Birzeit University
Department of Electrical & Computer Engineering
ENCS5343 Computer Vision
Assignment#2

Content-Based Image Retrieval (CBIR) System Using Color Features and other Techniques

Student: Rama Abdlrahman

Id : 1191344

Instructur : Dr. Aziz Qaroush

January-2024

Contents

1. Introduction

1.1 Overview of CBIR

Content-Based Image Retrieval (CBIR) systems are essential in the field of digital image processing and multimedia databases. CBIR refers to retrieving relevant images from large data sets based on the content of the images themselves, such as colors, shapes, textures, or any other information that can be derived from the image itself. This technology has a wide range of applications, including digital libraries, crime prevention, medical imaging, and even in the entertainment industry for visual effects and graphics.

1.2 Theoretical Background

- ### Color Histograms

A color histogram represents the distribution of colors in an image. It is one of the simplest yet most effective tools for image retrieval. In a CBIR system, color histograms serve as a feature vector to represent the color distribution of an image. The histogram is a count of the number of pixels in an image showing each distinct color. In more technical terms, it represents the frequency of the occurrence of various color bins (ranges of color values) in an image. These histograms are used to compare and retrieve images based on color similarity.

- ### Color Moments

Color moments are statistical measures that summarize the color distribution of an image. The most common moments used are the mean (which gives the average color), standard deviation (which measures the color variability), and skewness (which describes the asymmetry of the color distribution in an image). These moments can be calculated for each color channel and combined to form a feature vector representing the image in a CBIR system.

- ### CNN feature extraction

The VGG16 model is imported from the Keras library and configured to exclude its top classification layer. This allows the model to serve as a powerful feature extractor for images. The extracted features are then used to measure the cosine similarity between query images and the entire dataset, facilitating content-based image retrieval. The system evaluates its performance through metrics such as precision, recall, and F1-score, providing

insights into the effectiveness of the retrieval process. The CNN's ability to capture high-level image features enables the system to discern visual similarities among images in the dataset.

2. System Implementation

2.1 CBIR System Architecture

The architecture of a CBIR system typically consists of several key components:

1. **Image Database:** A collection of images which the system will search to retrieve relevant images.

2. **Feature Extraction Module:** This module processes each image in the database to extract meaningful features such as color histograms or color moments.
3. **Indexing and Storage Module:** Extracted features are indexed and stored for efficient retrieval.
4. **Query tests:** test query images or define features for retrieval. The same feature extraction process is applied to the query image.

5. **Similarity Measurement and Ranking Module:** Computes the similarity between the query features and database image features using distance measures like Euclidean distance. It then ranks the images based on similarity scores.

2.2 Programming Tools

The CBIR system can be implemented using various programming languages and libraries. Python is a popular choice due to its extensive libraries for image processing (such as OpenCV) and data handling (like NumPy and Pandas). In this work I have used the mentioned tools and libraries.

3. Experimental Setup and Results

3.1 Tasks 1 and Task 2: CBIR System Using Color Features (Experiment with Color Histograms)

In this section I will describe the experiments conducted with color histograms using 120, 180, and 360 bins. Including methodologies for calculating precision, recall, F1 score, and computational time. I will also explain the construction of the Receiver Operating Characteristic (ROC) curve and the calculation of the Area Under the Curve (AUC).

Note: In all tests below I am showing small samples of results (images and log output), for the full results you can go to the files or folders mentioned in each part.

