Kaggle 2018 Competition

**Useful information**

**Resources:**

Kaggle kernel Page: <https://www.kaggle.com/c/data-science-bowl-2018>

Video Introduction: <https://datasciencebowl.com/2018dsbtutorial/>

**TensorFlow:** we installed the TensorFlow 1.5.0, for more detail information <https://www.tensorflow.org/>

**Environment Configuration**

**[Windows 10]**

Creating a new python 3.5 environment in Anaconda. Then install:

TensorFlow (version 1.5.0).

OpenCV (version 3.3.1). Installed in Anaconda Navigator.

Sk-learn (version 0.19.1). Installed in Anaconda Navigator.

matplotlib (version 2.1.2). Installed in Anaconda Navigator.

pandas (version 0.22.0). Installed in Anaconda Navigator.

seaborn (version 0.8.1). This package is used for statistical data visualization. Installed in Anaconda Navigator.

**[IDE]**

Install Jupyter Notebook. Installed in Anaconda Navigator.

**EDA**

**Reference:**

[Some dataset exploration on Kaggle kernel]

1. Exploratory Analysis – *Jerry Thomas* <https://www.kaggle.com/jerrythomas/exploratory-analysis>
2. Exploratory Analysis - image stats – *Wesley Goi* <https://www.kaggle.com/etheleon/exploratory-analysis-image-stats>
3. Simple data science bowl 2018 exploration – *David Retana* <https://www.kaggle.com/davidretana/simple-data-science-bowl-2018-exploration>

Every one of the perspective is quite different. But they all look at the distribution of the nucleis. And they all notice the background problem.

**First look at the Dataset**

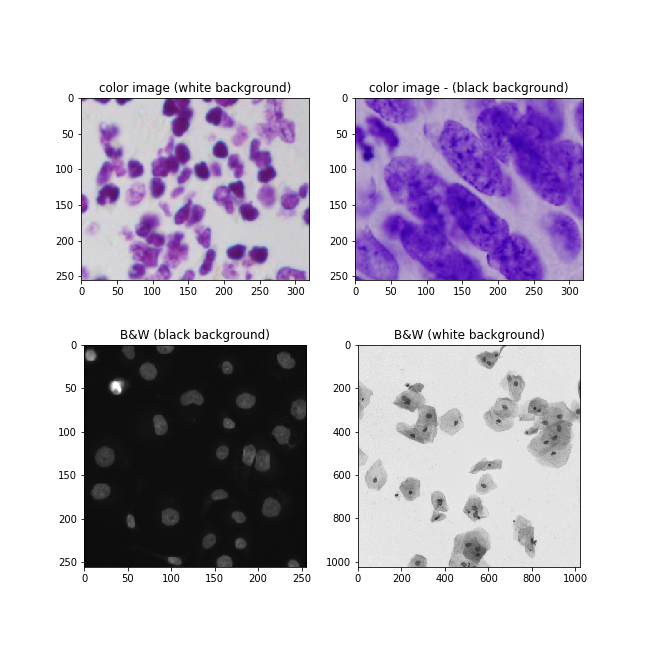
This file is to collect the training images and copy them into one single file. So we can actually see the data and have a better understanding of the pictures. Then we have a first look at the data set. We them display different types of the images with the following focuses:

* Color
* Number of nuclei
* Size and shape
* The different quality of the images

**Color**

Based on our observation, we sorted the images into four classes.

