Data Pre Processing Steps

Context

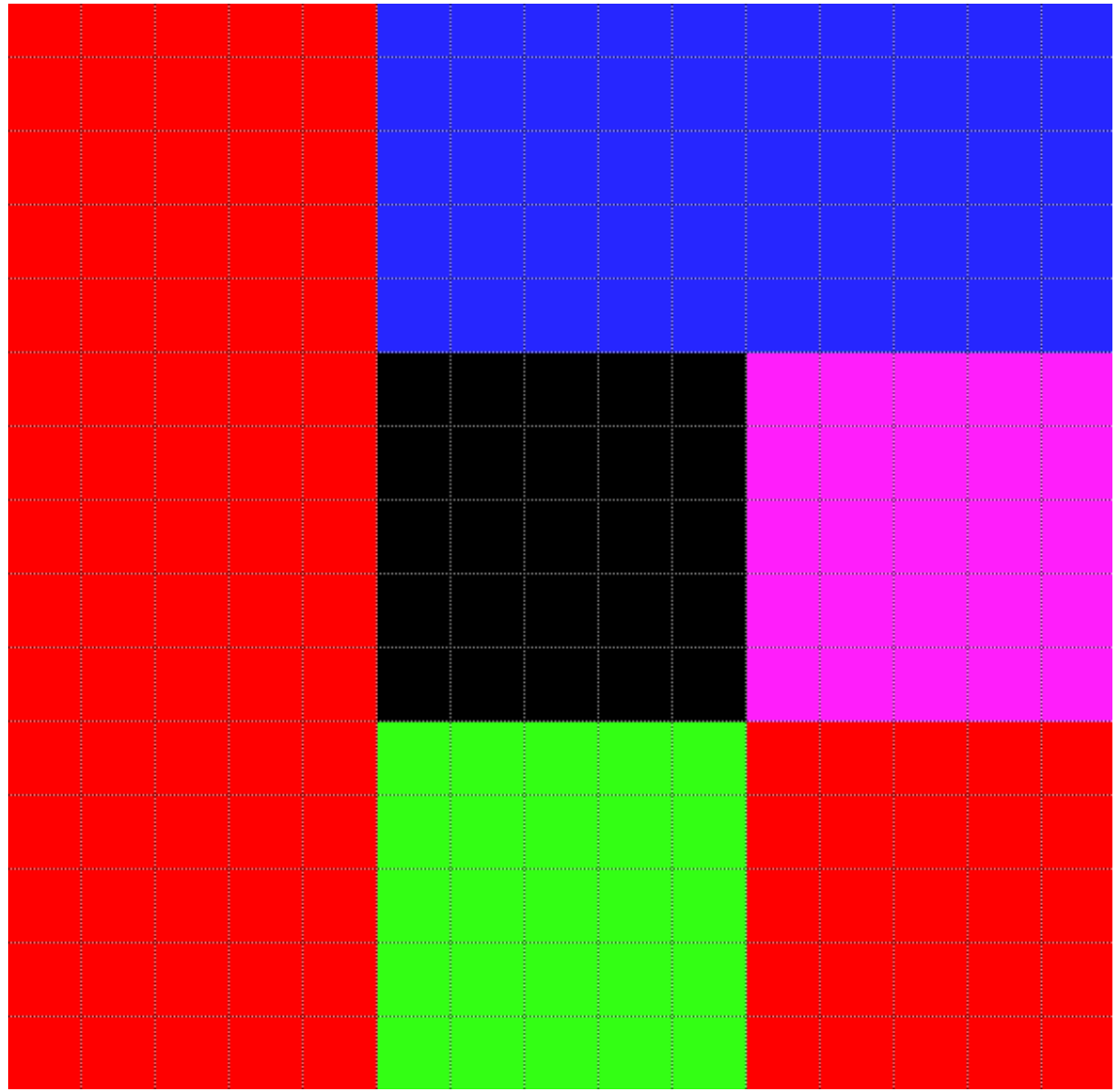
- We currently have 228 images of damaged walls captured in our campus areas
- Each image has 1 or more classes (cracked wall, damp wall, peeled paint, no defect) that we need to classify in our ML model
- These images also has various noise i.e. things which are not a wall
- So generating dataset first from these noisy images and then labelling them will take days of manual work
- In this presentation, we propose an easy way of generating dataset with least amount of manual work

Example

- Let us take an example image of one of our captured images

Divide the images into classes

- The image on the previous slide can be divided into classes as follows
  - Red = Not a wall
  - Blue = No defect
  - Black = Fungus (Damp wall)
  - Pink = Peeled paint
  - Green = Algae (Damp wall)
- There is no crack in this image
- We will consider fungus and algae in the same class as “Damp wall”

