Automated galaxy sizes in Euclid images using the Segment Anything Model

J. Vega-Ferrero[1,2,⋆], F. Buitrago[1,3], J. Fernández-Iglesias[4], S. Raji[1], B. Sahelices[4], and H. Domínguez Sánchez[2,5]

[1] Departamento de Física Teórica, Atómica y Óptica, Universidad de Valladolid, 47011 Valladolid, Spain
[2] Centro de Estudios de Física del Cosmos de Aragón (CEFCA), Plaza de San Juan, 1, 44001 Teruel, Spain
[3] Instituto de Astrofísica e Ciências do Espaço, Universidade de Lisboa, OAL, Tapada da Ajuda, PT1349-018 Lisbon, Portugal
[4] GIR GCME. Departamento de Informática, Universidad de Valladolid, 47011 Valladolid, Spain
[5] Instituto de Física de Cantabria (CSIC-UC), Avda. Los Castros s/n, 39005 Santander, Spain

Reporter: Yu Huajian

2025.5.16

Outline

1. Introduction

2. Dataset

3. Stellar disk truncations with SAM

4. Results

5. CONCLUSIONS

1 Introduction

In astrophysics, there are indeed several methods of measuring a galaxy’s size, but all come with significant challenges due to the diffuse and often irregular nature of galaxies, which do not have well-defined boundaries.

Stellar disk truncations, also referred to as galaxy edges, are key indicators of galactic size, determined by the radial location of the gas density threshold for star formation. This threshold essentially marks the boundary of the luminous matter in a galaxy. Accurately measuring galaxy sizes for millions of galaxies is essential for understanding the physical processes driving galaxy evolution over cosmic time.

Determining galaxy truncations is an object segmentation task within the field of computer vision, especially when approached from the perspective of analyzing astronomical images to identify and delineate specific features of galaxies.

Segment Anything Model (SAM1, Kirillov et al. 2023) addresses this challenge by leveraging a large synthetic dataset, trained with a prompt-based approach, enabling it to generalize to unseen objects and tasks without additional training.

In this study, we propose and validate the use of SAM, an AI foundational model for segmentation, to identify stellar disk truncations in Euclid-like images automatically. Subsequently, we infer physically motivated galaxy sizes using the truncations obtained with SAM from a set of approximately one thousand HST galaxies previously studied in Buitrago & Trujillo (2024). We demonstrate the enormous potential of SAM in estimating galaxy sizes for large samples rapidly and accurately, without the need for training the model with previously known truncations or performing any transfer learning or domain adaptation step.

2 Dataset

For our analysis, we utilized the sample of 1047 galaxies described in Buitrago & Trujillo (2024, hereafter BT24). This dataset consists of a redshift- and a mass-selected sample of disk- dominated galaxies.

We used the most up-to-date set of images (v1.0) in the CANDELS survey and in the GOODS-South field from the Hubble Legacy Field. Galaxies were selected according to their stellar masses and spectroscopic redshifts . This target selection was based on the CANDELS public catalogs , the high-quality spectroscopic redshifts from the LEGA-C DR2 redshifts and the DR3 data , the ZCOSMOS Final Data Release , and hCOSMOS .

Galaxies in this dataset were selected to be disk-dominated according to the ML morphological classifications derived by , and split into three categories: pure disks (DISK), disks with central spheroids (DISKSPH), and irregular disks (DISKIRR). In BT24, the authors derive the radial location of the gas density threshold for star formation as a size indicator for the whole dataset using a combination of surface brightness, color, and stellar mass density profiles. We used these stellar truncations and truncation masks as a reference dataset for evaluating our results.

3 Stellar disk truncations with SAM

SAM's architecture includes three main components: an image encoder, a prompt encoder, and a mask decoder. The image encoder is based on a Vision Transformer (ViT) that has been pretrained to handle high- resolution inputs efficiently. The prompt encoder can process various types of prompts, including sparse (points, boxes, text) and dense (masks) ones, facilitating flexible and accurate segmentation outcomes.

