**
-	Bicubic	EDSR	RCANN	PromptIR	HAT	SwimIR	Ours
KL↓	/	0.1941	0.1579	0.1207	0.1413	0.1455	/
JS↓	/	0.1654	0.1476	0.1338	0.1312	0.1359	/

Figure 3: We further demonstrate several sets of visualization results on the $\times 2$ Gaussian PSF experiment. Models with (*) are trained using FCL.

Table 6: 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**

the AstroSR dataset. Tab. 7 demonstrates that our method outperforms the original baseline models reported in the AstroSR paper, showcasing its architectural effectiveness on different data types.

Table 7: 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 8.

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 8.

Table 8: 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. 9 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 9: 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. 10, while performance degrades slightly, FISR 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 10: Performance of FISR 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. 11 show that FISR is computationally efficient, with an inference time comparable to other high-performing transformer-based models like SwinIR.

Table 11: Inference time per image (in seconds) for various SR methods.

Method	Bicubic	EDSR	SwinIR	RCAN	HAT	RealESRGAN	FISR (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) [6] or the upcoming Large Synoptic Survey Telescope (LSST) [7], 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 12: Ablation study on the penalty factor λ ($\times 2$ on Gaussian PSF + Airy PSF).

FCL Weight λ	**PSNR↑**	**SSIM↑**	**FE↓**
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 4.

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 5. 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.

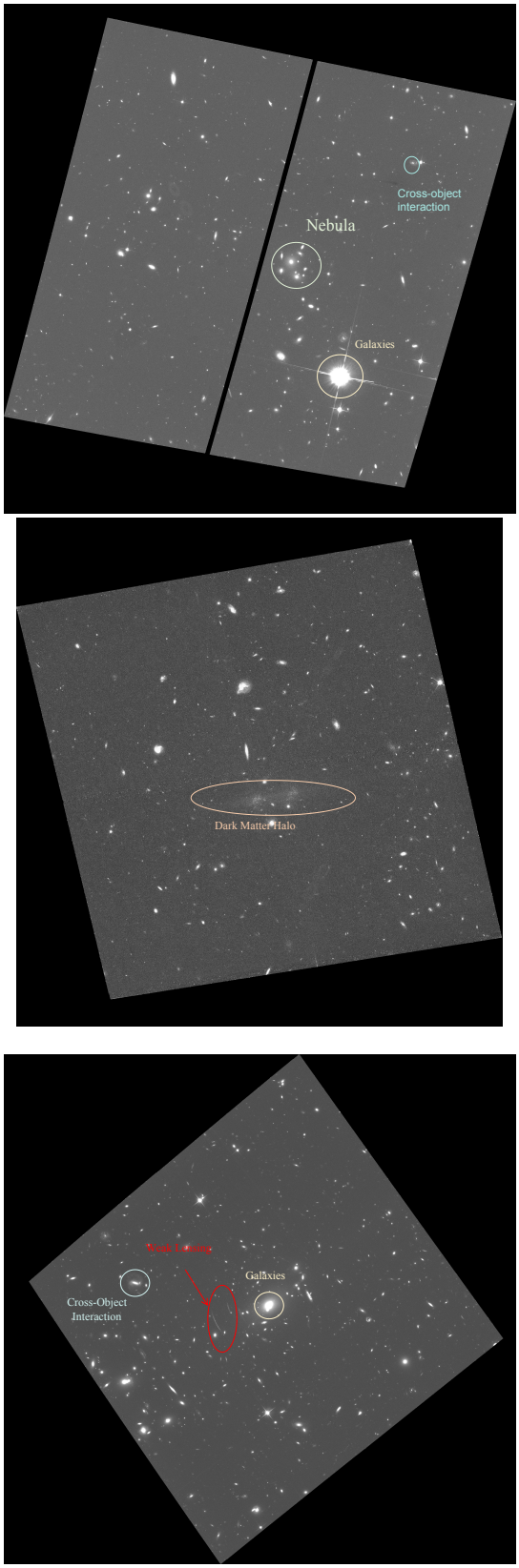


Figure 4: Examples of the original wide-field raw data from the HST WFC/ACS survey, which form the basis of the STAR dataset.

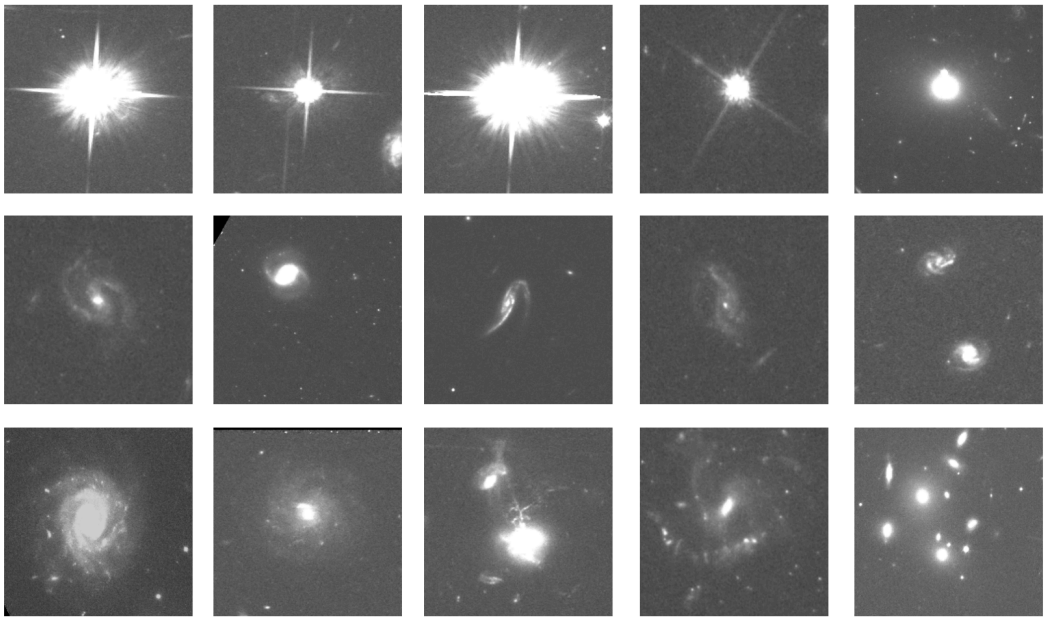


Figure 5: A selection of patches from the STAR dataset, showcasing its diversity. The examples include crowded stellar fields, regions with interacting galaxies, and complex nebulae, demonstrating the variety of astronomical scenes available for training robust models.

References

[1] John W Hardy. *Adaptive optics for astronomical telescopes*, volume 16. Oxford university press, 1998.

[2] François Roddier. V the effects of atmospheric turbulence in optical astronomy. In *Progress in optics*, volume 19, pages 281–376. Elsevier, 1981.

[3] 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.

[4] 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.

[5] 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.

[6] Klaus M Pontoppidan, Jaclyn Barrientes, Claire Blome, Hannah Braun, Matthew Brown, Margaret Carruthers, Dan Coe, Joseph DePasquale, Néstor Espinoza, Macarena Garcia Marin, et al. The jwst early release observations. *The Astrophysical Journal Letters*, 936(1):L14, 2022.

[7] 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.