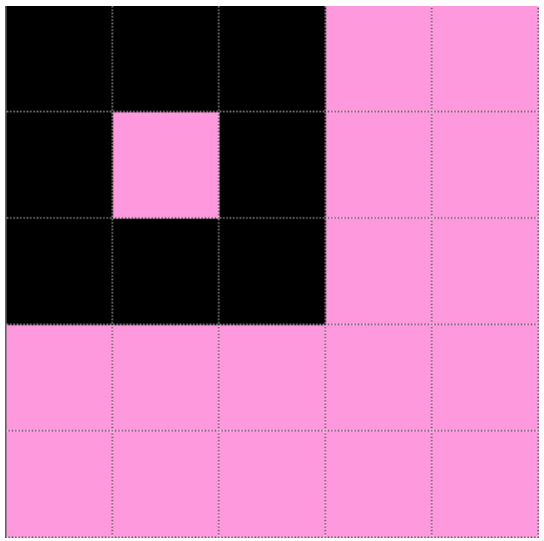
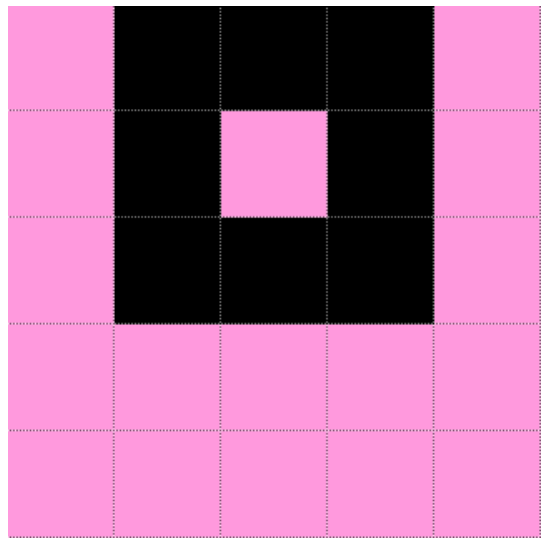
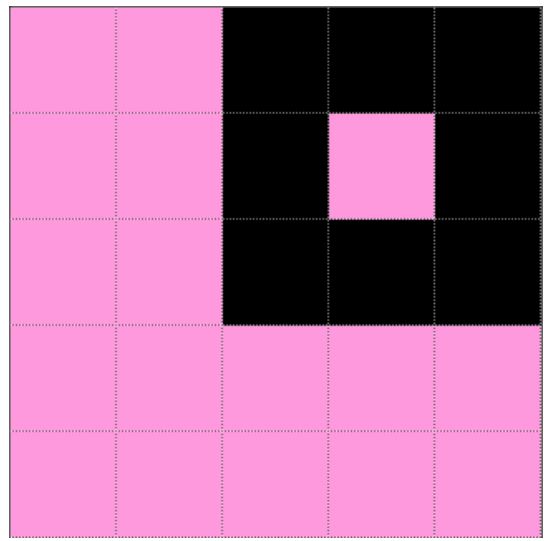
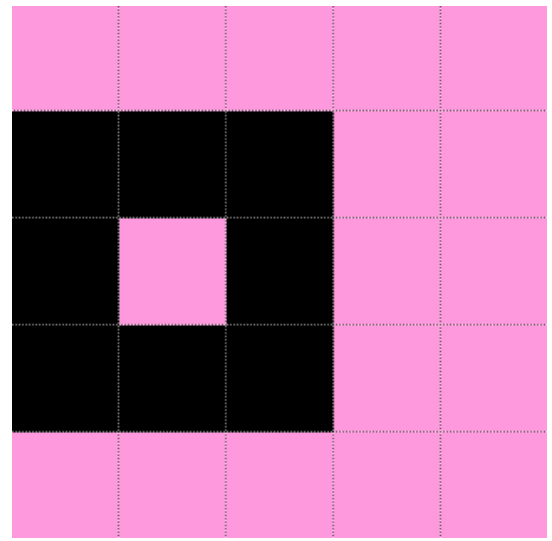


How to divide the image into classes

- The division of image is done manually
- But as the divided image would be as pure as possible (sub image comprising of only one class), the amount of noise would be reduced
- This cropped image would be placed in a folder of a specific class
- Currently we have taken 20 such sub images each of dimensions (300*300) for each class (total 80)
- The above 20 sub images are taken from different images to increase the variation in the dataset

Process each divided class

- For each sub image of 4 classes, we would generate all possible (256*256) window images
- These generated window images would be a part of our final dataset
- We will be using sliding window mechanism to generate these dataset (first few iterations illustrated in next slide)
- We are not applying any synthetic transformations such as rotation, scaling, etc.

Benefit and results

- For each (300*300) image, we are getting (300-256+1)^2 = 2025 (256*256) images each labelled with same class as the input (300*300) image
- Thus for 80 such labelled (300*300) images, we are obtaining (2025*80) = 1,62,000 labelled (256*256) images
- We can increase the number of samples (currently 80) to increase the volume of the dataset and reduce the cropping window size (currently 300*300) to increase the variation

Benefits and results (cont.)

- The sliding window mechanism generates images that are a column/row of pixel different, which although would look similar to humans but will be different to ML model
- We are working on building ML model to be trained from this dataset
- We are hoping to get good accuracy from this method