3.1 Image preprocessing

For the BT24 dataset, we produced squared stamps with a field of view of 12×12 arcsec2 (i.e., 200×200 px, with a pixel scale of 0.06 px/arcsec) in the F606W (V-band) and the F814W (I-band) from the ACS camera, and the F125W (J-band) and the F160W (H-band) from the WFC3 camera.

Our aim is to test the performance on Euclid data to be ready to apply our methodology for DR1. Therefore, until Euclid's wide-field survey data is released, we have produced mock Euclid ("euclidized") galaxy images by generating composite RGB images using H, J, and I+V HST filters, respectively.

It employs an asinh transformation to assign RGB colors to bright pixels (as was described in Lupton et al. 2004), while the faint ones are shown in an inverse gray scale, as follows:

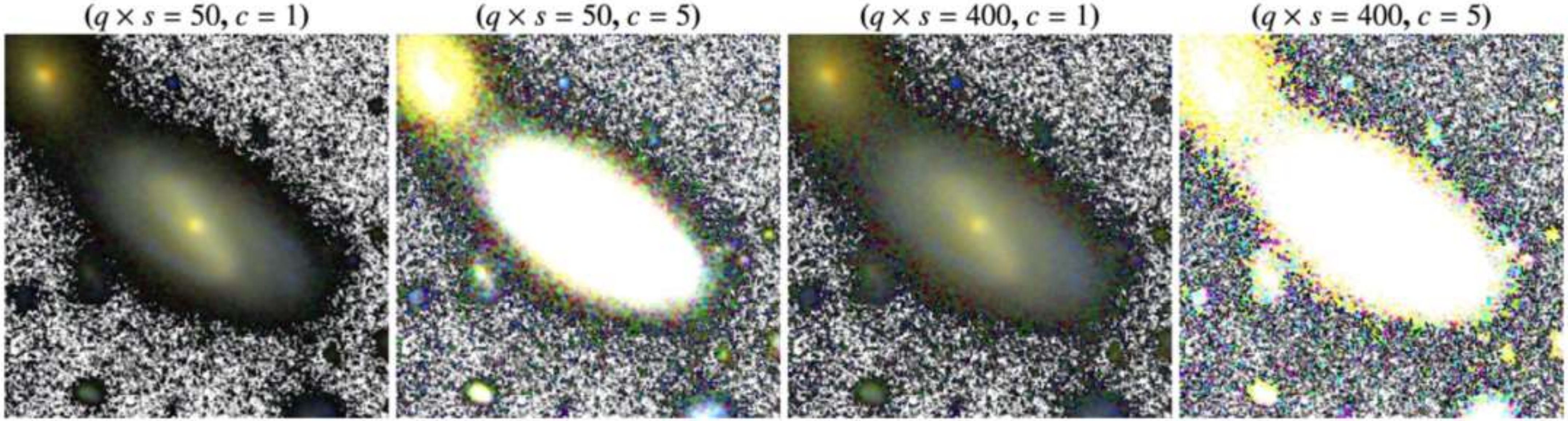
$$f(I) = \frac{a \sin h(q \times s \times I)}{q}$$

where I is the value in each pixel of the image, s is the "stretch" parameter (s → 0, equivalent to linear stretch), and q is the "bright threshold" that controls the coloring of brighter features. When combining the three images in the H, J, and I+V filters into a composite RGB image, the relevant quantity is the product q × s.

