PSFGAN: a generative adversarial network system for separating quasar point sources and host galaxy light

Dominic Stark, Barthelemy Launet, Kevin Schawinski, Ce Zhang, Michael Koss, M Dennis Turp, Lia Sartori, Hantian Zhang, Yiru Chen, Anna Weigel

Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **477,** 2513–2527 (2018)
Advance Access publication 2018 March 22

doi:10.1093/mnras/sty764

PSFGAN: a generative adversarial network system for separating quasar point sources and host galaxy light

Dominic Stark,[1]★ Barthelemy Launet,[1,2]★ Kevin Schawinski,[1]★ Ce Zhang,[3] Michael Koss,[1,4] M. Dennis Turp,[1] Lia F. Sartori,[1] Hantian Zhang,[3] Yiru Chen[5] and Anna K. Weigel[1]

[1]*Institute for Particle Physics and Astrophysics, ETH Zurich, Wolfgang-Pauli-Strasse 27, CH-8093 Zürich, Switzerland*
[2]*LERMA, Observatoire de Paris, PSL Research Univ., CNRS, Sorbonne Univ., UPMC Univ. Paris 06, F-75014 Paris, France*
[3]*Systems Group, ETH Zurich, Universitätstrasse 6, CH-8006 Zürich, Switzerland*
[4]*Eureka Scientific, 2452 Delmer Street Suite 100, Oakland, CA 94602-3017, USA*
[5]*Institute of Network Computing and Information Systems, Peking University, Yiheyuan Lu, Haidian District, Beijing 100871, China*

Accepted 2018 March 20. Received 2018 March 20; in original form 2017 December 12

ABSTRACT

The study of unobscured active galactic nuclei (AGN) and quasars depends on the reliable decomposition of the light from the AGN point source and the extended host galaxy light. The problem is typically approached using parametric fitting routines using separate models for the host galaxy and the point spread function (PSF). We present a new approach using a Generative Adversarial Network (GAN) trained on galaxy images. We test the method using Sloan Digital Sky Survey *r*-band images with artificial AGN point sources added that are then removed using the GAN and with parametric methods using GALFIT. When the AGN point source is more than twice as bright as the host galaxy, we find that our method, PSFGAN, can recover point source and host galaxy magnitudes with smaller systematic error and a lower average scatter (49 per cent). PSFGAN is more tolerant to poor knowledge of the PSF than parametric methods. Our tests show that PSFGAN is robust against a broadening in the PSF width of ±50 per cent if it is trained on multiple PSFs. We demonstrate that while a matched training set does improve performance, we can still subtract point sources using a PSFGAN trained on non-astronomical images. While initial training is computationally expensive, evaluating PSFGAN on data is more than 40 times faster than GALFIT fitting two components. Finally, PSFGAN is more robust and easy to use than parametric methods as it requires no input parameters.

Key words: methods: data analysis – techniques: image processing – quasars: general.

1 INTRODUCTION

Active galactic nuclei (AGN) are among the brightest continuously emitting objects in the Universe radiating in most wavelengths of light. The link between AGN and the host galaxy properties such as stellar mass (e.g. Hernán-Caballero et al. 2013; Vitale et al. 2013; Matsuoka et al. 2014; Reines & Volonteri 2015) and star formation rate (e.g. Kim, Ho & Im 2006; Schawinski et al. 2006; Santini et al. 2012; Shimizu et al. 2015) is critical to better understand the relationship between black hole growth and the host galaxy. These quantities are frequently inferred by modelling the spectral energy distribution (SED) from multiwavelength data (Simmons et al. 2011; Michałowski et al. 2014; Chang et al. 2015; Collinson et al. 2015). Unfortunately, especially in unobscured quasars, the light from the AGN far outshines the host galaxy emission. Investigating correlations between galaxy parameters and properties of the AGN thus requires a separate analysis of AGN and host galaxy components (Gabor et al. 2009; Pierce et al. 2010).

Extending photometric studies to host galaxies at higher redshift (e.g. Simmons & Urry 2008; Böhm et al. 2013) is critical to understanding their evolution across cosmic time. However for imaging data, if the host galaxy is very faint compared to the quasar and its angular size is close to the width of the point spread function (PSF), it can be hard to detect the host galaxy at all (e.g. Bahcall et al. 1997). Following the pioneering work of Bahcall, Kirhakos & Schneider (1995) the first studies of quasar hosts were conducted using the *Hubble Space Telescope* (*HST*; e.g. McLeod & Rieke 1995; Hooper, Impey & Foltz 1997; Kirhakos et al. 1999; Lehnert et al. 1999). The most widely used techniques were based on scaling and aligning a stellar PSF to the peak of the surface brightness distribution of the quasar. Other approaches included some constraints on the residual host galaxy emission such as monotonicity of the

★ E-mail: dostark94@gmail.com (DS); launet.barthelemy@gmail.com (BL); kevin.schawinski@phys.ethz.ch (KS)

radial light profile (Boyce, Disney & Bleaken 1999). These methods however systematically overestimate the quasar contribution and only yield a lower limit for the host galaxy flux. Later studies showed that fitting two-dimensional galaxy components simultaneously with the point source (PS) component yields the most robust method (Peng et al. 2002; Bennert et al. 2008).

One of the most popular methods used for two-dimensional surface profile fitting is GALFIT (Peng et al. 2002, 2010). Its ability to recover PS fluxes and host galaxy parameters has been demonstrated several times both for *HST* images (Kim et al. 2008; Simmons & Urry 2008; Gabor et al. 2009; Pierce et al. 2010) and for ground-based images (Goulding et al. 2010; Koss et al. 2011). GALFIT is a very powerful tool for detailed morphological decomposition of single cases but it was not designed for batch fitting (Peng et al. 2002). In the era of 'big data' in astronomy,[1] where large data sets have to be efficiently analysed without human interaction, parametric fitting might not be an efficient approach. Nevertheless there have been approaches (Vikram et al. 2010; Barden et al. 2012) to automate GALFIT by combining it with SEXTRACTOR (Bertin & Arnouts 1996), but these methods still depend on their input parameters.

Machine learning (ML) often accomplishes the demand for automation and scalability in data analysis. Various ML techniques have been applied to astronomy, for example in outlier detection (Baron & Poznanski 2017), galaxy classification (Dieleman, Willett & Dambre 2015; Sreejith et al. 2018), or detector characterization (George & Huerta 2018). The most recent developments in automated galaxy fitting use Bayesian inference (Yoon, Weinberg & Katz 2011; Robotham et al. 2017) or deep learning (Tuccillo et al. 2018).

By using a Generative Adversarial Network (GAN; Goodfellow et al. 2014) we develop the first ML-based method for separating AGN from their host galaxies. We adopt the GALAXYGAN algorithm (Schawinski et al. 2017) that was originally conceived to recover features in noisy ground-based imaging data. Our method is called PSFGAN as it subtracts point sources from CCD images. We test the effectiveness of PSFGAN at recovering the AGN (and the host galaxy) and compare our results to GALFIT. In Section 2 we describe the overall method, we describe the specific GAN architecture in Section 2.1, the training and testing procedure in Section 2.2, the model selection in Section 2.3, and in Section 2.4 the GALFIT fitting strategy we used for the comparisons. In Section 3 we test the performance of PSFGAN. Finally, in Section 4 we discuss applications and limitations.

Throughout this paper, we adopt a cosmology with $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$, and $H_0 = 70\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$.

2 METHOD

2.1 GAN architecture

In Fig. 1 , we show a graphical scheme of the architecture we used. A GAN consists of two neural networks: a generator and a discriminator. The generator creates artificial data sets, and the discriminator classifies a given set as 'real' or 'fake'. The generator and the discriminator are simultaneously trained. In an ideal case, the generator recovers the training data distribution (Goodfellow et al. 2014). Conditional GANs take a conditional input (Reed et al. 2016) and can be used for image processing (Isola et al. 2017). GALAXYGAN takes a degraded galaxy image as conditional input (Schawinski et al. 2017). During the training the generator tries to recover the original image from the degraded one. The discriminator learns to distinguish between the original image and the generator output. Both networks are trained at the same time to maximize the others loss and by this means the generator learns the inverse of the transformation that has been applied to the original image. In the testing phase, the generator is applied to degraded images it has never seen before, in order to recover the original ones. In this work we choose the processed image to be the original galaxy image with a simulated PS representing an unobscured AGN. Using this as the conditional input, the generator then learns the inverse transformation that is equivalent to subtracting the PS.

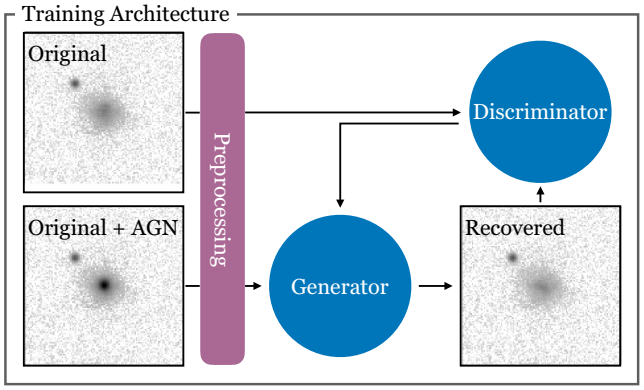

Figure 1. Scheme of the architecture used in this work. The generator takes as input the modified image (original galaxy image with a simulated PS in its centre) and tries to recover the original galaxy image. The discriminator distinguishes the original from the recovered image. Before feeding the images to the GAN they are normalized to have values in [0,1] and transformed by an invertible stretch function.

Adding a simulated PS to the centre of a galaxy image will primarily affect a few pixels at the centre of the image. We therefore adapt the generator to increase the weight of the central region in the loss computation.

2.2 Data preparation

We use *r*-band images from the SDSS (Blanton et al. 2017) as a test case though PSFGAN can be applied to any CCD imaging data in any filter. For this proof of concept we choose SDSS data because it is very homogeneous and has many galaxy images available for large training sets. We divide the data into a training set, a validation set for model selection, and a testing set to evaluate model performance. Each set consists of image pairs (original and conditional input). However, only during training PSFGAN uses both the original image without a PS and the conditional input (original image with added PS). In the validation set and the testing set we use exclusively the conditional input as we only run the trained generator on these samples. To avoid overfitting and ensure the generalization ability of our approach, throughout the whole project, we only use the testing set *once* for each of the final experiments. The development of models is conducted completely using the validation set.

We test PSFGAN on three redshift ranges corresponding to $z \sim 0.05$, $\sim$0.1, and $\sim$0.2, respectively. In these ranges we use 424×424 pixels ($168 \times 168\,\mathrm{arcsec}^2$) cut-outs of SDSS galaxies with some variation of redshift to $z \in [0.045, 0.055]$, $z \in [0.095,$

[1] Currently, the total data volume of Sloan Digital Sky Survey (SDSS) is >125 TB (Blanton et al. 2017). The Large Synoptic Survey Telescope (LSST) will produce 15 TB of data per year (https://www.lsst.org/about/fact-sheets).

0.105], and $z \in [0.194, 0.206]$, respectively. For each redshift sample we split the data into training set of 5000 images, a validation set of 200 images, and a testing set of 200 images.

In the following we describe the transformation that we apply to the original images to get the conditional input. We perform the following three steps:

(i) we extract the PSF from the SDSS data;
(ii) we scale the PSF to a value by a contrast ratio drawn from a predefined distribution;
(iii) we align the centroid pixel of the galaxy with the centroid pixel of the PSF and then add the images pixel wise.

Hence for a given original image, the corresponding conditional input was defined by two parameters: (a) the brightness of the PS and (b) the shape of the PSF we convolved it with.