- **Testing for 10 random queries:**

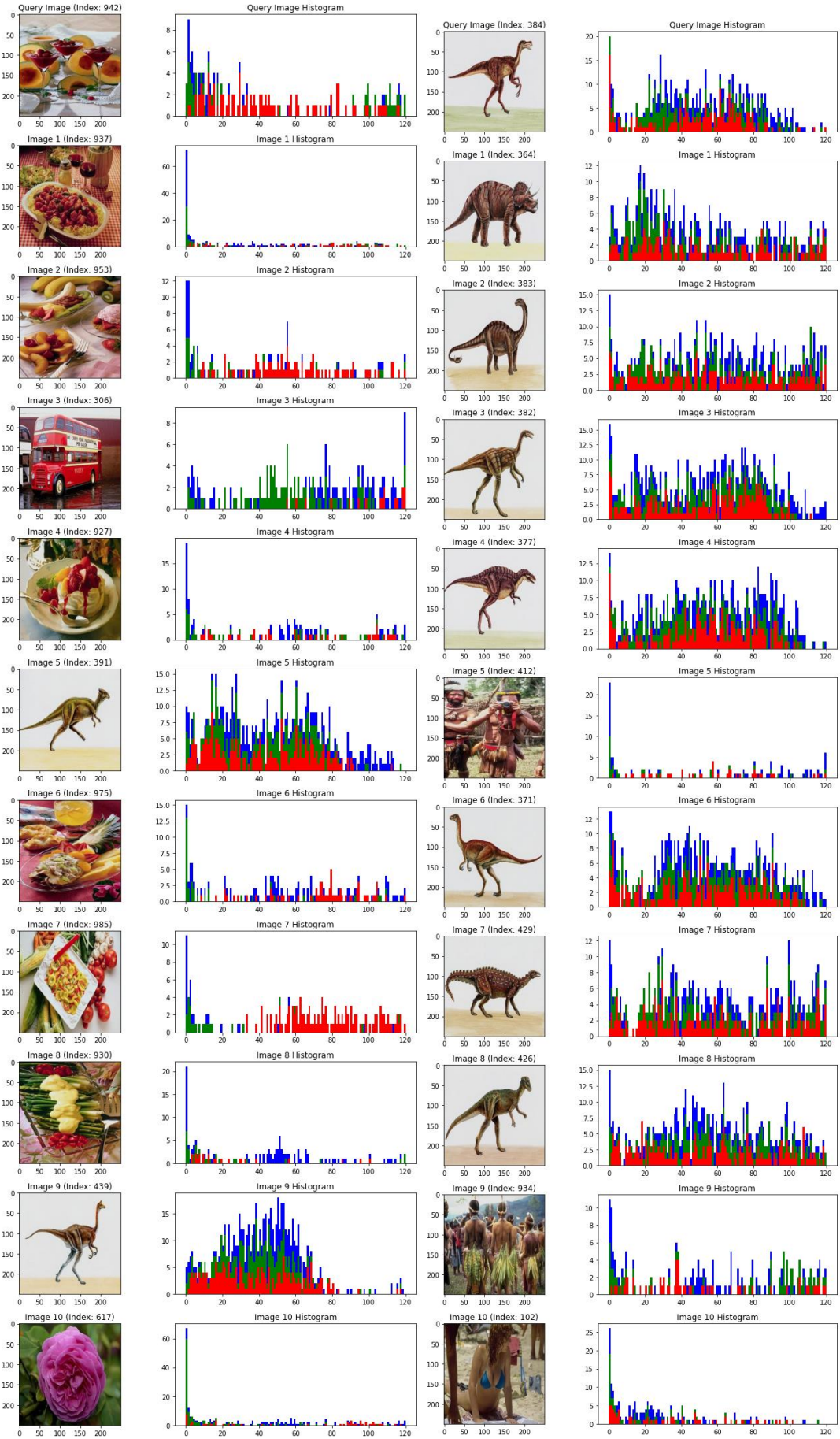
Here 10 random queries where chosen, and same threshold values applied on them. Some results will be shown for some tests, after showing the results, under each test results, we will have a short explanation.