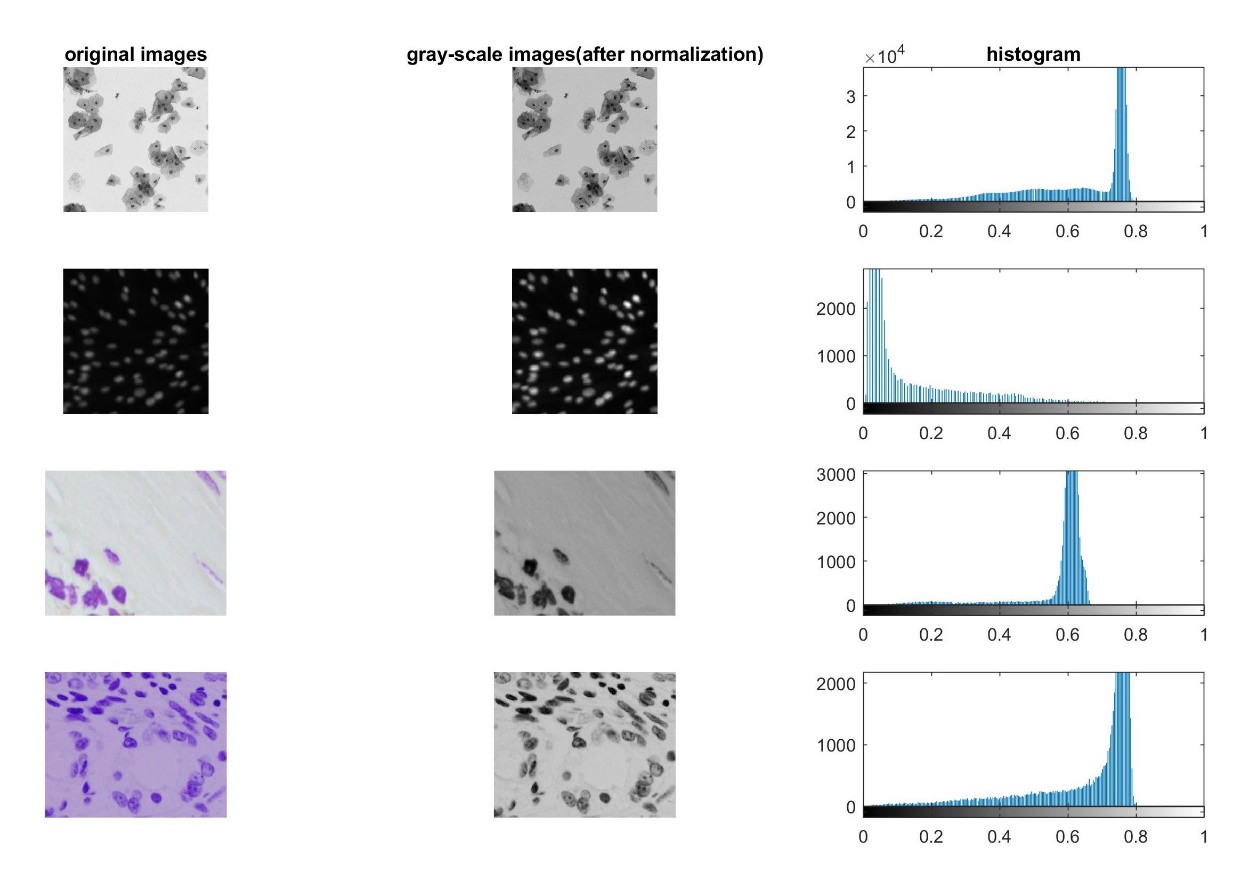
* Colored images with white background.
* Colored images with purple background.
* B&W images with black background.
* B&W images with white background.