3 Stellar disk truncations with SAM

3.2 SAM truncations

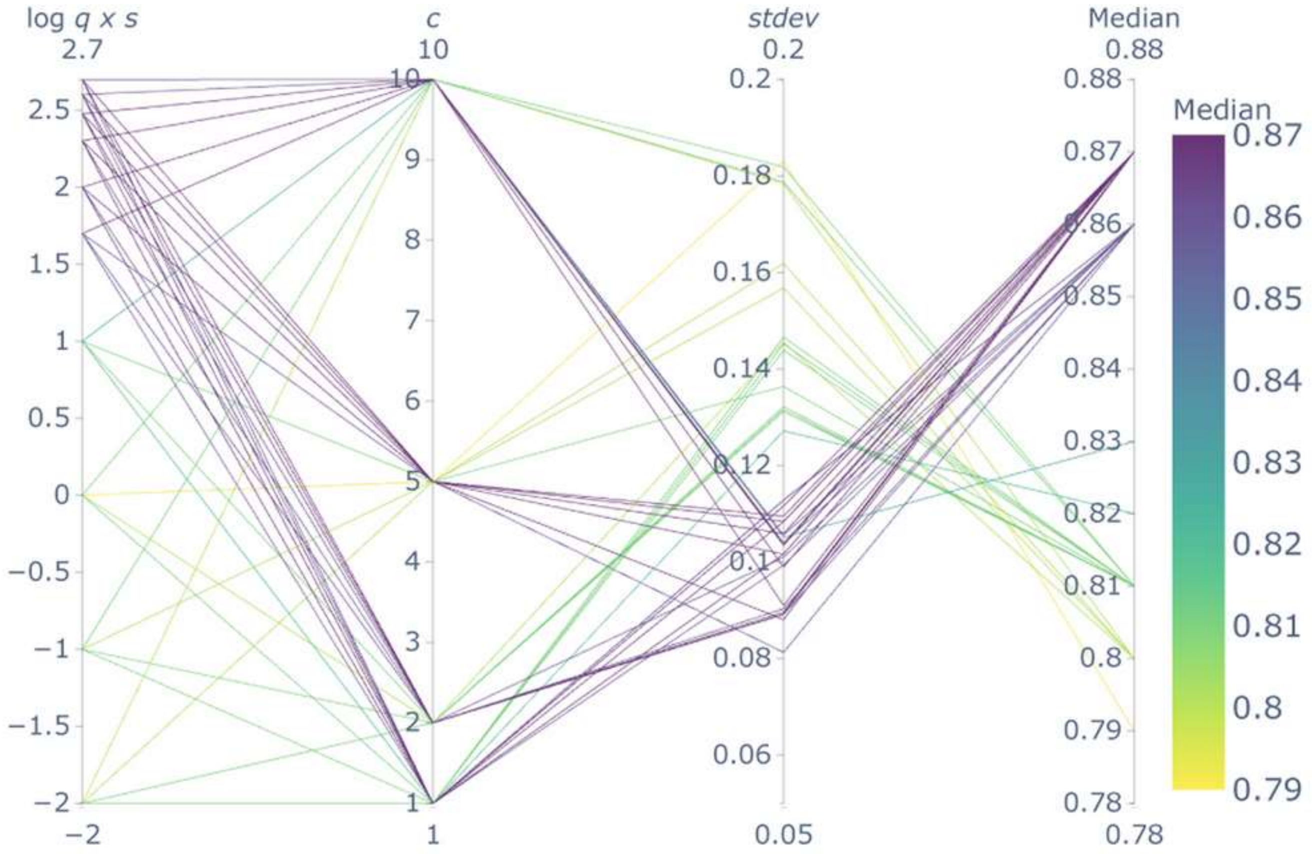
To check the robustness of SAM in retrieving stellar disk truncations, we applied it to the whole dataset of RGB images for a large range of combinations of q × s and c. We generated RGB images for q × s = (0.01, 0.1, 1, 10, 50, 100, 200, 300, 400, 500) and c = (1, 2, 5, 10). In total, we applied SAM over the whole dataset for 40 different input configurations of the input images.



Four RGB images of the DISK galaxy .The RGB channels correspond to the H, J, and I+V HST filters, respectively. Images were produced with the Gnuastro script astscript-color-faint-gray for q × s = 50 (first and second panels) and q × s = 400 (third and fourth panels), and for c = 1 (first and third panels) and c = 5 (second and fourth panels).

