Realistic galaxy generation using Variational AutoEncoders

Thomas Sainrat

Marc Huertas–Company, Hubert Bretonnière, Alexandre Boucaud
Cécile Roucelle, Eric Aubourg, Bastien Arcelin

data intelligence institute of Paris

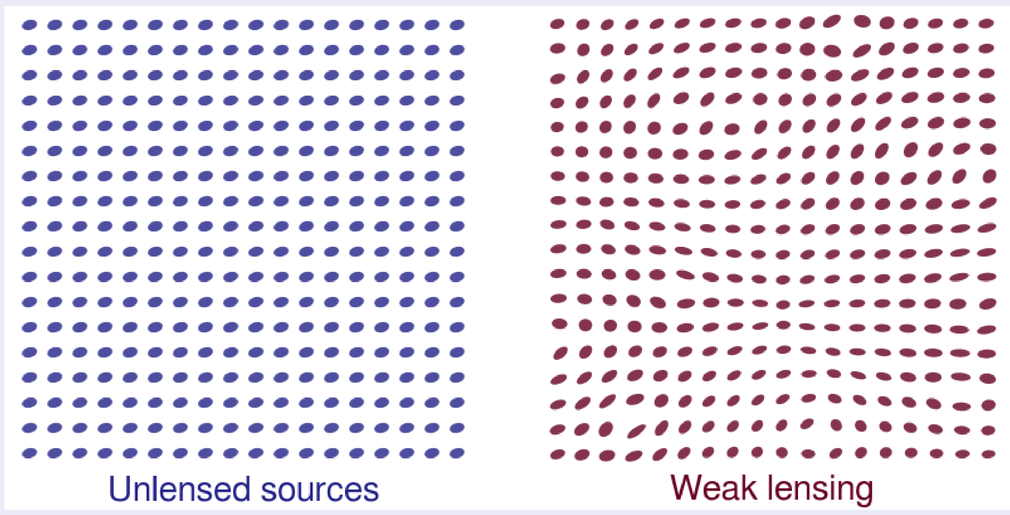
Instituto de Astrofísica de Canarias, Tenerife

thomas.sainrat@student-cs.fr

Motivations

One of the most important probes in cosmology today is weak lensing : when light passes near massive objects, it is deviated from its course according to Einstein's general relativity ; this effect is mostly apparent when observing distant galaxies. When the effect remains small (hence the *weak* lensing), we can consider that the galaxies have their shape modified by a small amount, called the *shear*. By measuring the shapes of the galaxies and their correlation (as the effect can only be measured statistically), we can recreate maps of the distribution of mass in the Universe, including the one from dark matter, and estimate cosmological parameters from them.

The main objective is thus to accurately measure the shape of galaxies ; this process needs to be very precise to avoid any kind of bias in the final results, and has to deal with the effects of the atmosphere, optics, noise and pixellisation from the camera, and biases due to the measurement algorithms. To adress those, simulations play a crucial role, both for calibrating the measurements and for testing algorithms, as we need to place ourselves in situations where we know the true shear.



Main work

Our goal is to simulate images of realistic galaxies, as we only know how to simulate galaxies using parametric models which cannot capture complex features in galaxies. Progress has already been made in this area by Lanusse et al. 2020 [1], using deep learning techniques (which we will develop in the next section) to generate realistic galaxies with user provided parameters. It was trained using images from Hubble in a catalog named COSMOS, which contains ~87000 monochromatic images of galaxies. We want to expand this work to generate images in several colors (bands), using the images from the CANDELS dataset, a polychromatic catalog of ~17600 galaxies, again from the Hubble telescope. Color information can be very important for several algorithms within the shape measurement pipeline, including for measuring the redshift of galaxies, and for deblending (separating several galaxies that overlap each other).
This presents additional challenges as we have to learn additional information on galaxies using a smaller dataset ; we also have to deal with the different observing conditions in the different bands, the cleaning of the dataset and missing data.