In detail, we implement this procedure differently for the training sets than for the validation and test sets.

In the training sets, we use the PSF tool provided by SDSS (Stoughton et al. 2002) to extract in each image the PSF and fit it with three 2D Gaussians in order to get an analytical PSF.[2] Its brightness is then scaled by a contrast ratio R defined with respect to the host galaxy luminosity (CMODELFLUX_R). This contrast ratio is drawn from a uniform distribution in linear space between 0.1 and 10, i.e. $R \sim U([0.1, 10])$. As we describe in Section 2.3, this distribution was chosen because it yields the best performance (among the tested distributions).

In the validation and testing sets, following the approach of Koss et al. (2011), we measure a semi-empirical PSF by median-stacking 40–60 stars from the neighbourhood of the galaxy. (The mismatch between training and testing PSF is necessary to take into account the lack of information about the exact PSF we would have in a real situation.) Because of the high dynamic range of contrasts we want to test for we draw R from a uniform distribution in logarithmic space, i.e. $\log(R) \sim U([-1, 1])$.

Table 1 shows an overview of the parameters chosen for different data sets.

Table 1. Overview of the data sets that were used for each of the three redshift groups. R is the contrast ratio that ranged between 0.1 and 10 in all cases. To simulate a real application case we introduced a discrepancy between the PSFs in the training and the testing set. The table shows how we simulated the AGN point sources in each distinct case. For a distribution of contrast ratios in the training set we refer to Section 2.3.2. The size of the data sets was determined heuristically and can be taken as a guideline of what might be appropriate for a general application.

	Training set	Testing and validation sets
PSF	Analytical fit of SDSS tool PSF	40–60 stars combined by median stacking
PDF of R	Uniform in linear space	Uniform in logarithmic space
Number of image pairs	5000	200

2.3 GAN models

To find a good model we train with different hyperparameters and then evaluate each trained model on the three validation sets. We then choose the model with the overall best performance. We emphasize that we do not perform an exhaustive hyperparameter search. Also the accuracy of a model depends on the random initialization of the weights at the beginning of the training. Therefore, the performance of the selected model represents a lower bound. An exhaustive search for the best hyperparameter and initialization would likely result in superior performance over our limited search.

To quantify the performance we define the recovery ratio as the ratio of recovered PS flux to the real PS flux[3] and compute its mean absolute deviation (MAD) from 1 in each validation set. As we will reuse this quantity for various tests in Section 3 we simply call it Δ:

$$\Delta := \text{average}\left(\left|\frac{\text{recovered PS flux}}{\text{real PS flux}} - 1\right|\right).$$

The average is taken over the 200 instances in the actual validation set. This yields a score for each redshift sample. We average these three scores again in order to obtain a measure for the general accuracy of a model. We then choose the model with the minimal average score.

While searching for the best GAN model we vary the following parameters:

(i) pre-processing: normalizing and redistributing pixel values by applying a non-linear stretching function;
(ii) distribution of contrast ratios in the training set;
(iii) learning rate (defined in Section 2.3.3).

Testing the whole parameter space is computationally expensive. Therefore we vary only one parameter at a time while holding the other two parameters at a fixed value.

We discuss the different models and their scores Δ. Users applying PSFGAN are advised to use our results as a starting point for optimizing the parameters for their specific data.

2.3.1 *Stretch function and scale factor*

It is a common practice to normalize and redistribute the input values of neural networks such that they are comparable across the training set (Sola & Sevilla 1997). This pre-processing is especially important for this work due to the high contrast between galaxy and PS brightness. If the data were just normalized and scaled linearly, the galaxy would have been interpreted as noise by the GAN in the cases where the PS is very bright.

Not only the input images themselves have a high dynamic range but also the maximum pixel values across the training set. We want to find a reversible transformation to rescale the images, i.e. redistribute the pixel values in a smaller range. The pixels in the transformed image should be distributed in a way that the GAN is sensible to both the PS and the host galaxy in all of the images. The transformation has to be unique so that it can be applied to all images before showing them to the GAN, and applied back on the output images to recover the full pixel scale. We test several stretching functions (see Table 2) while holding the learning rate constant at $\text{lr} = 9 \times 10^{-5}$ and using a uniform distribution in linear space for the contrast ratios in the training set.

[2] This tool generates a position-dependent, semi-empirical PSF by use of a Karhunen–Loève transform (Stoughton et al. 2002). The fitting step that is performed on the output of the tool is necessary to remove the noise in this PSF image. (The noise would be amplified when the PSF is scaled to high contrast ratios that would lead to unrealistic images.)
[3] The *real* PS flux is the flux of the PS that we put on to the original image.

Table 2. Overview of the stretch functions used. A is the scaling factor, max refers to the brightest pixel across the whole training set. max = 6140 for $z \sim 0.05$, max = 1450 for $z \sim 0.1$, and max = 1657 for $z \sim 0.2$.

asinh	$\frac{\operatorname{asinh}(A\,x)}{\operatorname{asinh}(A\,\mathrm{max})}$
log	$\frac{\log\left(\frac{A\,x}{\mathrm{max}}\right)}{\log A}$
pow	$\sqrt[A]{\frac{x}{\mathrm{max}}}$
sigmoid	$2\left(\frac{1}{1+\mathrm{e}^{\frac{A\,x}{\mathrm{max}}}} - \frac{1}{2}\right)$

We observe (Fig. 2) that the asinh stretch function with a scale factor $A = 50$ model has the smallest average Δ.

2.3.2 *Distribution of contrast ratios in the training set*

We test two different distributions of contrast ratios in the training set: a uniform distribution in linear space $R \sim U([0.1, 10])$ and a uniform distribution in logarithmic space $\log(R) \sim \boldsymbol{U}([-1, 1])$. We hold the stretch function constant at asinh, $A = 50$, and the learning rate at $\mathrm{lr} = 9 \times 10^{-5}$. In Fig. 2 we plot the scores resulting from evaluation on the validation sets. If PSFGAN is trained on a sample with contrast ratios distributed uniformly in linear space, it is more stable than if it is trained on a sample with contrast ratios distributed uniformly in logarithmic space.

2.3.3 *Learning rate*

The discriminator and the generator are neural networks. Therefore they minimize their loss functions by adapting the weights of their neurons. The learning rate determines how much the weights are adjusted in each training step. For a more technical description of the optimization algorithm we are using, we refer to Kingma & Ba (2015).

In Fig. 2 we plot the score Δ for six different learning rates. While varying the learning rates we hold the stretch function constant at asinh with a scale factor of $A = 50$ and the distribution of contrast ratios in the training set is a uniform distribution in linear space. The model with the lowest average Δ is the one with $\mathrm{lr} = 9 \times 10^{-5}$.

2.3.4 *Summary*

Within the subset of the parameter space that we test, we find that the best model is given by the following parameters:

(i) learning rate: $\mathrm{lr} = 9 \times 10^{-5}$;
(ii) distribution of contrast ratios in the training set: uniform in linear space;
(iii) pre-processing: asinh stretch function with a scale factor of $A = 50$.

2.4 GALFIT fitting strategy

In this section we explain the GALFIT fitting strategy we use for the comparisons. GALFIT simultaneously fits an arbitrary number of surface brightness profiles to an image (Peng et al. 2002). Besides various types of inbuilt, analytical function types, it can also fit a PSF provided by the user. A surface brightness component of a specific function type is defined by its geometrical shape and its radial surface brightness profile. For the shape we choose ellipsoids and for the radial surface brightness profile we choose the Sérsic profile as it is usually done in the literature (Simmons & Urry 2008; Koss et al. 2011; Schawinski et al. 2011). The Sérsic profile is defined as

$$\Sigma_r = \Sigma_\mathrm{e} \exp\left(-\kappa\left(\frac{r}{r_\mathrm{e}}\right)^{\frac{1}{n}} - 1\right),$$

where Σ_r is the surface brightness at radius r, r_e is the half-light radius, the Sérsic index n is a positive real number, κ is a parameter

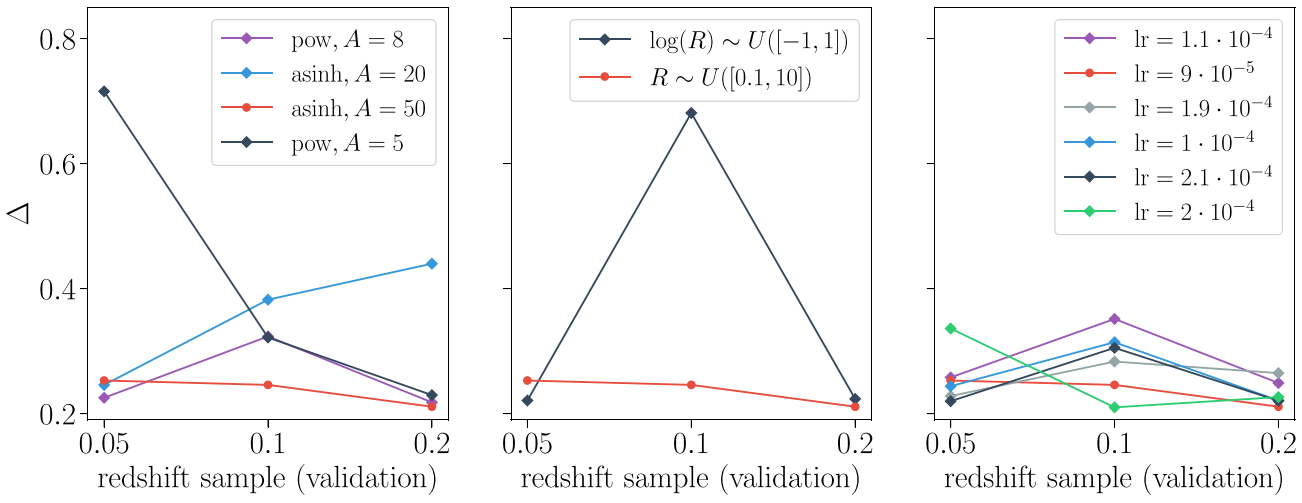

Figure 2. The GAN trained with different hyperparameters evaluated on the validation sets of the three redshift samples. We only search a subspace of the whole parameters space. While varying the stretch function we use $\mathrm{lr} = 9 \times 10^{-5}$ for the learning rate and a uniform distribution in linear space for the contrast ratios in the training set. Results for log and sigmoid functions are not plotted as their score was more than five times the average score plotted here. While varying the distribution of contrast ratios we use asinh, $A = 50$ as stretch function, and $\mathrm{lr} = 9 \times 10^{-5}$ for the learning rate. While varying the learning rates we use asinh, $A = 50$ as stretch function, and a uniform distribution in linear space for the contrast ratios. The quantity on the y-axis is the mean absolute deviation (MAD) of the recovery ratio (recovered PS flux/real PS flux) from 1.

that depends[4] on n, and r_e and the radius r for an ellipsoid is defined by

$$r = \left(|x|^2 + \left| \frac{y}{q} \right|^2 \right)^{\frac{1}{2}},$$

where q is the ratio of the minor to major axis of the ellipses describing the isophotes (Peng et al. 2002). To fit the PS component we provide a PSF image as input for GALFIT. We obtain this PSF in the same way as the PSF we use in the training set of PSFGAN: We run SDSS's PSF tool (Stoughton et al. 2002) and fit the output with three 2D Gaussians.

To let GALFIT runs in an automated way, we use an approach similar to that of Barden et al. (2012). We run the following algorithm on each galaxy of the testing set.

(i) Run GALFIT with only a PS component to very roughly subtract the PS. This yields an initial guess for the PS flux on the one hand and allows for the next step on the other hand.[5]

(ii) Run SEXTRACTOR to get initial guesses for host galaxy flux, geometrical parameters, and half-light radius.