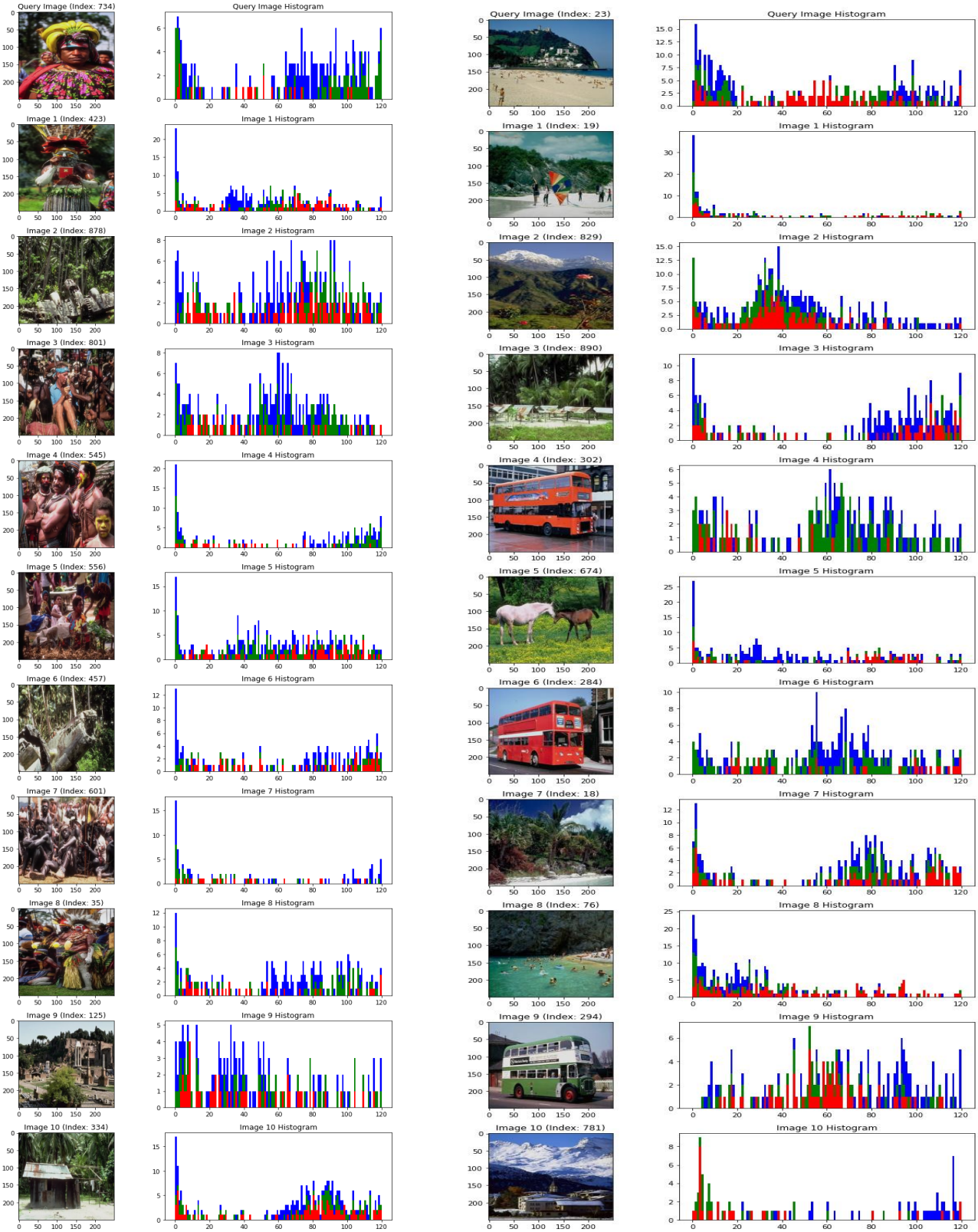
Results Using these thresholds for the 10 queries: (Test 1)