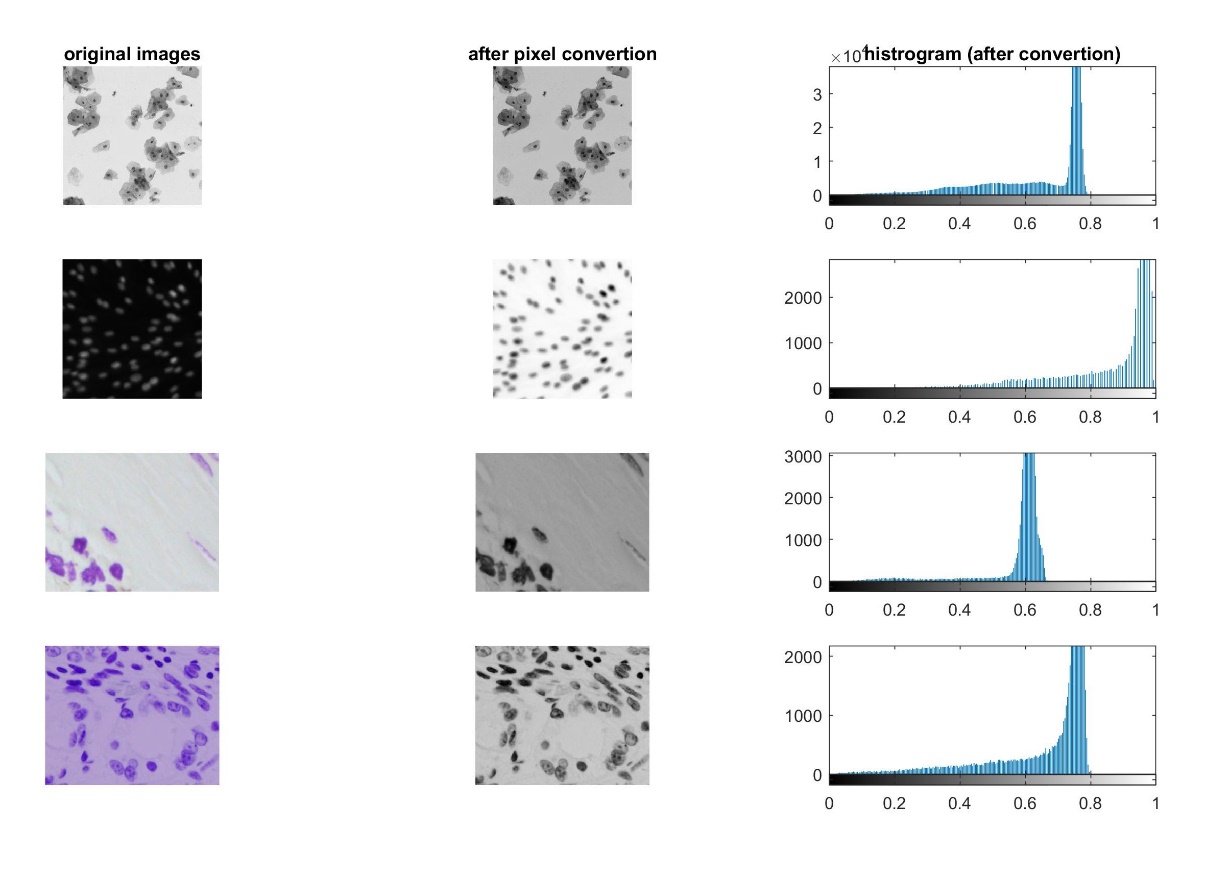


First we convert these color images into gray-scale images. When converting, we use:

to normalize the image.

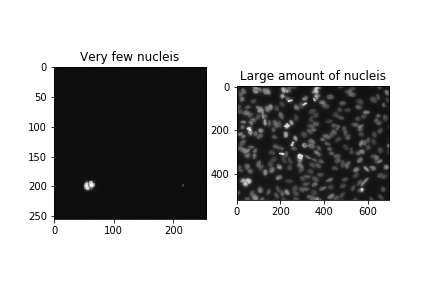
Then we look at the histograms of the four types of gray-scale images. We found that the histogram of B&W images with black background is different from the others. While others look very similar.





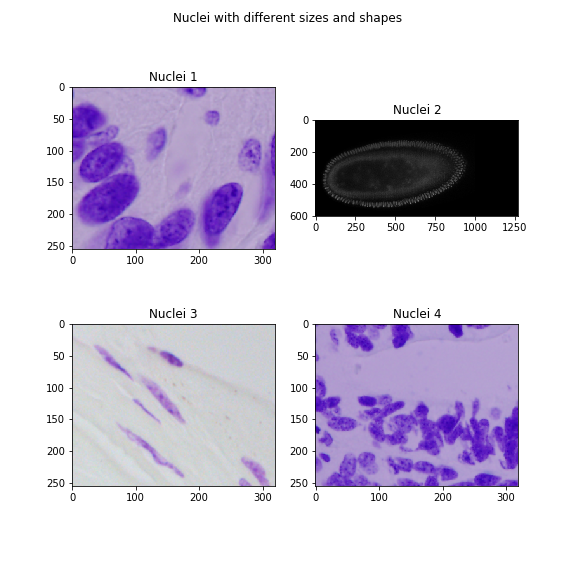
**Number of nuclei**

Number of the nuclei on every image varies. Some pictures only have few nuclei, while some have hundreds of nuclei.



**Size and shape**

The size and shape of the nuclei can be very different. Below shows some different sizes and shapes of the nuclei.