3 Stellar disk truncations with SAM

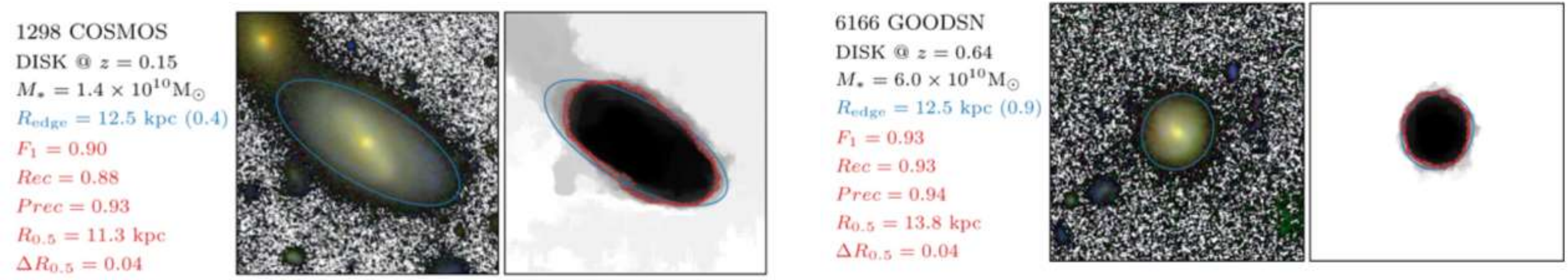
3.2 SAM truncations



Parallel coordinates plot for the 40 input configurations combining q × s (in log scale, first coordinate), c (second coordinate), standard deviation of the F1 score (third coordinate), and median values of the F1 score (fourth coordinate). Each configuration is color-coded according to the median F1 score. All the configurations with log(q × s) $\gtrsim$ 1.7 (q × s $\geqslant$ 50) (independently of the contrast value, c) lead to the largest median values of the F1 score with the smallest values of the standard deviation of the F1 score.

