1. Dataset Description

* Authors:
* Authors: O. Ulucan, D. Karakaya, M. Turkan
* Department of Electrical and Electronics Engineering, Izmir University of Economics, Izmir, Turkey
* Corresponding author: M. Turkan
* Contact Information: [mehmet.turkan@ieu.edu.tr](mailto:mehmet.turkan@ieu.edu.tr)
* Data Gathering Equipment and Data Augmentation
* Images were collected via 2 different cameras, Kodak Easyshare Z650 and Samsung ST60. Therefore, the resolutions of the images are 2832 x 2128, 1024 x 768, respectively.
* Then the dataset was resized to 590 x 445 by preserving the aspect ratio. After resizing the images, all images in the dataset were augmented (by flipping and rotating).
* At the end of the augmentation process, the number of total images for each class became 2000; 1000 for the RGB fish images and 1000 for their pair-wise ground truth labels.
* Description of the data in this data set
* The dataset contains 9 different seafood types. For each class, there are 1000 augmented images and their pair-wise augmented ground truths.
* Each class can be found in the "Fish\_Dataset" file with their ground truth labels. All images for each class are ordered from "00000.png" to "01000.png"

1. Future improvement (for writing in report):

* Add background removal to aid in image classifications (the images in this dataset are already on a mostly blue background).
* Because of the lack of background removal, images with (complex) background different from the one in the dataset can cause the models to produce the wrong prediction
* For background removal, we can use a python library like **OpenCV** or **rembg** to remove the background in the image preprocessing step.
* However, background removal can be computationally expensive. We can consider removing the background in advance and saving them to disk if the dataset is large. The training and testing step then is like what we did.
* For making prediction with new image, we will first remove the background in the image preprocessing step (remove the background before we resize and normalize the images) then fit it to the model.
* Expand the dataset to contain more images:
* Due to the way images of the dataset were collected, the dataset is not difficult to get a good accuracy score.
* The sample size for each label is also quite small and not diverse, (for example the shrimp label: only one species of shrimp while having several species of fishes)