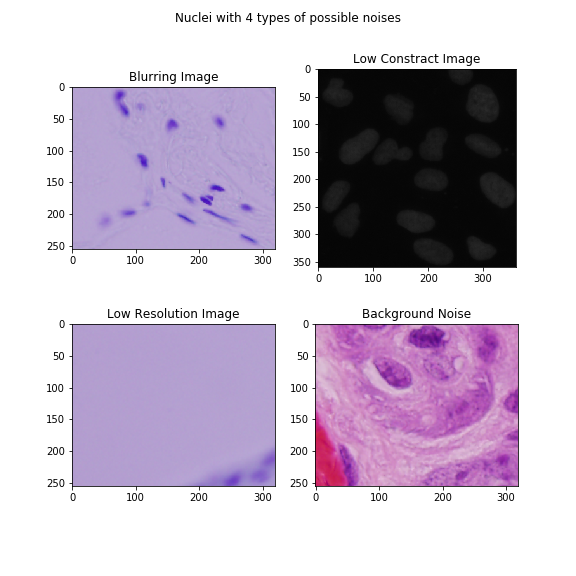


**The quality of images, possible noises**

According to the images, we identify four different types of the noises:

* blurring
* low resolution
* low contract
* background noise

The image below shows the four different types of the noises.

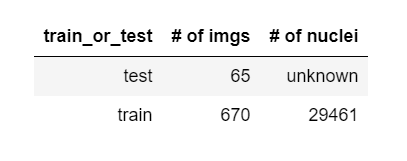


**Statistical Analysis of the Data**

After having a first glance at the data, we will do some statistical analysis on the dataset. We will going to explore more features in both training images and test images, such as the exact numbers of nuclei, the sizes of images, the distribution of nuclei and etc.

Most of our work are based on Jerry Thomas's instruction and work. Please see his Kaggle kernel here. <https://www.kaggle.com/jerrythomas/exploratory-analysis> We also add some our own work on it based on jerry’s work.

1. **Data Overview: How many images? How many nuclei?**



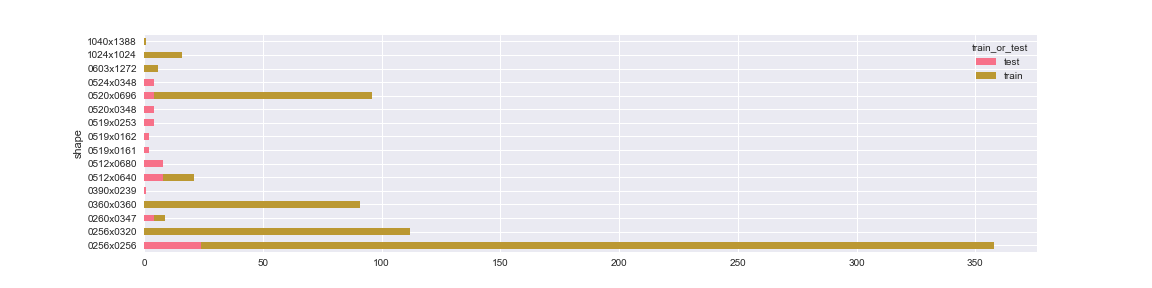
First, we tried to find out how many images in our training and test sets. As you can see from the table, the data set is not very big, there are only 670 images in training set and 65 images in test set. We know that the more data we can get, the better we can get during the training and learning process. So we probably need to somehow expand our training set later, in order to have better preformation.

And we also calculate the average nuclei per training image, the number is 44. All of these give us a general inspect into the dataset we get.

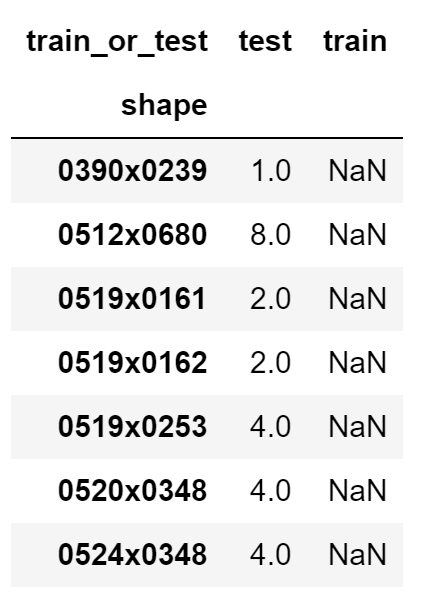
1. **Size matters: how many different types of the images are there?**

The sizes of the images are very important. From previous work (see the First look at the Dataset), we divide the images into 4 different types. Some images have white nuclei in black background, while some have black nuclei in white background. It would be good to convert them into one format.