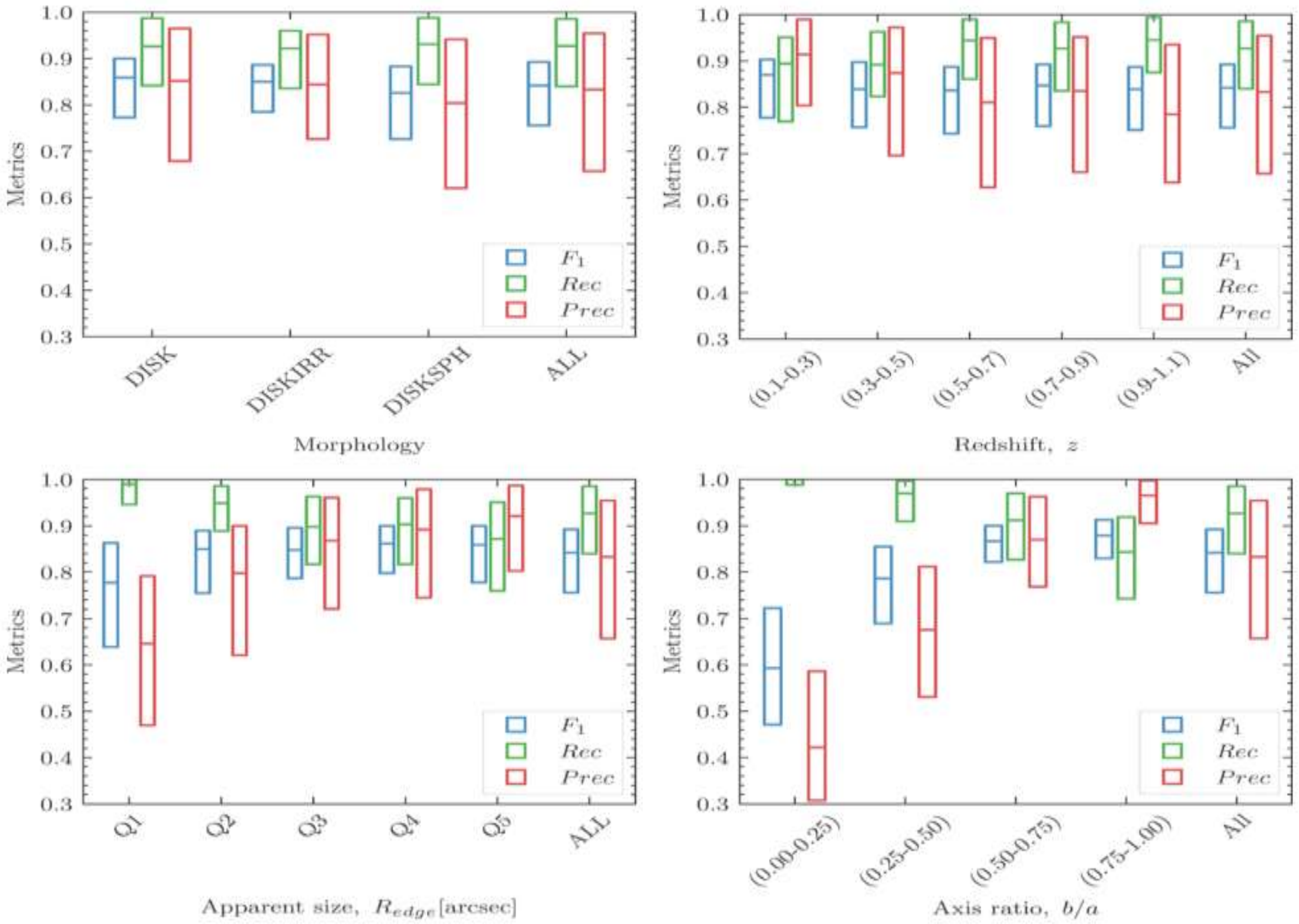
3 Stellar disk truncations with SAM

3.2 SAM truncations



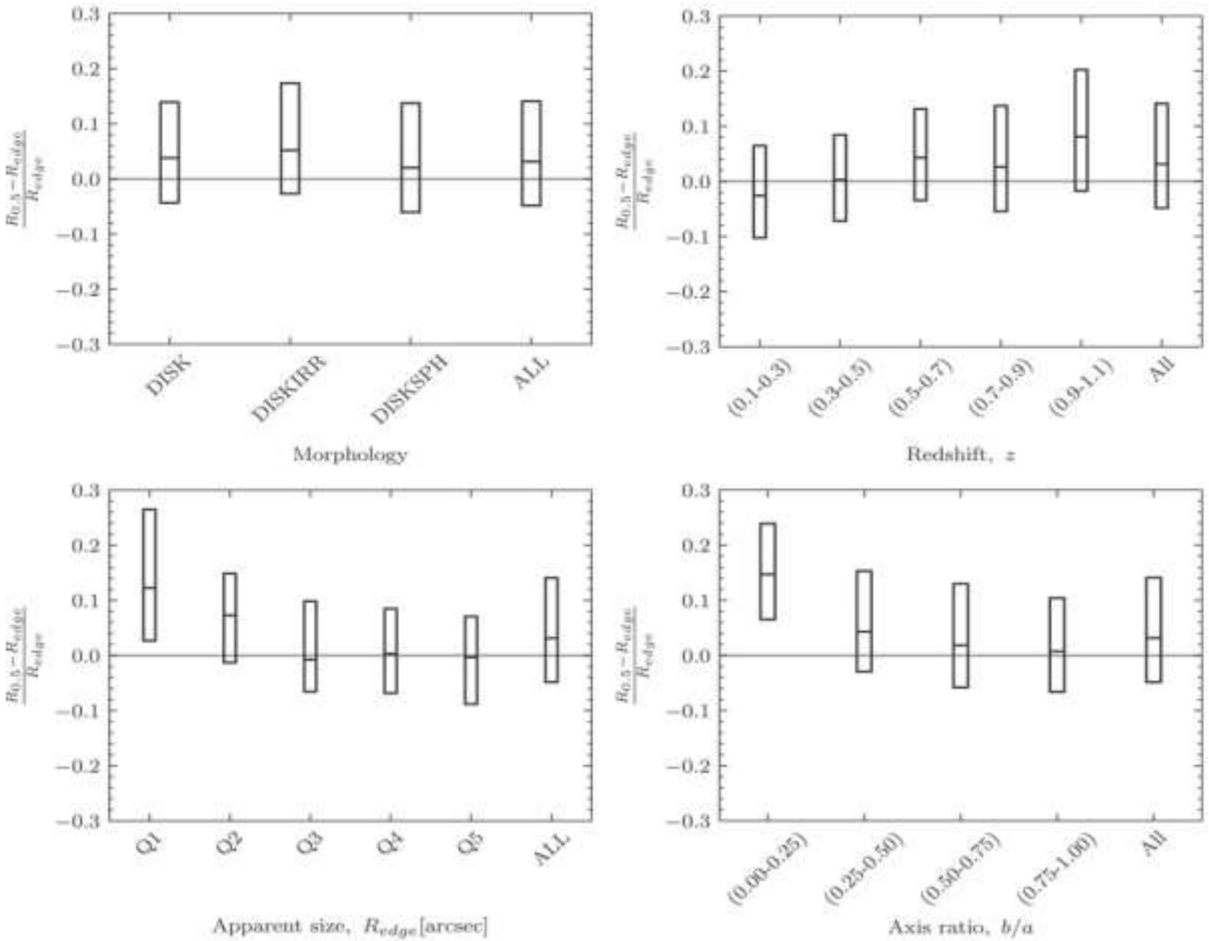
For each galaxy, we show an RGB image (left-hand panel) of the galaxy and the averaged segmentation map inferred with SAM (right-hand panel). The averaged segmentation map shows the SAM truncations (in gray) for different configurations of the input image, where darker regions indicate stronger agreement between different truncation estimates. The solid red contours depict the SAM truncation based on the majority voting criteria (above a 0.5 threshold), while the dashed red contours correspond to the regions above a 0.7 probability threshold. The blue ellipse corresponds to the truncation derived by BT24 parameterized by the semi-major axis (denoted as Redge) and the axis ratio, b/a (shown between brackets after the Redge value). We indicate in red the different metrics (F1, Rec, and Prec) when comparing the SAM and the BT24 truncation estimates, the size of the truncation (R0.5), and the relative error in the size of the truncation and the contour level at a 0.7 probability, denoted as ΔR0.5.