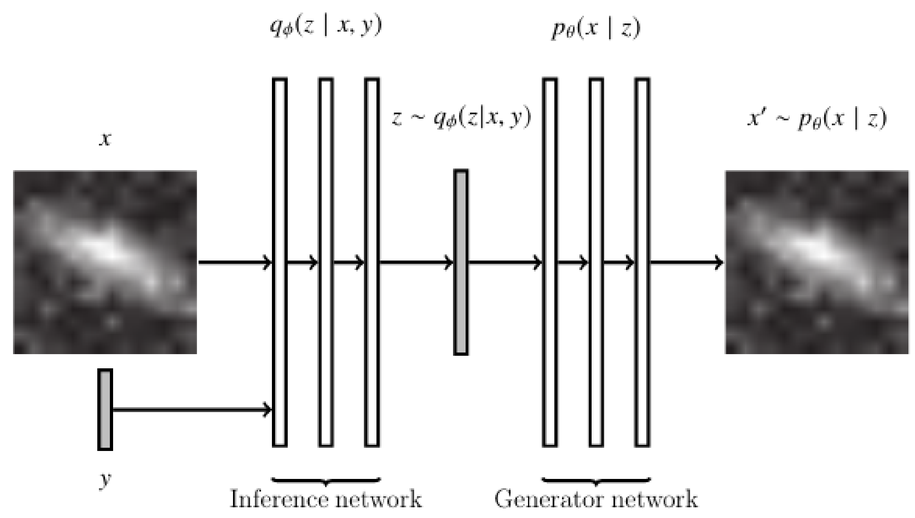
Variational AutoEncoders

In order to generate galaxy images, we want to rely on generative deep learning methods, which are mostly divided in two categories : Generative Adversarial Networks (GANs) and Variational AutoEncoders (VAE) ; we focus on the second one because it ensures a certain smoothness in the images, which is more important to us than getting very precise details. They use a "bottleneck" design like regular autoencoders, which basically compress the image in a low-dimensionnal space (called the latent space) and then reconstruct them ; however, instead of encoding the image as a single point in the latent space, the VAE encodes it as a distribution, from which you can sample to get some different possible reconstructions of the galaxy. It also forces the distribution to look like a Gaussian, ensuring that the latent space is smooth.

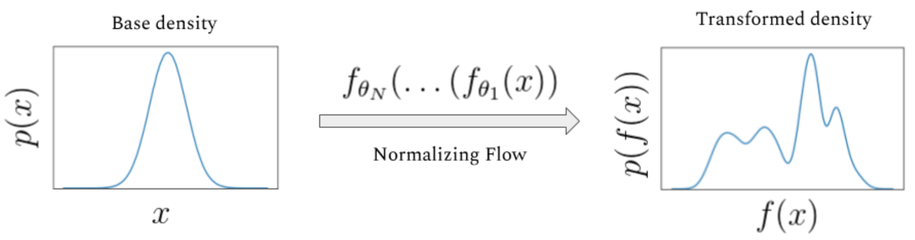
Normalizing Flows

On top of this, we use a normalizing flow (more precisely a Masked Autoregressive Flow), which is essentially a function transforming a distribution into another ; this flow can be conditionned with some parameters. This flow is used to transform a (n-dimensionnal) Gaussian into a distribution on the latent space matching the parameters we give to the flow.

The global workflow is then to provide parameters of interest (ellipticity,brightness,size,...), sample from a Gaussian, use the normalizing flow to transform the sample into a sample of the latent space of the VAE, and use the decoder part of the VAE to recover an image.