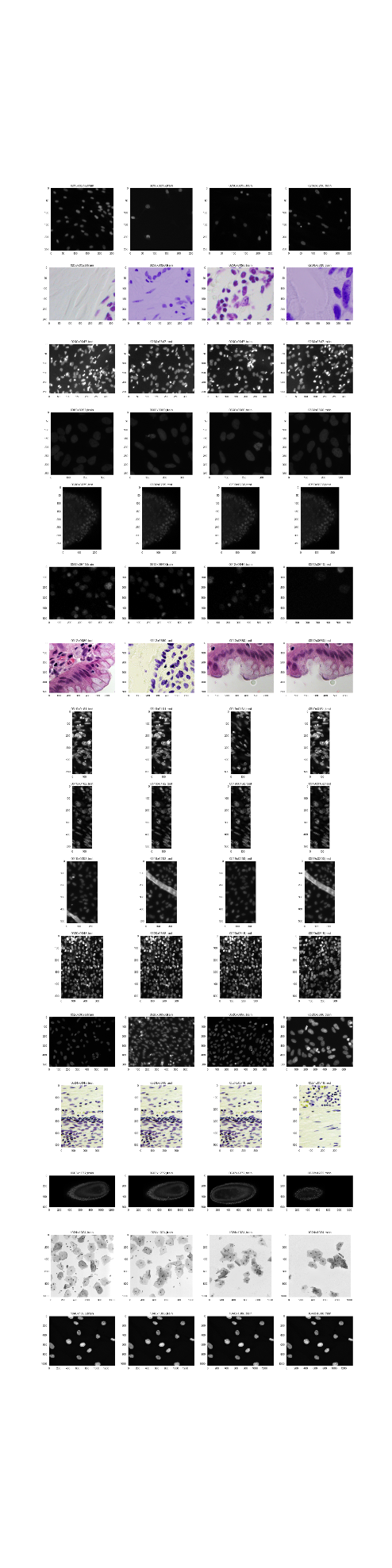
Based on Jerry’s method, we stored the properties of the image in .json file.



There are 16 different shapes of images, though the total different types of height and weight are 11 and 15 respectively. Nearly half of the training images are of size 256 x 256, and most of the test images are in this size. Notice that there are some sizes in test set but not in training set.