(iii) Find stars above the 5σ limit using the algorithm DAOSTARFINDER (Stetson 1987) and mask them out.

(iv) Run GALFIT with a Sérsic component and a PS component. Let the Sérsic index n be a free parameter within 0 and 4. Constrain the magnitude of the host galaxy to be within ± 1 from the initial guess. Moreover restrict the fitting region to a box of 60 kpc around the galaxy. Leave all the other parameters free.

3 RESULTS

We choose the GAN model that works best on the validation set, and evaluate it on the testing sets to produce the results that we present in the following. Section 3.1 contains the comparison to GALFIT. In Section 3.2 we test the dependence of PSFGAN on the brightness distribution underlying to the PS. In Section 3.3 we test the sensitivity of PSFGAN on the correct modelling of the PSF and in Section 3.4 we investigate the ability of PSFGAN to recover host galaxy structure. We further test the dependence on the size of the training set in Section 3.5 and the performance on lower quality data in Section 3.6. Finally, in Section 3.7 we explore the behaviour of our pre-trained models on higher redshift *Hubble* near-infrared (IR) data.

3.1 Comparison of GAN and GALFIT

We quantify the performance by comparing the recovery error in magnitude of both the PS and the host galaxy: we compute the flux of the recovered PS (host galaxy), divide it by the flux of the input PS (host galaxy), and then convert this ratio to magnitudes. That yields the difference between input PS (host galaxy) magnitude and output PS (host galaxy) magnitude.

To measure the flux of the recovered PS we subtract the output image (the residual after subtracting the PS component) from the input image and then sum up the pixel values inside a box of 40×40 pixels centred on the centre of the galaxy. To measure the flux of the recovered host galaxy we subtract the original image from the output image, sum up the pixel values inside a box of 40×40 pixels centred on the centre of the galaxy, and then add the resulting value to the input host galaxy flux that has already been measured by the SDSS pipeline (Stoughton et al. 2002). We sum up the pixels using a restricted box because PSFGAN also modifies other sources in the image and we do not want to count those modifications as contributions to the PS flux. As the input host galaxy flux we take the quantity CMODELFLUX_R measured by the SDSS pipeline (Stoughton et al. 2002). We plot the median magnitude error in different bins of contrast ratios and the 68 and 90 percentiles. We define the nth percentile as the distance from the median that (in both directions) encloses n per cent of the data points.

Fig. 3 shows the comparison of PSFGAN to GALFIT at the three redshift ranges. Figs 4–7 show example images of the original galaxy, the original galaxy with the simulated PS on top of it, the output images (by PSFGAN and GALFIT), and residuals (the output subtracted from the original galaxy image).

Figs 4–6 show examples of randomly selected contrast ratios in each of the redshift samples. Fig. 7 shows one high contrast example in each redshift sample.

Our results show that for contrast ratios $R < 1.8$ the median PS magnitude error of GALFIT in general is smaller than that of PSFGAN and reverse for contrast ratios higher than that. For contrast ratios below $R = 1.8$ the 68 percentiles of PSFGAN's PS magnitude errors are 1.6–4.7 times those of GALFIT. For contrast ratios $R > 1.8$ the 68 percentiles of PSFGAN's PS magnitude errors are 0.2–1.4 times those of GALFIT. This result is consistent with all redshift samples. For the host galaxy magnitudes we again observe that PSFGAN has smaller systematic error and smaller scatter above $R = 1.8$. For $R < 1.8$ the 68 percentiles of PSFGAN are 1.0–4.6 times those of PSFGAN. For $R > 1.8$ PSFGAN has percentiles smaller than GALFIT with factors between 0.3 and 1.2.

In Table 3 we compare runtime and robustness of PSFGAN and GALFIT. We find that the fitting time of GALFIT is ∼3.6 times the evaluation time of PSFGAN if they are run on the same machine. By running PSFGAN on GPUs it can be further accelerated such that (in our specific case) it is ∼48.3 times faster than GALFIT. We also find that GALFIT crashes in ∼2 per cent of the cases if it is wrapped by our script.

3.2 Dependence on the underlying brightness profile

In order to test whether PSFGAN actually uses information of host galaxy brightness distribution we create a comparison sample consisting of pictures of cats and dogs. We add simulated AGN to the centres of the images at different contrasts: we normalize the animal image in such a way that the sum of the pixel values inside a box of 10×10 pixels around the centre is equal to the sum of the pixel values inside a box of the same length in the original galaxy image. Although contrast ratio is not well defined in the case of animals, we plot the PS magnitude recovery against the contrast ratio the PS would have if it was added to the galaxy it corresponds to. We train PSFGAN once on animals and once on galaxies and then evaluate both on each testing set (again one consisting of animals and one consisting of galaxies).

Fig. 8 contains example images and Fig. 9 shows the cross-comparisons. We conclude that the underlying brightness distribution of the objects does indeed matter: PSFGAN trained on animals is better at subtracting point sources from animals and PSFGAN trained on galaxies is better at subtracting point sources from galaxies. However as the contrast increases this effect gets less significant. For evaluation on galaxies both versions of PSFGAN have the same

[4] The parameter κ ensures that half of the total flux is always within r_e (Peng et al. 2002).

[5] If we let SEXTRACTOR run before subtracting the PS, all the host galaxy parameters would be totally biased by the bright PS.

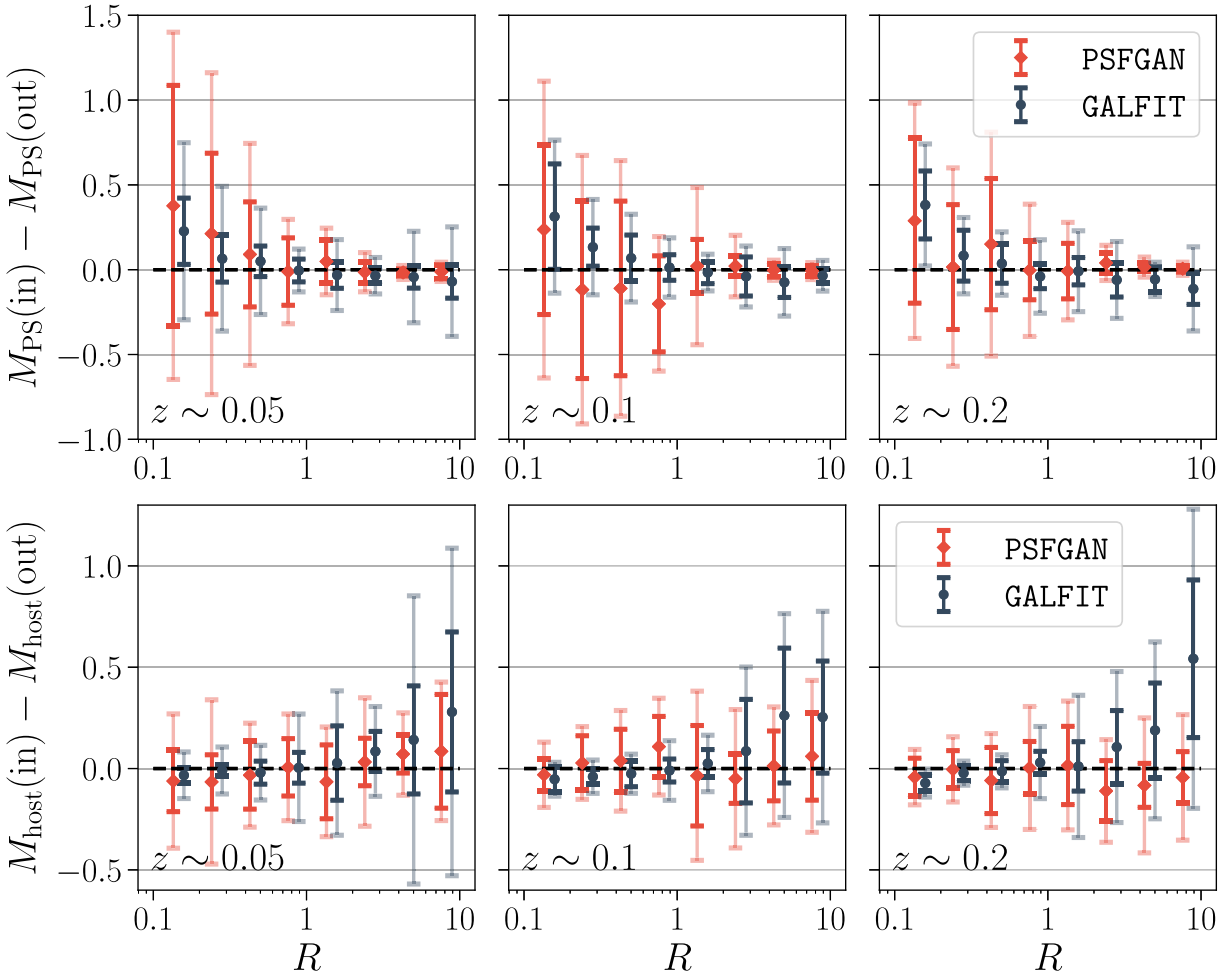

Figure 3. We compare how well PSFGAN and GALFIT can recover PS and host galaxy magnitudes by computing the magnitude difference between the input and the output (recovered) for both the PS and the host galaxy. The plotted quantity is the median in its respective bin and the solid (transparent) error bars indicate the distance from the median where at least 68 per cent (90 per cent) of the data points are enclosed. The dotted line indicates perfect recovery ($M_{\rm PS}({\rm in}) = M_{\rm PS}({\rm out})$ or $M_{\rm host}({\rm in}) = M_{\rm host}({\rm out})$). At $z \sim 0.05$ we exclude five galaxies from the plot because GALFIT crashed on them. For the same reason we exclude two galaxies from the $z \sim 0.1$ plot and three galaxies from the $z \sim 0.2$ plot.

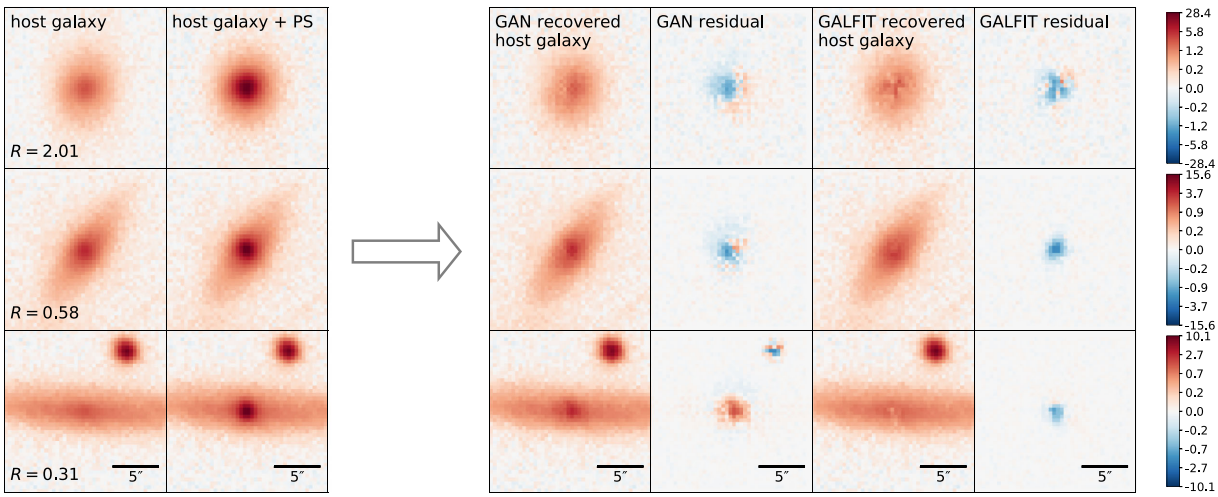

Figure 4. Examples at $z \sim 0.05$ with different contrast ratios. In each row we plot the original host galaxy and the host galaxy with the simulated PS in its centre. We then plot the output images of PSFGAN and GALFIT to see how they differ from the original galaxy image. We call the output images recovered host galaxy image. Moreover we show residuals (recovered original) for both methods.

