List the important parameters around which we are going to build our DB. Discuss the amount of comics needed/ the years of releas/ the color/line/shading artists/ the stories in question (interesting experience, test globalization on comics with characters never seen in the traning samples [that could

Create 1 (or more) DB

Creating the DB :

## **The questions**

* the impact of the year of release
* the impact of the on the art style
* the of the overall colorscheme

For the first db we wont mind any of these questions but will simply try to gather enough samples to have a first working model. From there we will test out different variations of databases to figure out what parameters that go in it’s creation (year of realeas of the CB, artists, overall colorscheme, presence of noise) affect our model in significant ways.

One of the main concerns we have as we are trying to build our db is that the classification we are trying to achieve is extremely vague! There are hundreds of thousands of comics published by both houses that range from world ending superhero plots to scoobydoo. Furthermore each publisher works with a wide array of artists colorists and graphic desingers that are not exclusive to one publisher. In fact many comic book artists have worked with both DC and Marvel. The elements that differentiate both publishing houses are mostly the character/heros in these comics that compose the IP of both houses. However, as stated before, these charaters all have been represented in a wide variety of ways.

Potential solutions to the variation in all possible inputs would be reducing the variantion in astetic by only selecting comics with the same pencilers/inkers/colorist, or making sure that each set of samples (marvel/Dc) are as varied when I comes to the artstyle and color palet so that the model wont associate one pub to a simple pattern (if we only inputted batman for DC during training, any dark color schemed comic would be interpreted as DC). We could also try to use image preprocessing to isolate important elements of our data (eg. only show the characters/ the backgrounds ect). However by doing this we may remove noise that could harm the model.

But maybe our dependencie on recognisable IP to differentiate both publishers will not show in the model.

We can at first try to focus on

We wrote a script that randomly creates n crops of a comic page scan to generate large amounts of samples without having to gather unrealistic amounts of comic books. The script simply takes as input the number of crops per pages and the resolution of said crops. Being that image translation changes entirely the information inputted in the model, each crop of a given image will be considered as a distinct sample from the las(even if the visual information stays relatively similar). When it will come to training our model we will have to calibrate the resolution and numbers of crops per page.