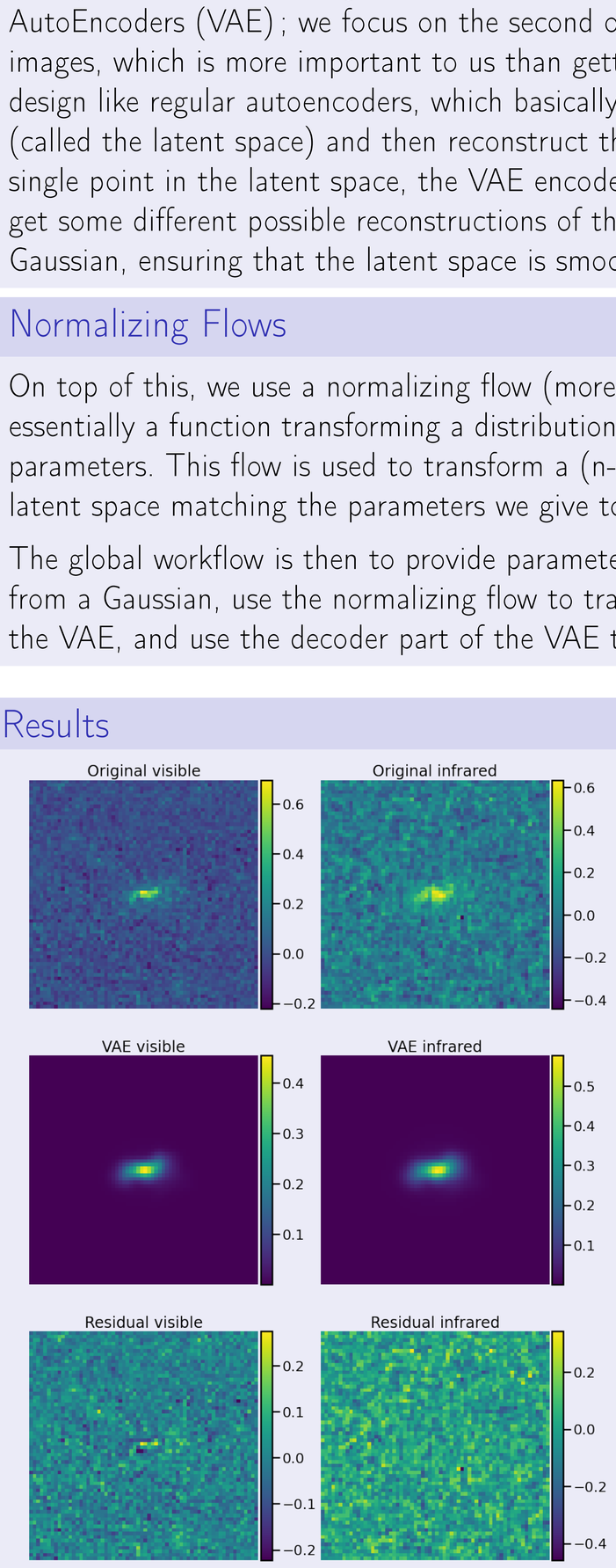


Architecture of a VAE (credit : Lanusse et al. 2020 [1])

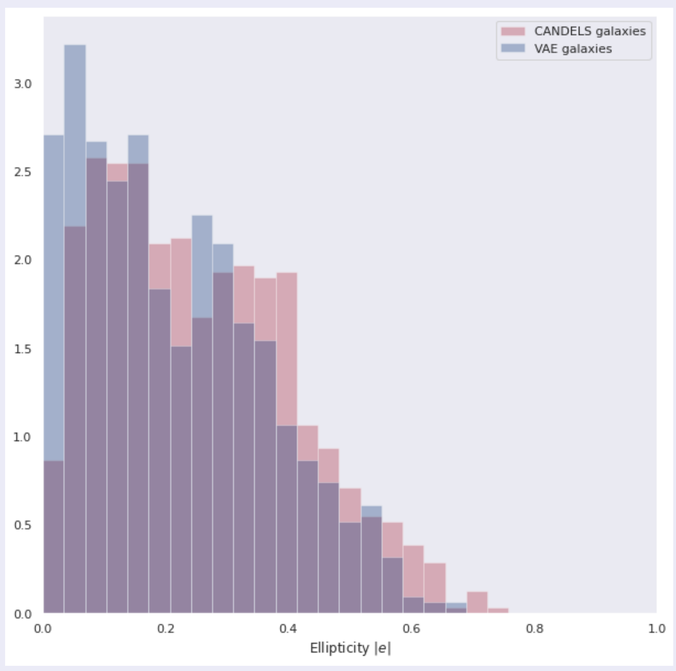


Representation of a normalizing flow (credit : Simon Boehm)

Results



Reproduction of a galaxy with VAE



Distributions of the ellipticity of galaxies

Future perspectives

While the shape information is correctly retrieved, we are currently working on the color information ; it seems to be at least partly retrieved but it is harder to interpret and to ensure that it is well calibrated, as it can depend on several other parameters such as the type of galaxy or the redshift. We also need to work more on the flow to complete the generative aspect of the work, as it currently tends to generate only round galaxies.
Once this is completed, we can look into integrating it in actual simulations ; in particular, I have been working on a framework called Blending ToolKit for generating blended galaxy images, which is one of the domains where color information can be decisive.

References

- François Lanusse, Rachel Mandelbaum, Siamak Ravanbakhsh, Chun-Liang Li, Peter Freeman, and Barnabás Póczos.
  Deep generative models for galaxy image simulations.
  Monthly Notices of the Royal Astronomical Society, 504(4) :5543–5555, 05 2021.

Project realized with the financial support of diiP, IdEx Université de Paris, ANR-18-IDEX-0001.