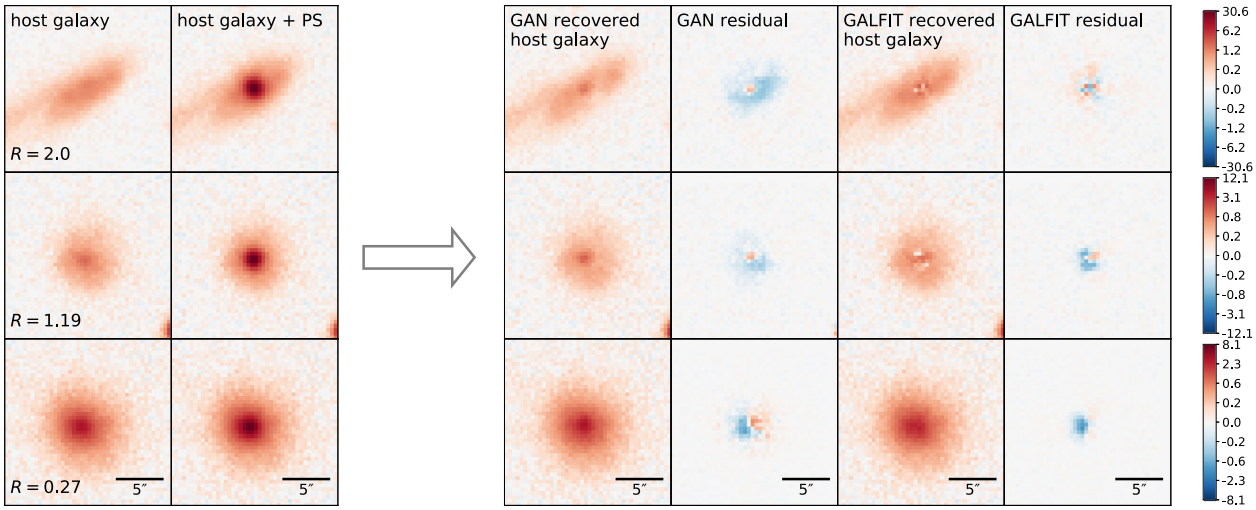

Figure 5. Examples at $z \sim 0.1$ with different contrast ratios. The format of the plots is the same as in Fig. 4.

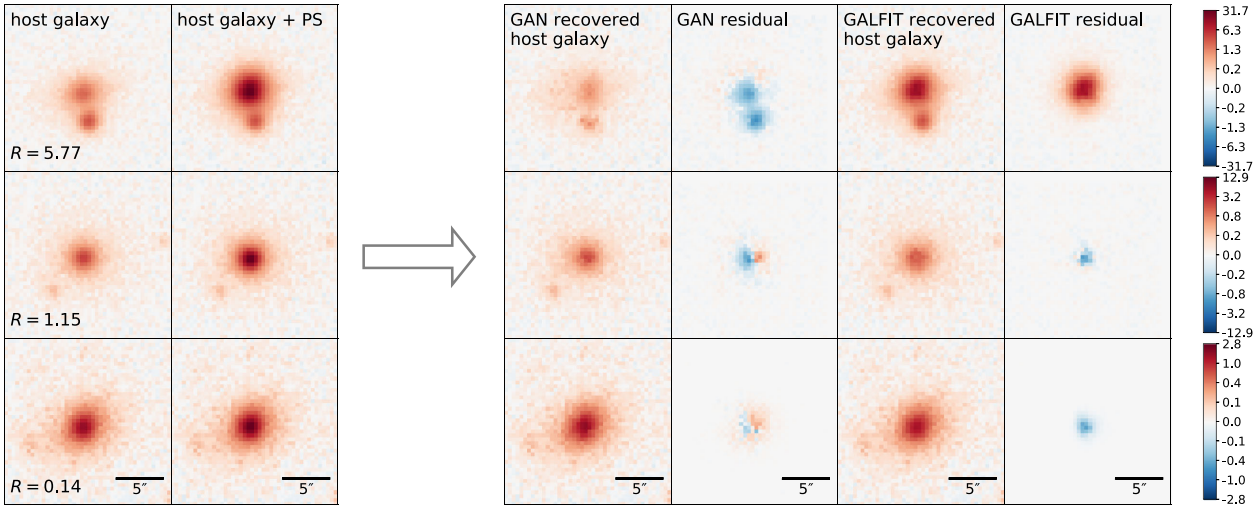

Figure 6. Examples at $z \sim 0.2$ with different contrast ratios. The format of the plots is the same as in Fig. 4.

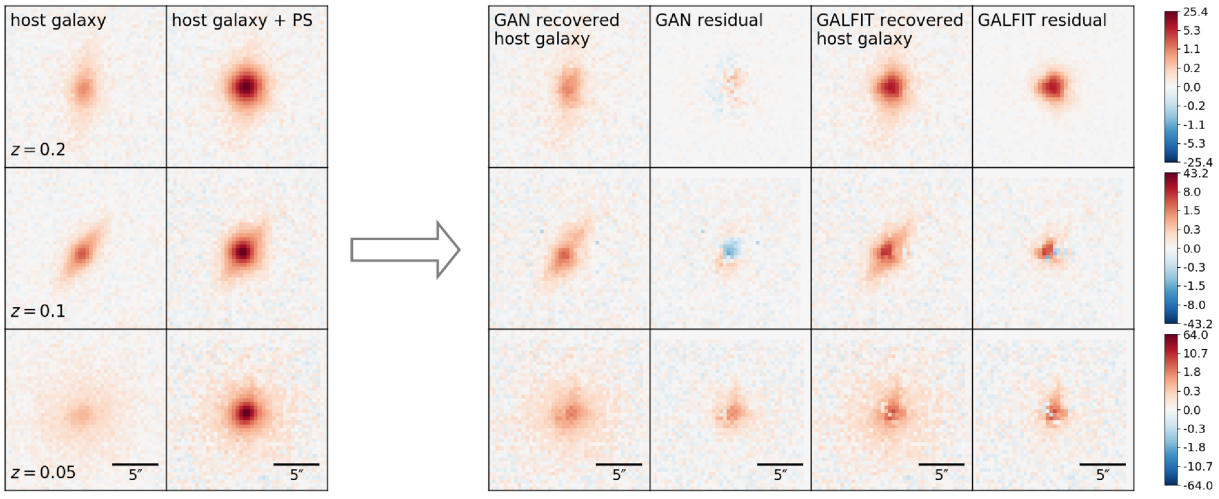

Figure 7. High contrast examples in all three redshift samples. The format of the plots is the same as in Fig. 4.

Table 3. Comparison of PSFGAN and GALFIT in terms of runtime and robustness. The specified runtime are measured on the $z = 0.1$ test sample with images of size 424 × 424 pixels. GALFIT however only fits a cut-out of 83 × 83 pixels (∼60 kpc). We note that for redshift $z = 0.2$ ($z = 0.05$) the GALFIT fitting time is smaller (larger) as we fix the fitting region to a box of fixed physical length.

	PSFGAN	GALFIT
Training time	Takes ∼8 h for 5000 images on one NVIDIA Titan Xp GPU	N/A
Inference/fitting time	Takes ∼2 s per image on a Macbook Air with a 1.7 GHz Intel Core i5 CPU and 4 GB RAM (total runtime of 8.3 h for 500 images, ∼560 h for 10^6 images)	Takes ∼7.25 s per image on the same Macbook (1 h for 500 images, ∼2000 h for 10^6 images)
	Takes ∼0.15 s per image on one NVIDIA Titan Xp GPU (total runtime of 8 h for 500 images, 49.7 h for 10^6 images)	Not compatible with GPU technology
Crashes	Always outputs an image (by construction)	2.5 per cent crashes for $z \sim 0.05$, 1 per cent crashes for $z \sim 0.1$, and 1.5 per cent crashes for $z \sim 0.2$

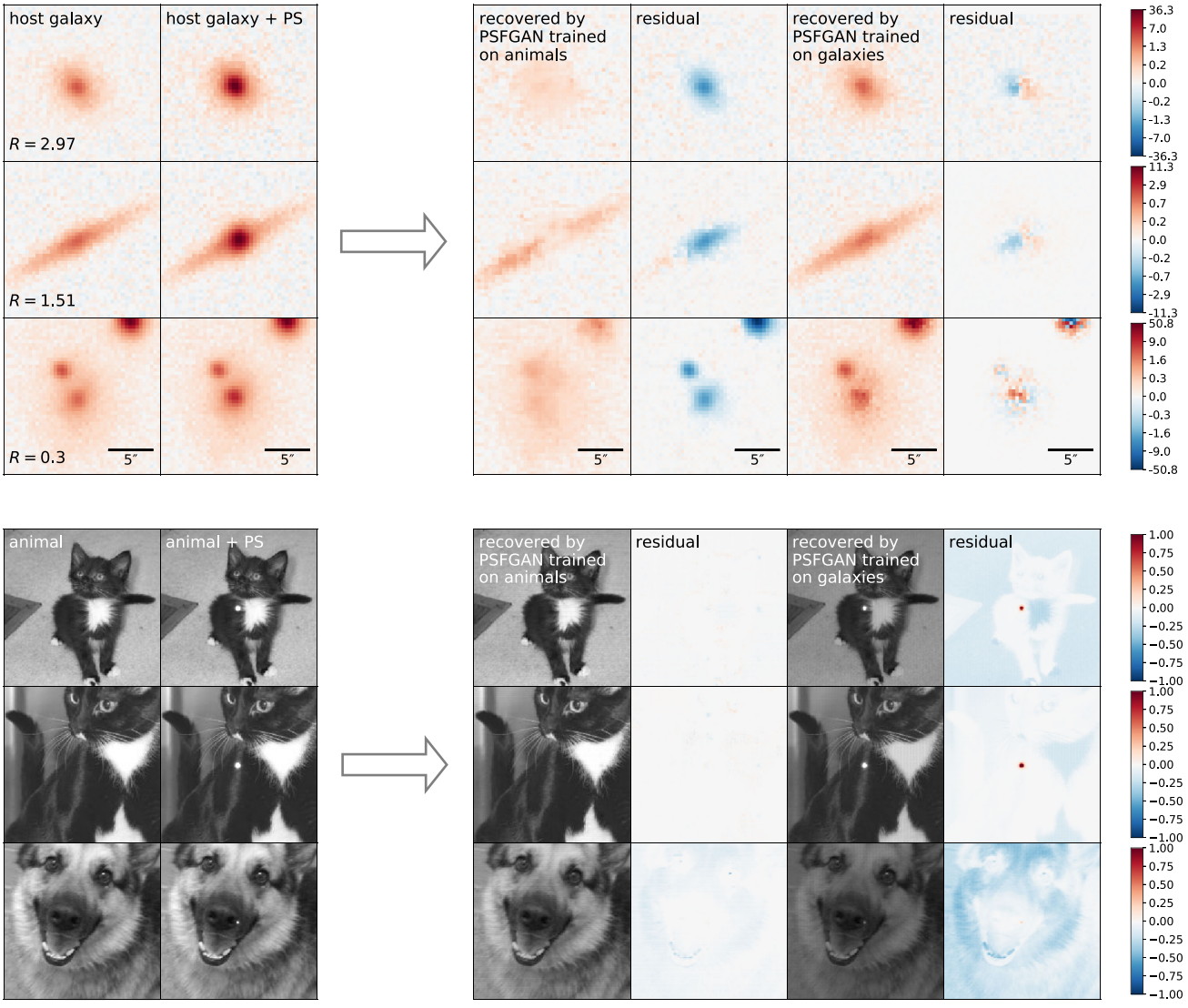

Figure 8. We validate the hypothesis that the visual structure of galaxies helps PSFGAN for PS subtraction (i.e. the neural network learns, intuitively, what galaxies look like and uses this information for PS subtraction). To validate this, we apply PSFGAN to very different domains: cats and dogs. Top: with galaxy images as test images, we compare the outputs and residuals of PSFGAN trained on animals to the ones of PSFGAN trained on galaxies. Bottom: with cats and dogs images as test images, we compare the outputs and residuals of PSFGAN trained on animals to the ones of PSFGAN trained on galaxies. We note that the colour map describes only the residuals. We scaled them differently from the animal images in order to visualize both over- and undersubtraction.