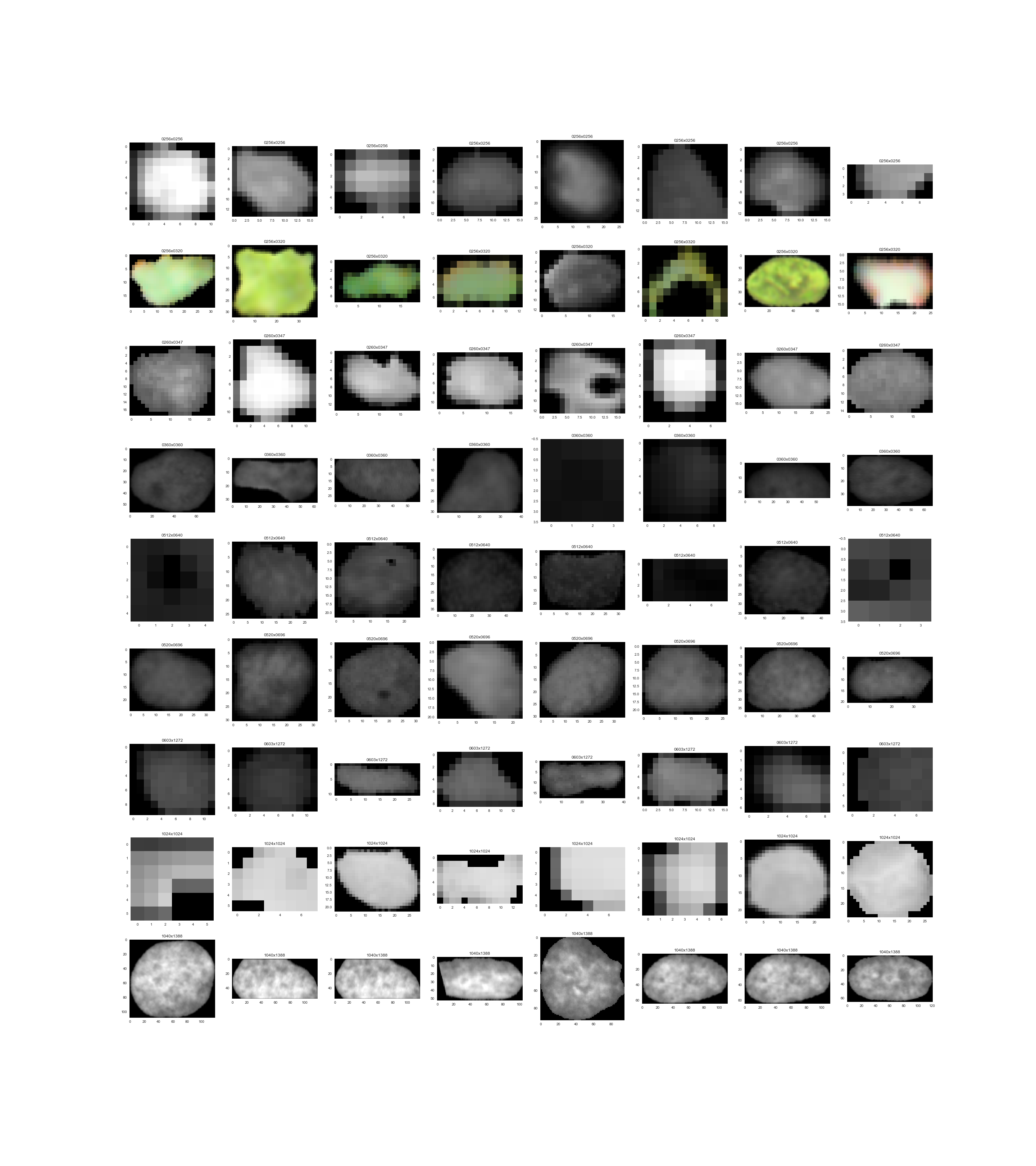


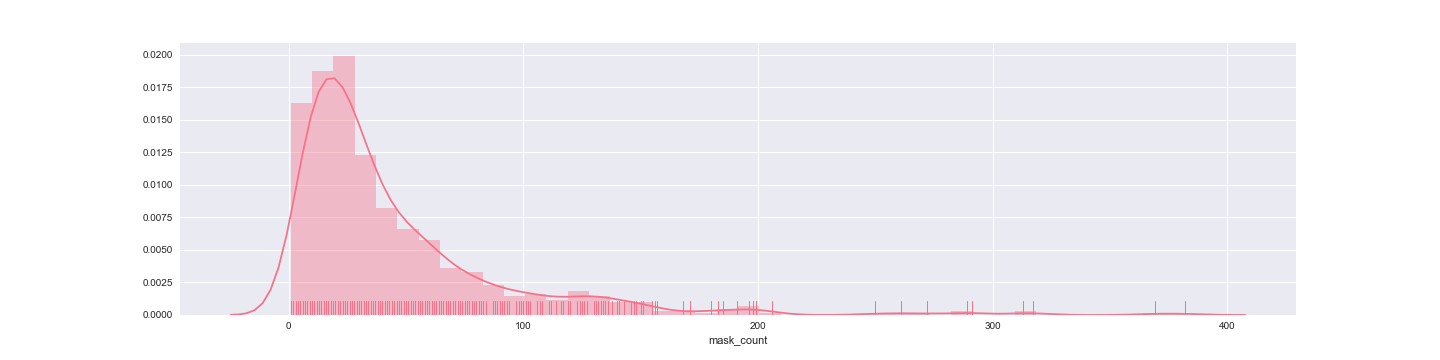
We also notice that the images within each type of size looks similar, maybe they are acquired using the same devices or by the same lab. The size of input is vary important in the training process and need to be fixed, so we may need to adjust or resize them for further usage.



1. **Nuclei: number and size**

In this section, we investigate some statistics of nuclei. The first thing we have noticed is that the shape and size of nuclei are not identical, though most of them are of circle shape. Here is what we got from the images.



The Number Distribution 

Size Matters: Width & Height Distribution