4 RESULTS



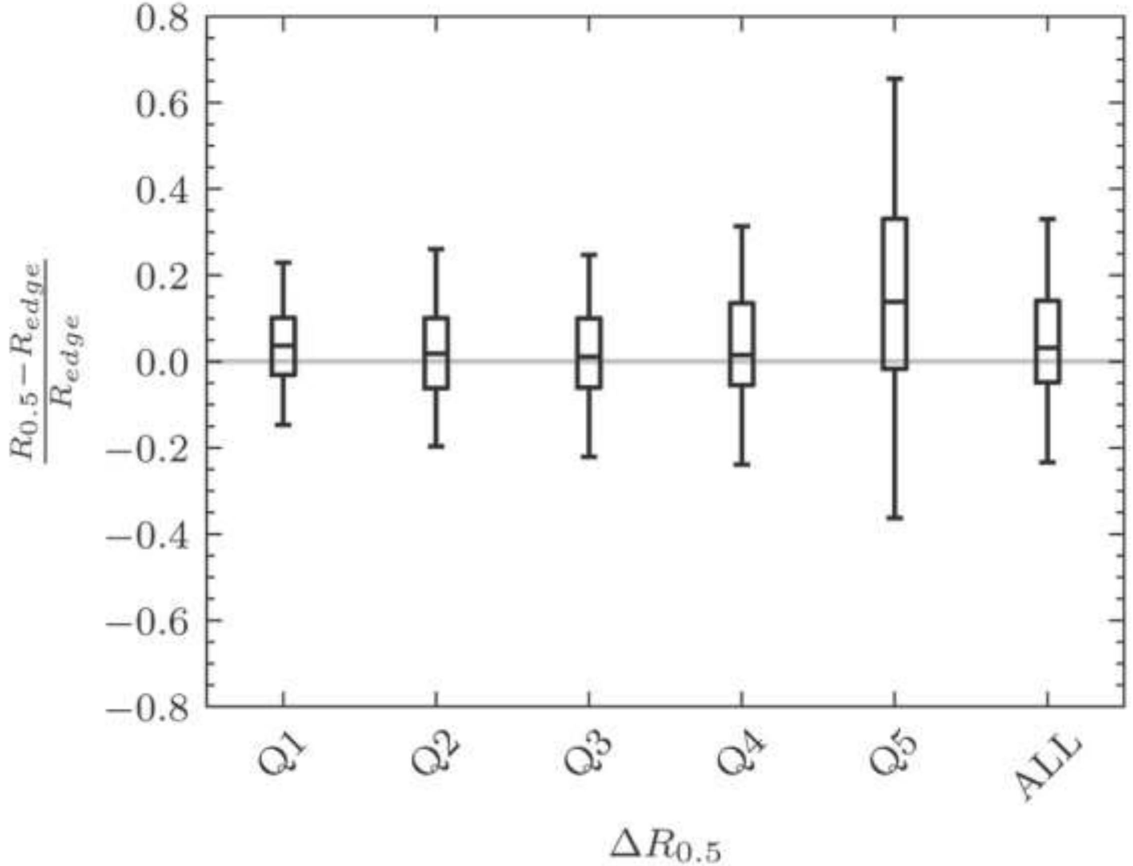
From left to right, SAM's performance as a function of: morphological type, redshift, apparent size (Redge , semi-major axis of the BT24 truncation) split in quintiles, and axis ratio of the BT24 truncation. Metrics are shown in different colors: F1 in blue, Rec in green, and Prec in red. Boxes represent the IQR, while the horizontal lines indicate the median value. F1 and Prec values decrease, while keeping high Rec values, for small and/or elongated galaxies.