Table 4. In this table we show how an overall single Gaussian model would be affected by the broadening of the central Gaussian component of our PSF. In the left-hand column we list the relative broadening of the central component of the triple Gaussian PSF. To obtain intuition, we fit the PSF with a single Gaussian before and after broadening it. We then compute the change in full width at half-maximum (FWHM; of the single Gaussian) after broadening relative to the FWHM (of the single Gaussian) before broadening and take the average over the whole set. These are the values in the second column.

Relative change in FWHM of central Gaussian (per cent)	Relative change in mean FWHM of single Gaussian fit (per cent)
±0	+0
+5	+2
−5	−2
+10	+5
−10	−4
+15	+7
−15	−6
+20	+10
−20	−8
+30	+15
−30	−11
+50	+27
−50	−13
+60	+33
−60	−21
−70	−34
−80	−46
−85	−52
+100	+57
+200	+127
+300	+200
+500	+354

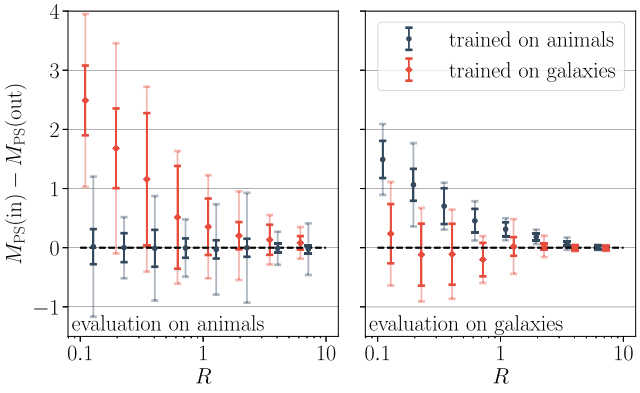

Figure 9. A quantitative summary of the same experiment in Fig. 8. The right-hand image shows evaluations on a sample of 200 galaxies (it is the test sample of $z \sim 0.1$) with artificial point sources, and the left-hand image shows evaluations on a set of 200 animals with artificial point sources.

68 percentiles in the highest contrast bin $R > 5.6$. Also for evaluation on animals the systematic error and the scatter of PSFGAN gradually decrease with increasing contrast ratios. In the highest contrast bin the version of PSFGAN trained on galaxies has smaller 90 percentiles. Its 68 percentile in this bin is twice the 68 percentile of the version trained on animals.

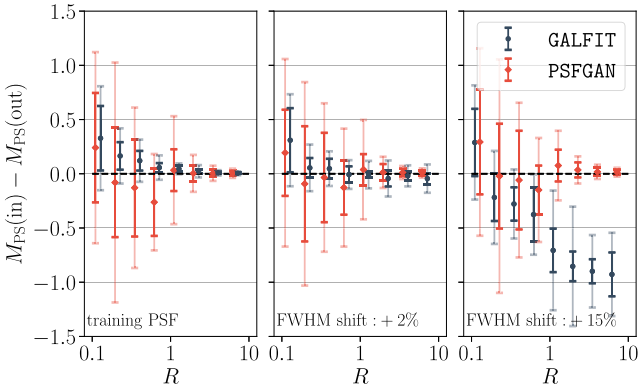

Figure 10. Comparison of PSFGAN and GALFIT with increasing broadening of the PSF. We fit the three-Gaussian PSF with one single Gaussian before and after broadening its central component and then compute the relative change of the FWHM of this single Gaussian fit. We use this relative change as a measure for the PSF broadening (or narrowing).

3.3 PSF dependence

To test the sensitivity of PSFGAN on the PSF shape we simulate atmospheric seeing variation that occurs in ground based imaging data. We change the width of the core of the PSF that we use to simulate the AGN. As a starting point we choose the 3-gaussians fit of the PSF generated by the SDSS tool. This is the same PSF that we use to simulate the AGN in the training set and also the same PSF that GALFIT uses to fit the PS component. The FWHM[6] of its central gaussian component (the one of those three with smallest FWHM) varies across the images in the testing set with a standard deviation of 0.13″ around a mean of 0.98″. Furthermore the central gaussian on average contains 48 % of the total flux of the PSF (with a standard deviation of 9 %). To simulate seeing variation we broaden this FWHM by a certain percentage for each PS in the testing set of $z \sim 0.1$. When we broaden the FWHM we keep the amplitude constant in order to avoid the central gaussian from becoming negligible (compared to the other two components) when it is broadened. In order to get some intuition we compute (in Table 4) what the changes in FWHM of the central gaussian component would correspond to if the PSF was modeled by a single gaussian.

As an example we show a comparison of PSFGAN and GALFIT for broadenings in the single Gaussian full width at half-maximum (FWHM) of +0, +2, and +15 per cent in Fig. 10. Fig. 11 shows the score Δ for the whole range of FWHM broadenings we tested. We also compare to a version of PSFGAN that was trained on a single PSF. We randomly choose one of the PSF's generated by the SDSS tool and constantly use this one as to simulate the AGN in each galaxy.

The results show that GALFIT has very high accuracy if its input PSF is the same that the one used to simulate the AGN. As soon as there is some discrepancy introduced between those two PSFs GALFIT starts to have large systematic errors. PSFGAN starts to have problems only for broadenings >100 per cent (in the FWHM of a single Gaussian model). PSFGAN trained on a single PSF is in general

[6] The gaussians we use to fit the PSF are slightly elliptical with an average axis ratio (for the central component) of 0.91 and a standard deviation of 0.06. We use the same definition of FWHM that is used by (Peng et al. 2002): The radial surface brightness profile of a two-dimensional gaussian is parametrized by the same r that we defined in subsection 2.4. The FWHM is then given by $2\sqrt{2 * \ln 2} \cdot \sigma$, which is analogous to the one dimensional case.

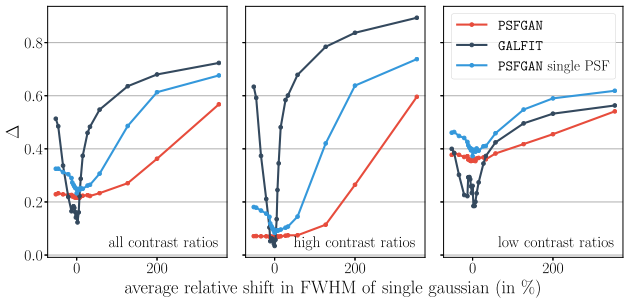

Figure 11. Score of PSFGAN and GALFIT for different PSF broadenings. To get the quantity on the x-axis we fit the three-Gaussian PSF before and after broadening its central component. We then take the relative FWHM change of this single Gaussian fit as a measure for the PSF broadening (or narrowing). The plots show that GALFIT is more sensitive on the PSF than PSFGAN.

(though not for low contrasts) also more robust to PSF variation in the test set. Its Δ is however higher than that of the normal PSFGAN. Judging from the score Δ, GALFIT can handle a seeing variation of approximately $+8$ and -20 per cent. However at high contrast ratios $R > 1$, PSFGAN has already a lower score for -13 and $+5$ per cent. We conclude that PSFGAN is more robust against seeing variation and improper modelling of the PSF. Moreover we can infer that PSFGAN learns the variation of the PSF during training if it is trained on a variety of PSFs.

3.4 Host galaxy structure recovery

By now we have only tested the recovery of magnitudes. To test how well PSFGAN recovers structure of the host galaxy we use the structural similarity index (SSIM). The SSIM is a distance metric for two images that takes into account spatial correlations between different pixels (Wang et al. 2004). The SSIM of two images that are the same is 1 and it decreases as one of the two images is degraded. As the SSIM was designed to coincide with the quality assessment of the human eye (Wang et al. 2004), we consider it useful for quantifying the loss and recovery of structural information of AGN host galaxies. For this test we created an additional test sample consisting of spiral galaxies. We compare the structure recovery on this sample to the structure recovery on the normal test samples that we use in this work. It serves as a comparison sample as it consists of mixed types of galaxies.

To get a sample of spiral galaxies we select galaxies with $z \in [0.04, 0.06]$, $z \in [0.09, 0.11]$, and $z \in [0.19, 0.21]$ that are neither in the training set nor in the validation set and have Galaxy Zoo vote fractions (for either spiral clockwise or spiral anticlockwise) above 70 per cent (Lintott et al. 2008, 2011). The reason for using a slightly wider redshift range here is that there are not enough sources matching our criteria in the redshift range that we use in the other tests. We finally get a total of 129, 168, and 59 sources, respectively.

In Figs 12 and 13 we compute the SSIM between the original image and the recovered images for both GALFIT and PSFGAN. In order to only extract the relevant information we compute the SSIM on

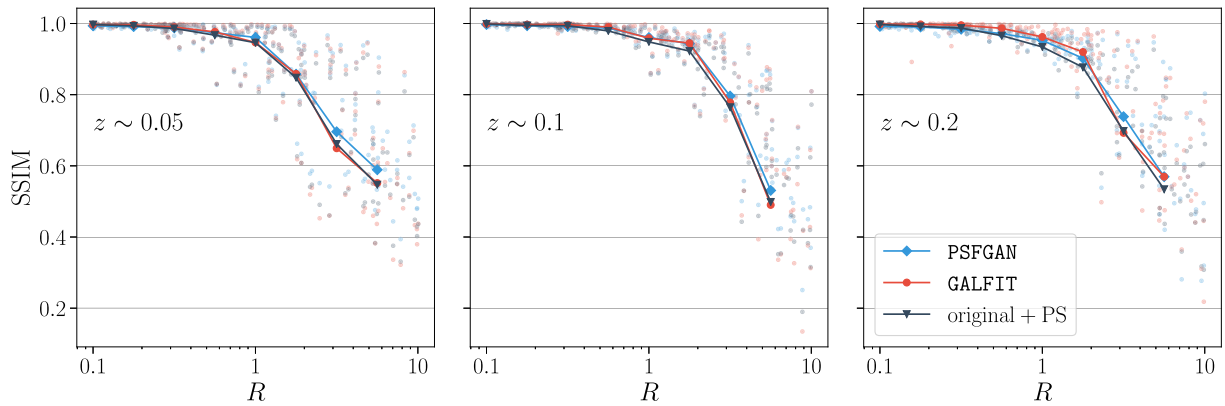

Figure 12. SSIM for the mixed type galaxy sample. We observe that PSFGAN's median SSIM is only higher than that of GALFIT for high contrast ratios.