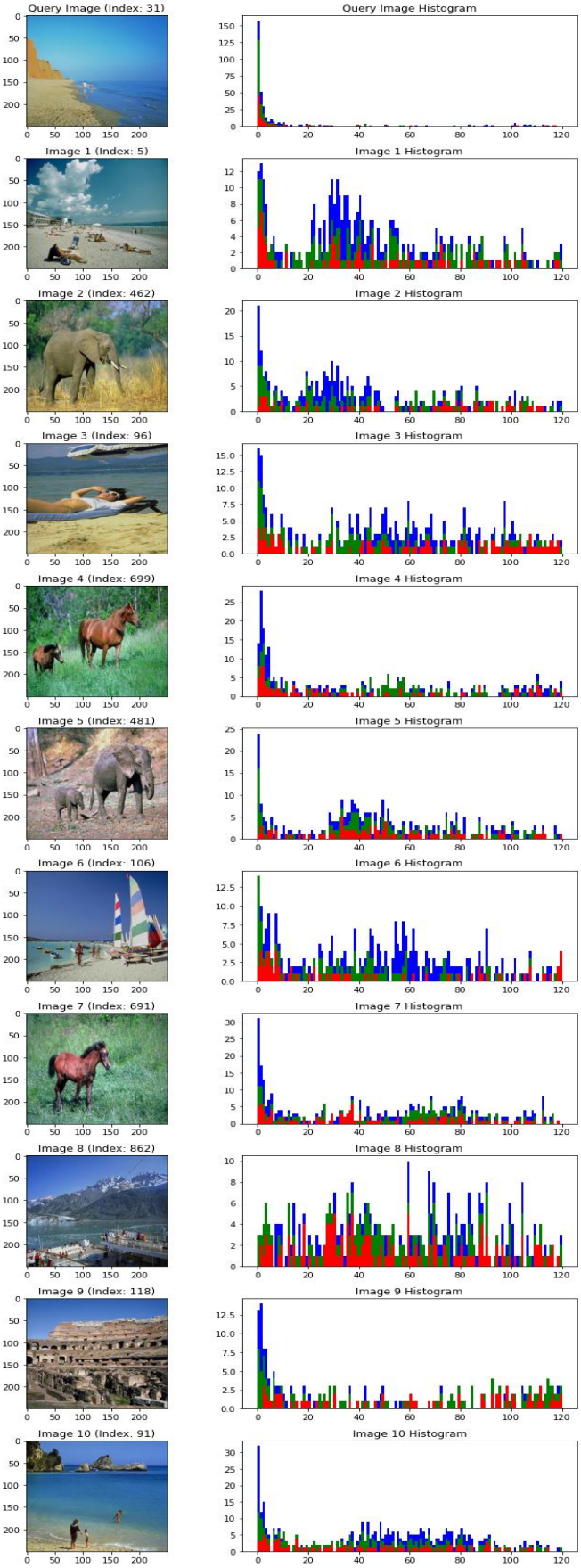
quantiles = [0.0111, 0.03,0.05,0.059,0.07, 0.09,0.099,.119,0.15,0.188, 0.19,0.2,0.246,0.3,0.333,0.36]

Results for 120 bins:

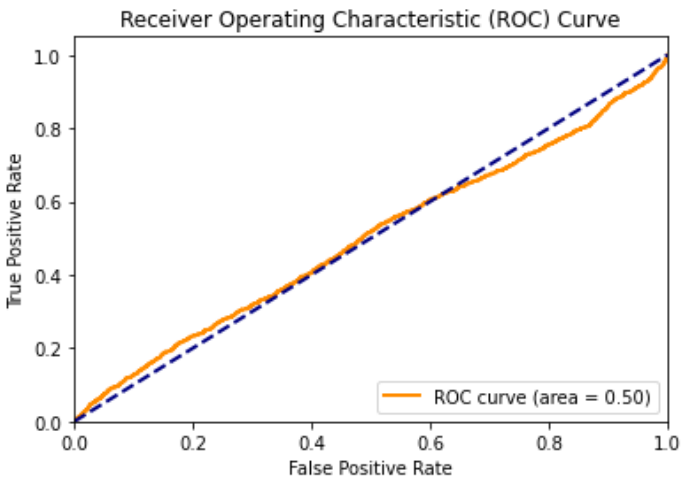
Here I am showing samples, rest of images for this test could be found in folder RandomQuiriesTest1 .