4 RESULTS



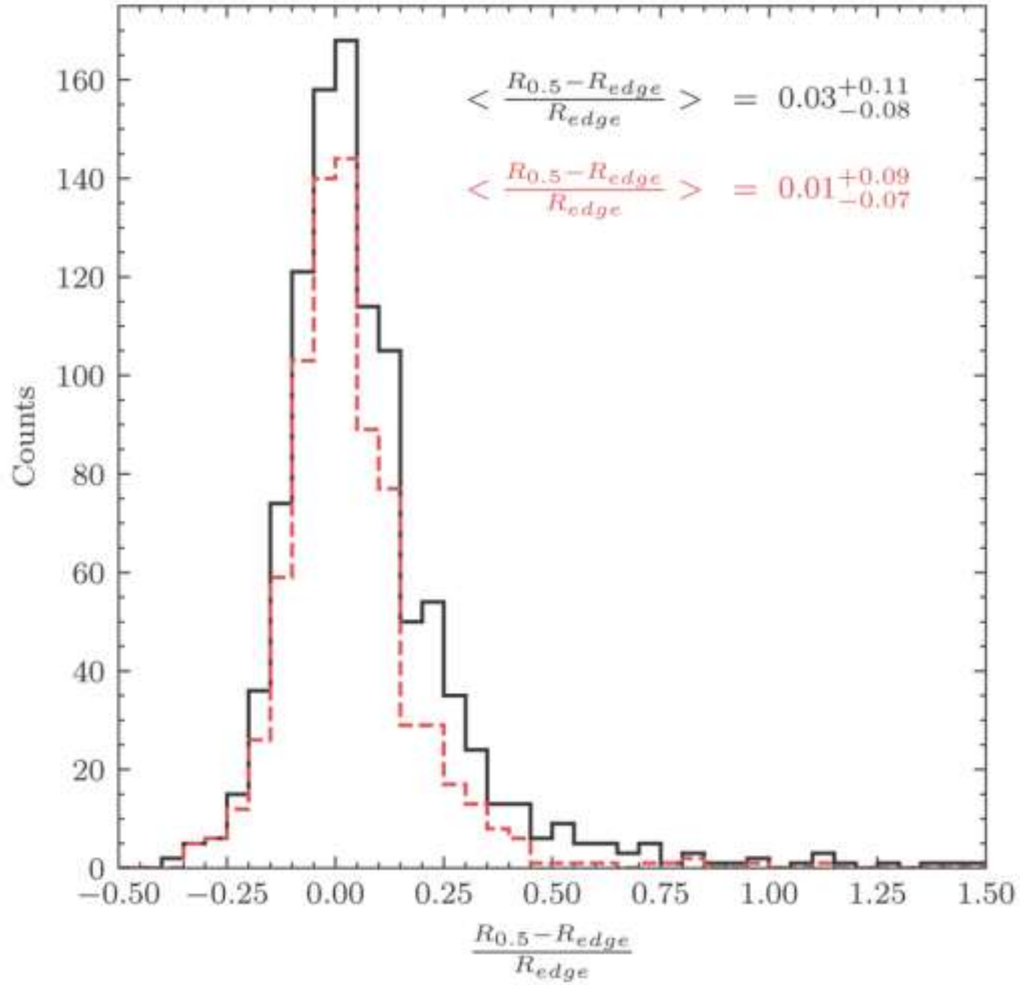
From left to right, relative difference between the size of the galaxy measured from the SAM truncation (denoted as R0.5) and the semi-major axis of the truncation ellipse in BT24 (denoted as Redge) as a function of: morphological type, redshift, apparent size (Redge , semi-major axis of the BT24 truncation) split in quintiles, and axis ratio of the BT24 truncation. Boxes represent the IQR, while the horizontal lines indicate the median value. SAM truncations are on average overestimated with respect to BT24 truncations for the most elongated (in projection) and/or for the (apparently) smallest galaxies.

4 RESULTS



Relative error on the size of the truncation inferred with SAM with respect to the BT24 measurements as a function of the relative error on the size of the truncation derived from the averaged SAM segmentation maps ($\Delta R_{0.5}$) split in quintiles. Boxes represent the IQR and whiskers extend from the box to the farthest data point lying within 1.5 × IQR from the box, while the horizontal lines indicate the median value. Median values (and the scatter) of the relative error on the size obtained with SAM compared to BT24 are below the ~5% level for all quintiles with the exception of Q5. For Q5, where the differences between the size of the truncation at 0.5 and 0.7 probability thresholds are the largest, the median value of the relative error on the size of the truncation inferred with SAM surpasses the ~10% level, with a clear increase in the scatter.

4 RESULTS



Distributions of the relative error between the size of the galaxy measured from the SAM truncation (denoted as R0.5) and the semimajor axis of the truncation ellipse in BT24 (denoted as Redge). The solid black histogram corresponds to the whole sample of 1047 galaxies, while the dashed red histogram shows the distribution for 774 galaxies with ΔR0.5 ≲ 0.17 (i.e., excluding those within the Q5 quintile) and b/a > 0.25. The median and IQR of the relative error between the two estimates for both distributions are also shown.

5 CONCLUSIONS

Our main results are summarized below:

By leveraging SAM, we can achieve rapid and reliable galactic size estimations, which is crucial for large-scale astronomical surveys like the Euclid Wide Survey.

In particular, SAM will allow us to estimate truncations for millions of galaxies up to z ~ 2 when applied to the upcoming Euclid Wide Survey, paving the way for understanding the physical processes that shape galaxy evolution across cosmic time.

Within the context of Euclid, precise size estimates will contribute to mapping the distribution of galaxies across large-scale structures and understanding the interplay between galaxy evolution and dark matter halos, key to Euclid’s cosmological goals.

Thanks