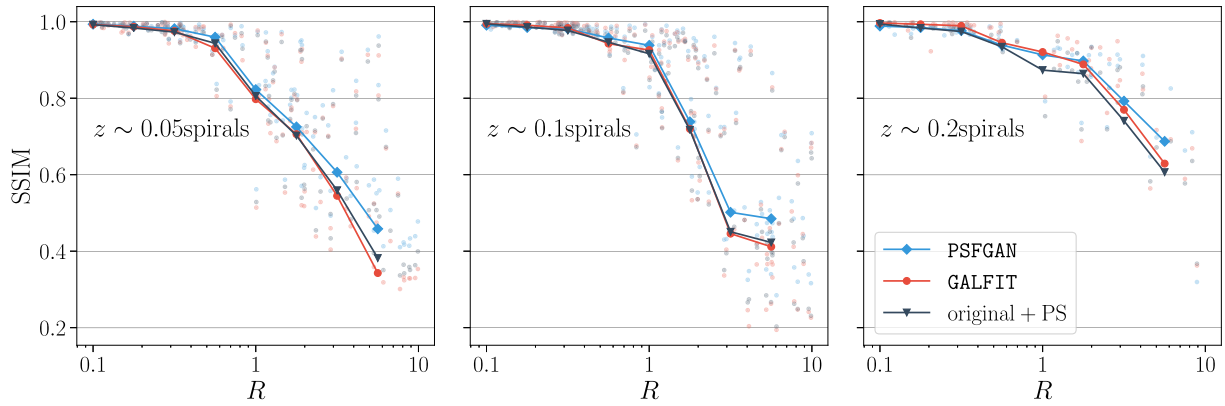

Figure 13. SSIM for spirals galaxies. We observe that PSFGAN's median SSIM is closer to one for moderate and high contrast ratios. For $z \sim 0.05$ and $z \sim 0.1$, PSFGAN has a lower median SSIM only for contrast ratios $R < 0.6$.

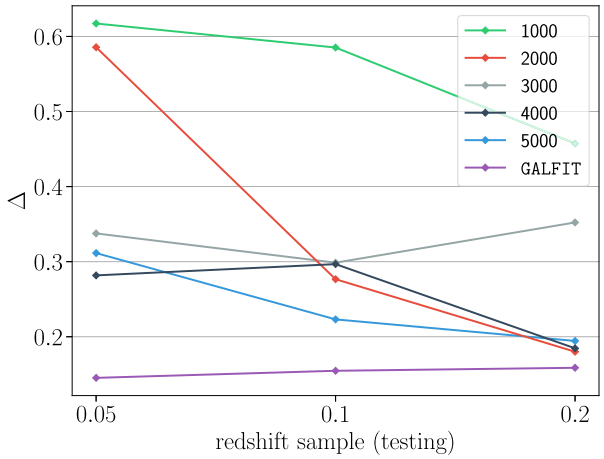

Figure 14. We train PSFGAN on training sets with different sizes 1000, 2000, 3000, 4000, and 5000. As expected the MAD score increases with decreasing training set size. We conclude that we indeed need approximately 5000 galaxies to be able to train PSFGAN.

cut-outs of the images. We cut out a quadratic box around the centre of the galaxy and we chose the length of the box by hand such that the galaxy fills the cut-out (therefore we have different box lengths for the different redshift samples). In order to get an intuition for the significance of the different performances we also compute the SSIM between the original galaxy image and the image with the added PS. After plotting each individual SSIM we calculate the median in eight bins of contrast ratio and connect the median points with a straight line.

For the sample of mixed morphologies we find results consistent with the analysis of magnitude recovery. We observe that only above contrast ratio $R \sim 1.8$ PSFGAN has a higher median SSIM than GALFIT. For the spiral galaxies we find that PSFGAN has a higher SSIM already for lower contrast that in the comparison sample of mixed morphologies. We conclude that PSFGAN is less confused by spirals arms.

3.5 Dependence on the size of the training set

To test the dependence of PSFGAN on the size of the training set we train (for each redshift) on training sets of size 1000, 2000, 3000, 4000, and 5000 images at different redshifts.. We then evaluate on the test samples and compute the MAD of the recovery ratio from 1 (which we defined as Δ). Fig. 14 shows how the different models perform. As expected, decreasing the training set leads to a decrease in accuracy.

3.6 Performance on low-quality data

The large amount of high-quality imaging data provided by SDSS make it easy to train a GAN. For many applications, the data may be noisier and the resolution poorer. Moreover, finding 5000 galaxies for the training set is not necessarily feasible for many surveys and wavelengths.

We now show that models trained on SDSS data can perform well on lower quality data. We train a model on degraded SDSS images and compare it to GALFIT and to the model trained on non-degraded images. We compare the models by evaluating them on a degraded test sample. To degrade the images we convolve the original image with a Gaussian kernel of size 5×5 pixels and FWHM, $\mathrm{FWHM}_{\mathrm{kernel}}$. We then add white noise with a variance such that the noise variance of the degraded image σ_{d} is larger than the initial noise variance σ_{i} of the original image. For each redshift we create three differently degraded tests with $(\mathrm{FWHM}_{\mathrm{kernel}}, \sigma_{\mathrm{d}}/\sigma_{\mathrm{i}}) = (1.2, 1.5)$, $(1.2, 1.8)$, $(2.0, 1.8)$. The way we degrade the training PSF is different from the way we degrade the PSF in the test sets. For the test sets we convolve the PSF image obtained by median combining stars with the same kernel and add the resulting image to the degraded galaxy image. We do not add white noise to the PSF image as we already add noise to the whole original image.

In the training set we degrade the PSF image obtained from the SDSS tool by applying the same transformation as for the original images. Then we fit the degraded PSF image with two Gaussians. Fitting three Gaussians is not possible here because the convolution smoothes out the images.

To compare the models we again use the one-dimensional score Δ from Section 2.3. We estimate the performance of the model trained on non-degraded images and the model trained on degraded images by evaluating them on a degraded test set. We then run GALFIT on the degraded test set where we provide a degraded PSF image as input. To get the input PSF we perform the same steps as for creating the degraded PSF in the training sets. We apply Gaussian blurring and add white noise to the image that is outputted by the SDSS PSF tool and then fit the resulting image with two Gaussians. We choose the variance of the white noise such that the noise of the PSF image gets increased by a factor $\sigma_{\mathrm{d}}/\sigma_{\mathrm{i}}$.

Fig. 15 shows the scores for the three degraded test sets and compares them to the performance of PSFGAN and GALFIT on the non-degraded test set. The plots show that both GALFIT and PSFGAN have

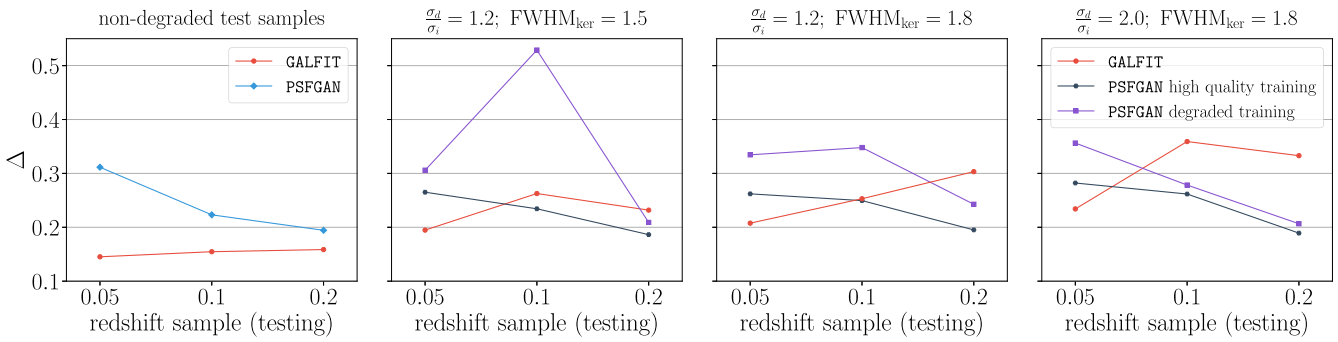

Figure 15. We degrade the test sample of each redshift by applying convolution with a Gaussian kernel and adding white noise. We create three different degraded test sets according to different kernel FWHMs and white noise variances. The quantity $\sigma_{\mathrm{d}}/\sigma_{\mathrm{i}}$ is the noise variance of the degraded image divided by the noise variance of the original image.

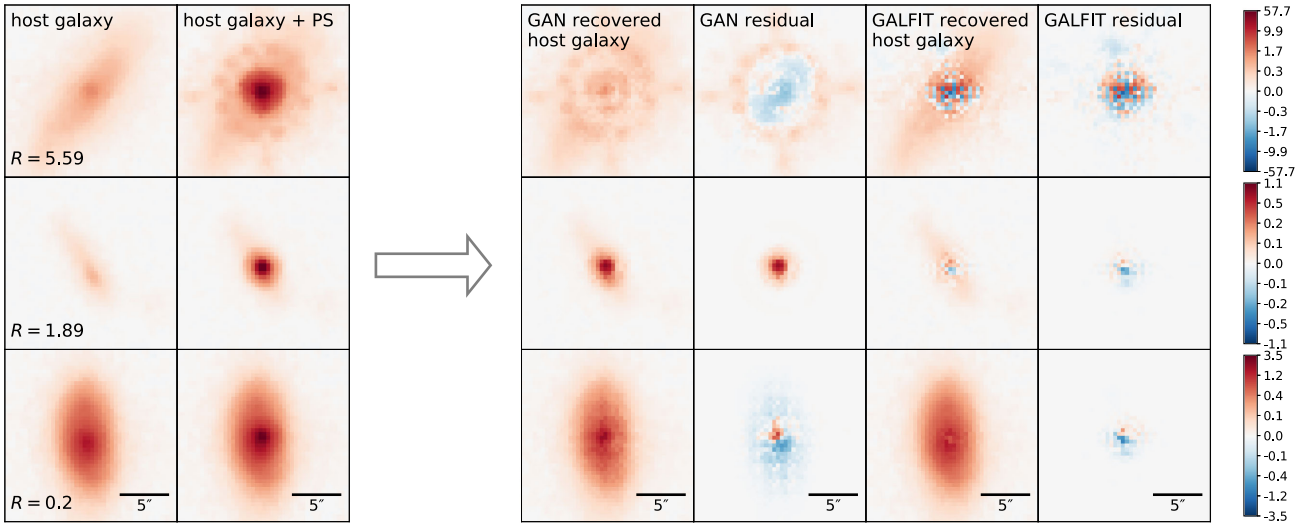

Figure 16. Examples of *Hubble* WFC3 images in the *F160W* filter of galaxies with $z \in [0.4, 0.5]$. We randomly choose three examples with different contrast ratios. The format of the plots is the same as in Fig. 4.