Query Image (Index: 942)
Query Image Histogram
Query Image (Index: 384)
Query Image Histogram
Image 1 (Index: 937)
Image 1 Histogram
Image 1 (Index: 364)
Image 1 Histogram
Image 2 (Index: 953)
Image 2 Histogram
Image 2 (Index: 383)
Image 2 Histogram
Image 3 (Index: 306)
Image 3 Histogram
Image 3 (Index: 382)
Image 3 Histogram
Image 4 (Index: 927)
Image 4 Histogram
Image 4 (Index: 377)
Image 4 Histogram
Image 5 (Index: 391)
Image 5 Histogram
Image 5 (Index: 412)
Image 5 Histogram
Image 6 (Index: 975)
Image 6 Histogram
Image 6 (Index: 371)
Image 6 Histogram
Image 7 (Index: 985)
Image 7 Histogram
Image 7 (Index: 429)
Image 7 Histogram
Image 8 (Index: 930)
Image 8 Histogram
Image 8 (Index: 426)
Image 8 Histogram
Image 9 (Index: 439)
Image 9 Histogram
Image 9 (Index: 934)
Image 9 Histogram
Image 10 (Index: 617)
Image 10 Histogram
Image 10 (Index: 102)
Image 10 Histogram


Query Image (Index: 734)
Query Image Histogram
Image 1 (Index: 423)
Image 1 Histogram
Image 2 (Index: 878)
Image 2 Histogram
Image 3 (Index: 801)
Image 3 Histogram
Image 4 (Index: 545)
Image 4 Histogram
Image 5 (Index: 556)
Image 5 Histogram
Image 6 (Index: 457)
Image 6 Histogram
Image 7 (Index: 601)
Image 7 Histogram
Image 8 (Index: 35)
Image 8 Histogram
Image 9 (Index: 125)
Image 9 Histogram
Image 10 (Index: 334)
Image 10 Histogram
Query Image (Index: 23)
Query Image Histogram
Image 1 (Index: 19)
Image 1 Histogram
Image 2 (Index: 829)
Image 2 Histogram
Image 3 (Index: 890)
Image 3 Histogram
Image 4 (Index: 302)
Image 4 Histogram
Image 5 (Index: 674)
Image 5 Histogram
Image 6 (Index: 284)
Image 6 Histogram
Image 7 (Index: 18)
Image 7 Histogram
Image 8 (Index: 76)
Image 8 Histogram
Image 9 (Index: 294)
Image 9 Histogram
Image 10 (Index: 781)
Image 10 Histogram