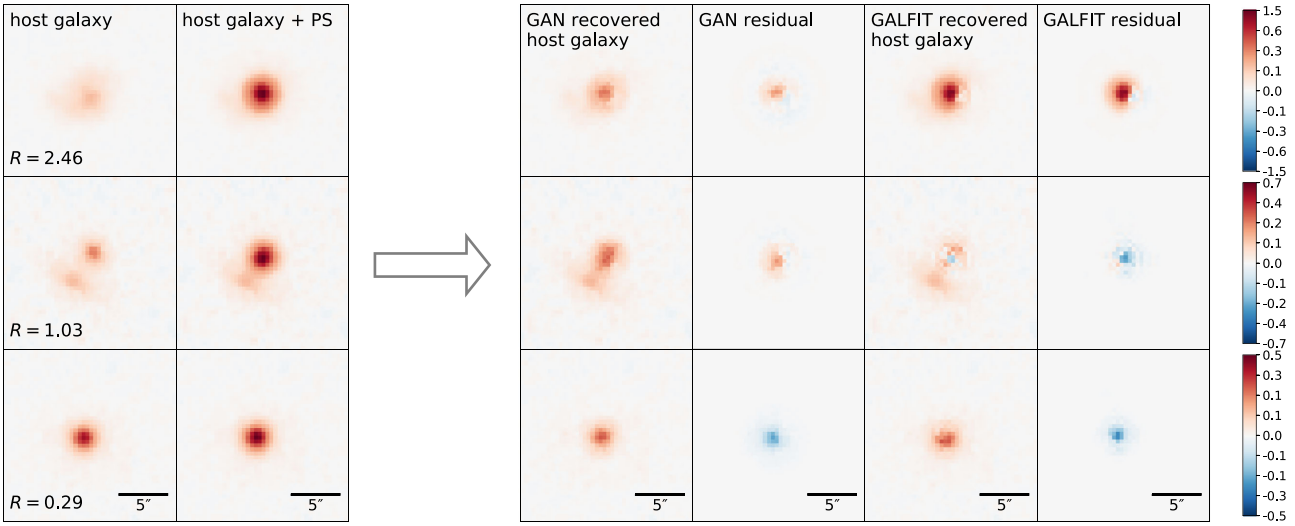

Figure 17. Examples of *Hubble* WFC3 images in the *F160W* filter of galaxies with $z \in [1.0, 1.5]$. We randomly choose three examples with different contrast ratios. The format of the plots is the same as in Fig. 4.

a larger Δ if they are run on the degraded test samples. However PSFGAN is more stable. For the non-degraded images GALFIT has a lower score for all three redshifts. For the most strongly degraded images ($\sigma_d/\sigma_i = 2.0$, $\mathrm{FWHM_{ker}} = 1.8$) GALFIT only has a lower score for redshift $z = 0.05$. For the other two samples both PSFGAN models have a lower score.

3.7 Applying PSFGAN to *Hubble* data

To demonstrate that PSFGAN can be used even if there is not enough training data available, we apply it to the Great Observatories Origins Deep Survey-South (GOODS-S) Wide Field Camera 3 (WFC3) data in the *F160W* filter (Grogin et al. 2011; Koekemoer et al. 2011). We use the fully calibrated, drizzled images. We create two test sets with different redshift ranges. We use the GOODS-S Cosmic Assembly Near-IR Deep Legacy Survey (CANDELS) stellar mass catalogue (Santini et al. 2015) to select detections with SEXTRACTOR's flag 'STAR_CLASS' <0.8 and observed AB magnitude in the *F160W* filter $m < 24$. We exclude detections with 'AGN_FLAG' <0.1. This yields a set consisting of 164 galaxies with $z \in [0.4, 0.5]$ and another set consisting of 195 galaxies with $z \in [1.0, 1.5]$. We simulate the AGN point sources by stacking 10–30 stars from the neighbourhood of the galaxy. We combine the stacked stars by taking the weighted median in each pixel where we distribute the weights according to the signal-to-noise ratio (S/N).

We then evaluate our pre-trained PSFGAN models and compare them to the GALFIT script we described in Section 2.4. The PSF image we provide as input for GALFIT is a cut-out of the brightest star with S/N > 100 we can find in the whole field. In Figs 16 and 17 we show example images of the original host galaxy, the host galaxy with the PS in its centre, PSFGAN and GALFIT recovered host galaxies, and both method's residuals. The examples show that PSFGAN is not able to subtract the extended wings of the *Hubble* PSF that is intuitive given the fact that it was trained on the SDSS PSF.

Fig. 18 shows the PS magnitude errors and the host magnitude errors for the test sample with $z \in [0.4, 0.5]$ and $z \in [1.0, 1.5]$. For all

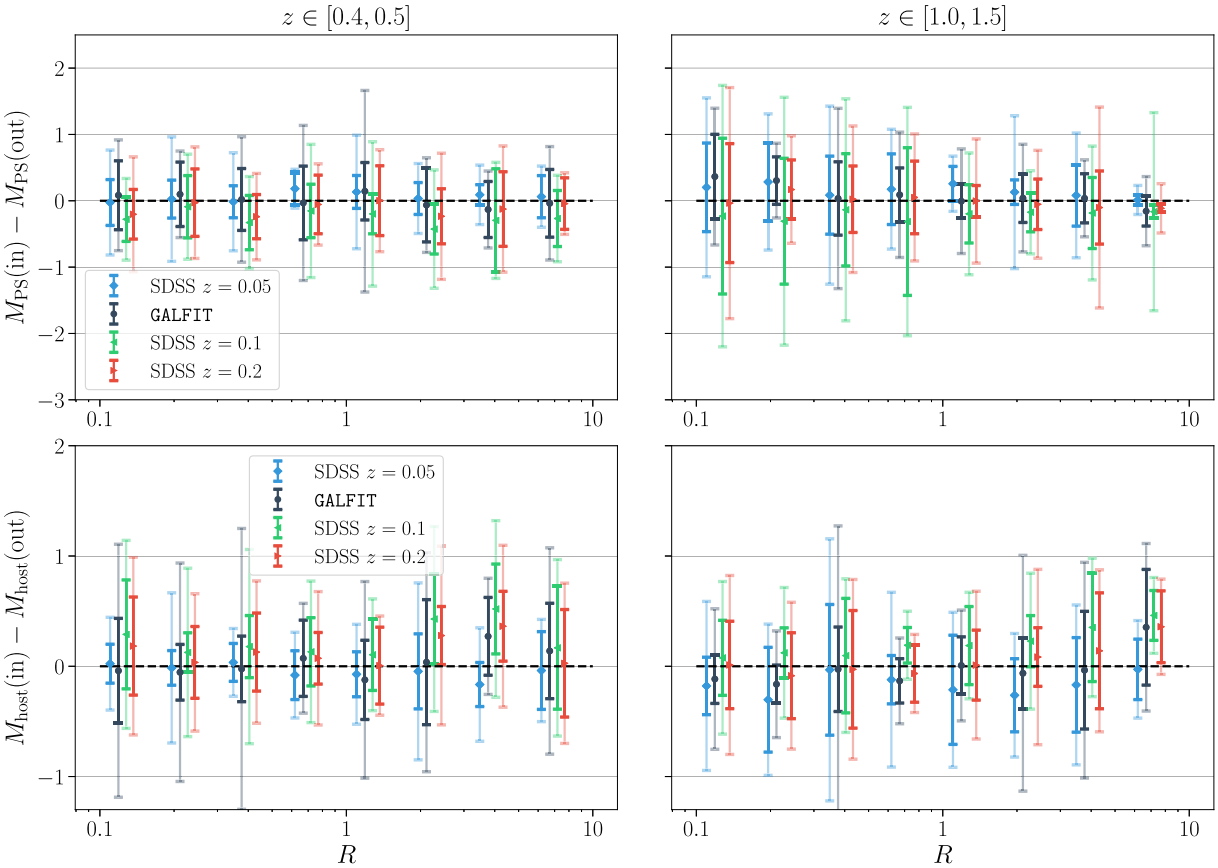

Figure 18. PS magnitude and host magnitude errors of evaluation on *Hubble* WFC3 galaxies in two different redshift samples (left- and right-hand plots). The plots show that the PSFGAN models trained on SDSS data (with the SDSS PSF) have similar bias and variance as the GALFIT script we used.

models we exclude 18 galaxies from the left-hand plots and 30 from the right-hand plots because GALFIT crashed on them. To compute the medians and the percentiles we only use those galaxies where the recovered flux is positive for both the PS and the host galaxy for all of the models. For some cases (<10 per cent) there is another source within the restricted box we use to compute the recovered PS flux. In the case where PSFGAN increases the brightness of this close source the computation of the recovered PS flux can result in a negative flux value (<2 per cent). The recovered host galaxy flux can be negative if either PSFGAN or GALFIT massively oversubtracts (<3 per cent for GALFIT and no observed cases for PSFGAN). All in all we have to exclude another four galaxies for $z \in [0.4, 0.5]$ and six for $z \in [1.0, 1.5]$.

The tests show that the models pre-trained on SDSS data can indeed be applied to different data and they even have accuracy comparable to GALFIT. Evaluated on the test sample with $z \in [0.4, 0.5]$, the $z \sim 0.05$ SDSS model has a smaller scatter and similar systematic error as our GALFIT script. At $z \in [1.0, 1.5]$ it is difficult to read a significant difference by eye just from the plots of the magnitude errors. Therefore we also list the Δ scores in Table 5. At first we notice that the score is higher for all models than if they are evaluated on SDSS data. The Δ is increased by a factor of 1.1 for the $z \sim 0.05$ model, by a factor of 1.7 for the $z \sim 0.1$ model, and by a factor of 2.0 for the $z \sim 0.2$ model. Evaluation on the high-redshift test sample yields an increase by factor of 2.1, 2.2, and 2.8 for the respective model of redshift $z \sim 0.05$, 0.1, and 0.2. However GALFIT's score increases as well and a thorough comparison reveals

Table 5. For both *Hubble* WFC3 test sets we compute the Δ score for GALFIT and the SDSS models trained on the three redshift samples. (Before computing Δ we exclude all the galaxies where GALFIT crashed in the results of the PSFGAN.) We find that all pre-trained models have a lower Δ for both test samples.

	$z \in [0.4, 0.5]$	$z \in [1.0, 1.5]$
GALFIT	0.51	0.68
$z \sim 0.05$ SDSS model	0.34	0.64
$z \sim 0.1$ SDSS model	0.37	0.49
$z \sim 0.2$ SDSS model	0.38	0.55

that all the SDSS models have a lower score than our GALFIT script for both test samples.

4 DISCUSSION

We have shown that GANs can be used to make photometric measurements. PSFGAN is able to separate AGN point sources from their host galaxies. We have shown that PSFGAN intuitively learns the light distribution of galaxies and applies this knowledge to subtract the PS. For contrast ratios above $R = 1.8$ it recovers PS and host galaxy fluxes with a smaller median magnitude error and a lower scatter than a single Sérsic+PS fit performed by GALFIT. We observe that for low contrast ratios ($R < 1.8$) PSFGAN's scatter in PS magnitude recovery is 1.6–4.7 times larger than GALFIT's scatter and for high

contrast ratios ($R > 1.8$) GALFIT's scatter is up to five times the scatter of PSFGAN. We have found that – in terms of SSIM – PSFGAN can recover host galaxy structure of spiral galaxies at least as good as a single Sérsic+PS fit performed by GALFIT while being better with higher contrast ratios. For $z \sim 0.05$ and $\sim$0.1 PSFGAN has a higher median SSIM already for $R = 0.6$. To conclude that PSFGAN can handle complicated morphologies better than parametric fitting in batch mode further tests should be conducted.

Parametric fitting is very powerful for well-resolved galaxies and low contrast ratios. However it struggles at high contrast ratios $R > 1.8$ because of the degeneracy between PS magnitude and host magnitude. Indeed, in this contrast range, GALFIT artificially increases the Sérsic index that causes the PS to be underestimated. This behaviour is documented in the literature (Kim et al. 2008; Koss et al. 2011).

The fact that PSFGAN performs well at high contrast ratios makes it a promising tool for studying AGN and their host galaxies at higher redshift where classical methods tend to break down. Indeed with increasing redshift the contrast ratio tends to be higher as the intrinsic emission emerges from a bluer part of the SED where the AGN is dominant. Also the host galaxy is affected by surface brightness dimming, while the PS is not (Falomo, Kotilainen & Treves 2000). This again increases the probability of finding high contrast systems with increasing redshift.

We have shown that PSFGAN is more stable with noisier and lower resolution imaging data. Evaluated on differently degraded data we find that GALFIT always has a lower Δ than PSFGAN for $z \sim 0.05$. However for $z \sim 0.1$ and $\sim$0.2 the accuracy of GALFIT declines faster (with the decline in quality) than PSFGAN's accuracy. For a kernel width $\mathrm{FWHM_{ker}} = 1.8$ and noise variance $\sigma_\mathrm{d} = 2.0\sigma_\mathrm{i}$, the Δ score of GALFIT increases by more than a factor of 2 compared to the evaluation on non-degraded images. The Δ of PSFGAN (trained on high-quality data) increases by a factor of less than 1.2. Furthermore we find that PSFGAN trained on non-degraded images has a lower Δ on degraded images than if it was trained on degraded images. We conclude that it can better learn the light distribution of galaxies if the training data are of high quality.

We find that it is indeed necessary to have a training set size of $\sim$5000 images. However if not enough data are available and a training cannot be performed, the user can also apply the PSFGAN trained on SDSS data. We demonstrate that our pre-trained models can be applied on *Hubble* IR data up to redshift $z = 1.5$. Although the accuracy is lower on this data than it was on SDSS data, it compares well to our GALFIT script. For the *Hubble* test sample with $z \in [0.4, 0.5]$ the best model is the one trained on $z \sim 0.05$ SDSS data. Its Δ score is 67 per cent of that ofGALFIT. For the *Hubble* test sample with $z \in [1.0, 1.5]$ the best model is the one trained on $z \sim 0.1$ SDSS data with a score of 72 per cent of that of GALFIT. We find that in agreement with Section 3.2 the $z \sim 0.05$ SDSS model performs best on the more nearby sample, and the $z \sim 0.1$ SDSS model performs best on the more distant higher redshift sample.

The inference phase of PSFGAN is faster on a CPU and one can accelerate it further by running it on GPUs. Run on a Macbook Air with a 1.7 GHz Intel Core i5 CPU and 4 GB RAM it is $\sim$3.6 times faster than GALFIT run on the same machine. By running PSFGAN on GPUs it can be accelerated such that its inference phase is more than $\sim$40 times faster than GALFIT.[7] The strength of PSFGAN however lies in its ability to apply the same trained model to many images. If a low number of galaxies is considered, GALFIT may have a speed advantage due to PSFGAN's training time of $\sim$8 h (on a GPU). However, e.g. for 10^6 galaxies the total runtime of PSFGAN (training+evaluation) is only 2.5 per cent of GALFIT's runtime.

The lack of input parameters during evaluation is another strong advantage of PSFGAN. Unlike parametric fitting methods that are very sensitive on their input parameters, PSFGAN is very robust and requires no human interaction once it is trained. Also it requires fewer physical assumptions than parametric fitting. The only physical knowledge that goes into PSFGAN is the training PSF. A user has to model the PSF of the data to simulate the point sources in the training set. We have however found that for using PSFGAN it is less important to correctly model the PSF than for using GALFIT. PSFGAN is thus especially powerful to analyse ground-based data where the seeing is variable.

Although we have trained PSFGAN to subtract AGN point sources in SDSS data, it is neither limited to AGN nor to SDSS data. PSFGAN is a general framework for subtracting point sources from CCD images in an automated way. In order to apply PSFGAN to some specific case a user should go through the following procedure.

(i) Create a training set consisting of pairs of images (original, original+point source). We used real observations of galaxies but if there is not enough data available a user could also simulate the ground truth (e.g. use simulated galaxies). If ground-based images are used, make sure PSFGAN sees a variety of PSFs during the training.

(ii) Look at the histogram of pixel values of the whole training set. Then decide which stretching function might be appropriate. We recommend starting with asinh and trying different scale factors.

(iii) Test the set-up on a separate testing set to estimate the accuracy.

We propose a number of applications of our method. One task that PSFGAN may be suited for is subtraction of foreground stars from galaxy images. The only difference from subtracting quasar point sources is the position of the point source relative to the galaxy. Another task where PSFGAN could be applied to is separating supernovae from their host galaxies. Given that this is usually done by fitting galaxy templates, PSFGAN could both simplify and accelerate those measurement processes. Lastly, we propose to apply our method to quasar spectra. Like images of quasar host galaxies, their spectra are as well contaminated the AGN. Indeed the architecture of PSFGAN can easily be adapted for taking spectra as input. However the training process might be less straightforward than in our case where the quasar was a point source and thus had a (more or less) constant shape.

The code of PSFGAN is described at http://space.ml/proj/PSFGAN and available at https://github.com/SpaceML/PSFGAN/. Moreover we will provide the pre-trained models at $z \sim 0.05$, $\sim$0.1, and $\sim$0.2.

ACKNOWLEDGEMENTS

We thank the anonymous referee for helpful feedback. KS, LFS, and AKW acknowledge support from Swiss National Science Foundation Grants PP00P2_138979 and PP00P2_166159, and KS also from the ETH Zurich Department of Physics. CZ and the DS3Lab gratefully acknowledge the support from the Swiss National Science Foundation NRP 75 407540_167266, IBM Zurich, Mercedes-Benz Research & Development North America, Oracle Labs, Swisscom, Zurich Insurance, Chinese Scholarship Council, the Department of Computer Science at ETH Zurich, and the cloud computation resources from Microsoft Azure for Research award program. MK acknowledges support from NASA through ADAP award

[7] These numbers should serve as rough estimation as they are specific for our implementation and hardware.

NNH16CT03C and the Swiss National Science Foundation through the Ambizione fellowship grant PZ00P2 154799/1. Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the US Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS web site is http://www.sdss.org/. The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington. Finally this work is based on observations taken by the CANDELS Multi-Cycle Treasury Program with the NASA/ESA *HST*, which is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS5-26555.

REFERENCES

Bahcall J. N., Kirhakos S., Schneider D. P., 1995, ApJ, 450, 486
Bahcall J. N., Kirhakos S., Saxe D. H., Schneider D. P., 1997, ApJ, 479, 642
Barden M., Häußler B., Peng C. Y., McIntosh D. H., Guo Y., 2012, MNRAS, 422, 449
Baron D., Poznanski D., 2017, MNRAS, 465, 4530
Bennert N., Canalizo G., Jungwiert B., Stockton A., Schweizer F., Peng C. Y., Lacy M., 2008, ApJ, 677, 846
Bertin E., Arnouts S., 1996, A&AS, 117, 393
Blanton M. R. et al., 2017, AJ, 154, 28
Böhm A. et al., 2013, A&A, 549, A46
Boyce P. J., Disney M. J., Bleaken D. G., 1999, MNRAS, 302, L39
Chang Y.-Y., van der Wel A., da Cunha E., Rix H.-W., 2015, ApJS, 219, 8
Collinson J. S., Ward M. J., Done C., Landt H., Elvis M., McDowell J. C., 2015, MNRAS, 449, 2174
Dieleman S., Willett K. W., Dambre J., 2015, MNRAS, 450, 1441
Falomo R., Kotilainen J., Treves A., 2000, The Messenger, 101, 15
Gabor J. M. et al., 2009, ApJ, 691, 705
George D., Huerta E. A., 2018, Phys. Lett. B, 778, 64
Goodfellow I. J., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y., 2014, NIPS, 2672 (arXiv:1406.2661)
Goulding A. D., Alexander D. M., Lehmer B. D., Mullaney J. R., 2010, MNRAS, 406, 597
Grogin N. A. et al., 2011, ApJS, 197, 35
Hernán-Caballero A., Alonso-Herrero A., Pérez-González P. G., Cava A., Cardiel N., the SHARDS Team 2013, preprint (arXiv:1305.0641)
Hooper E. J., Impey C. D., Foltz C. B., 1997, ApJ, 480, L95
Isola P., Zhu J.-Y., Zhou T., Efros A. A., 2017, IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, p. 5967
Kim M., Ho L. C., Im M., 2006, ApJ, 642, 702
Kim M., Ho L. C., Peng C. Y., Barth A. J., Im M., 2008, ApJS, 179, 283
Kingma D. P., Ba J., 2015, 3rd International Conference for Learning Representations, San Diego
Kirhakos S., Bahcall J. N., Schneider D. P., Kristian J., 1999, ApJ, 520, 67
Koekemoer A. M. et al., 2011, ApJS, 197, 36
Koss M., Mushotzky R., Veilleux S., Winter L. M., Baumgartner W., Tueller J., Gehrels N., Valencic L., 2011, ApJ, 739, 57
Lehnert M. D., van Breugel W. J. M., Heckman T. M., Miley G. K., 1999, ApJS, 124, 11
Lintott C. J. et al., 2008, MNRAS, 389, 1179
Lintott C. et al., 2011, MNRAS, 410, 166
McLeod K. K., Rieke G. H., 1995, ApJ, 454, L77
Matsuoka Y., Strauss M. A., Price T. N., III, DiDonato M. S., 2014, ApJ, 780, 162
Michałowski M. J., Hayward C. C., Dunlop J. S., Bruce V. A., Cirasuolo M., Cullen F., Hernquist L., 2014, A&A, 571, A75
Peng C. Y., Ho L. C., Impey C. D., Rix H.-W., 2002, AJ, 124, 266
Peng C. Y., Ho L. C., Impey C. D., Rix H.-W., 2010, AJ, 139, 2097
Pierce C. M. et al., 2010, MNRAS, 405, 718
Reed S., Akata Z., Yan X., Logeswaran L., Schiele B., Lee H., 2016, in Balcan M. F., Weinberger K. Q., eds, Proc. The 33rd International Conference on Machine Learning, Vol. 48. PMLR, New York, p. 1060
Reines A. E., Volonteri M., 2015, ApJ, 813, 82
Robotham A. S. G., Taranu D. S., Tobar R., Moffett A., Driver S. P., 2017, MNRAS, 466, 1513
Santini P. et al., 2012, A&A, 540, A109
Santini P. et al., 2015, ApJ, 801, 97
Schawinski K. et al., 2006, Nature, 442, 888
Schawinski K., Treister E., Urry C. M., Cardamone C. N., Simmons B., Yi S. K., 2011, ApJ, 727, L31
Schawinski K., Zhang C., Zhang H., Fowler L., Santhanam G. K., 2017, MNRAS, 467, L110
Shimizu T. T., Mushotzky R. F., Meléndez M., Koss M., Rosario D. J., 2015, MNRAS, 452, 1841
Simmons B. D., Urry C. M., 2008, ApJ, 683, 644
GOODS TeamSimmons B. D., Van Duyne J., Urry C. M., Treister E., Koekemoer A. M., Grogin N. A., 2011, ApJ, 734, 121
Sola J., Sevilla J., 1997, IEEE Trans. Nucl. Sci., 44, 1464
Sreejith S. et al., 2018, MNRAS, 474, 5232
Stetson P. B., 1987, PASP, 99, 191
Stoughton C. et al., 2002, AJ, 123, 485
Tuccillo D., Huertas-Company M., Decencière E., Velasco-Forero S., Domínguez Sánchez H., Dimauro P., 2018, MNRAS, 475, 894
Vikram V., Wadadekar Y., Kembhavi A. K., Vijayagovindan G. V., 2010, MNRAS, 409, 1379
Vitale M. et al., 2013, A&A, 556, A11
Wang Z., Bovik A. C., Sheikh H. R., Simoncelli E. P., 2004, IEEE Trans. Image Processing, 13, 600
Yoon I., Weinberg M. D., Katz N., 2011, MNRAS, 414, 1625

This paper has been typeset from a TEX/LATEX file prepared by the author.