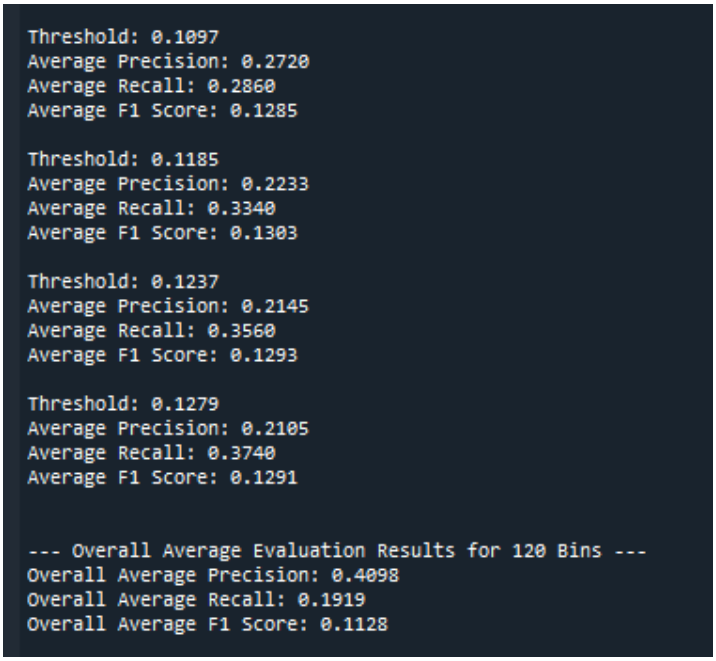

Query Image (Index: 31)
Query Image Histogram
Image 1 (Index: 5)
Image 1 Histogram
Image 2 (Index: 462)
Image 2 Histogram
Image 3 (Index: 96)
Image 3 Histogram
Image 4 (Index: 699)
Image 4 Histogram
Image 5 (Index: 481)
Image 5 Histogram
Image 6 (Index: 106)
Image 6 Histogram
Image 7 (Index: 691)
Image 7 Histogram
Image 8 (Index: 862)
Image 8 Histogram
Image 9 (Index: 118)
Image 9 Histogram
Image 10 (Index: 91)
Image 10 Histogram

I have used 10 queries but I can't show them all here since it will take too much space, you can check the folder I mentioned above.



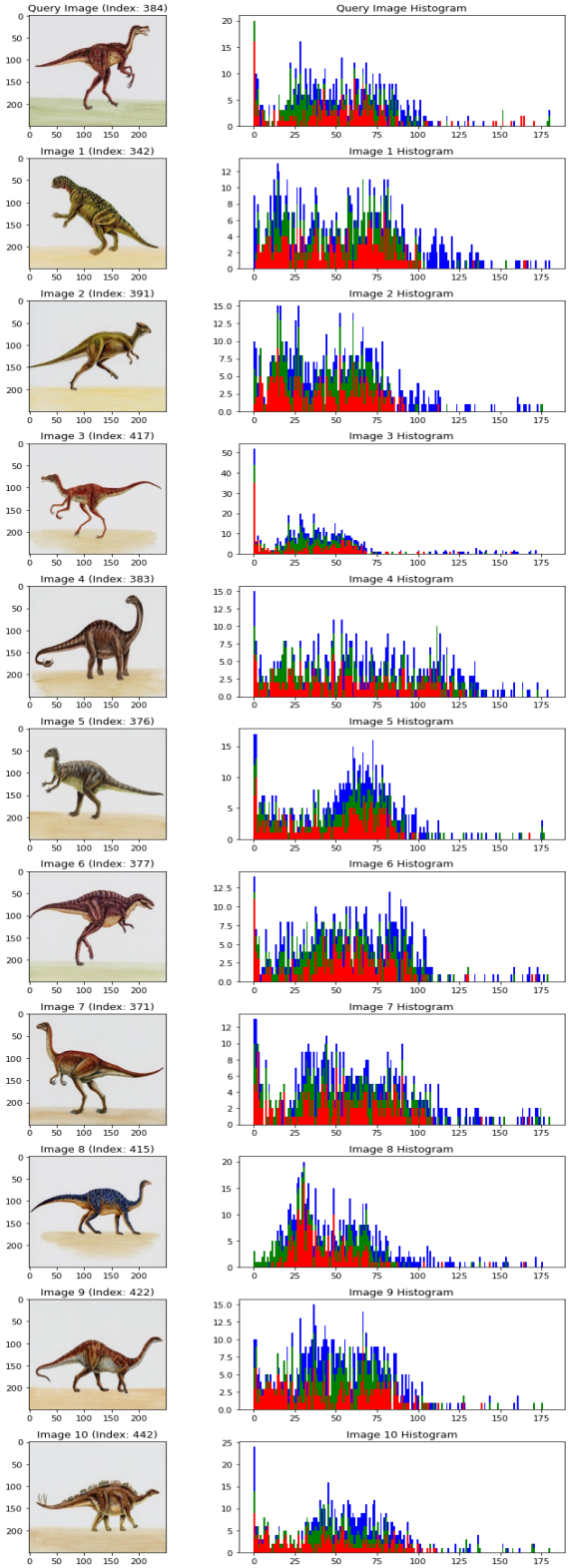
Some results are shown here, the full results including precision, recall, f1score, execution time per query at each threshold for each experiment(bins numbers), and the average values of these measures per query could be found in file **test1.log**

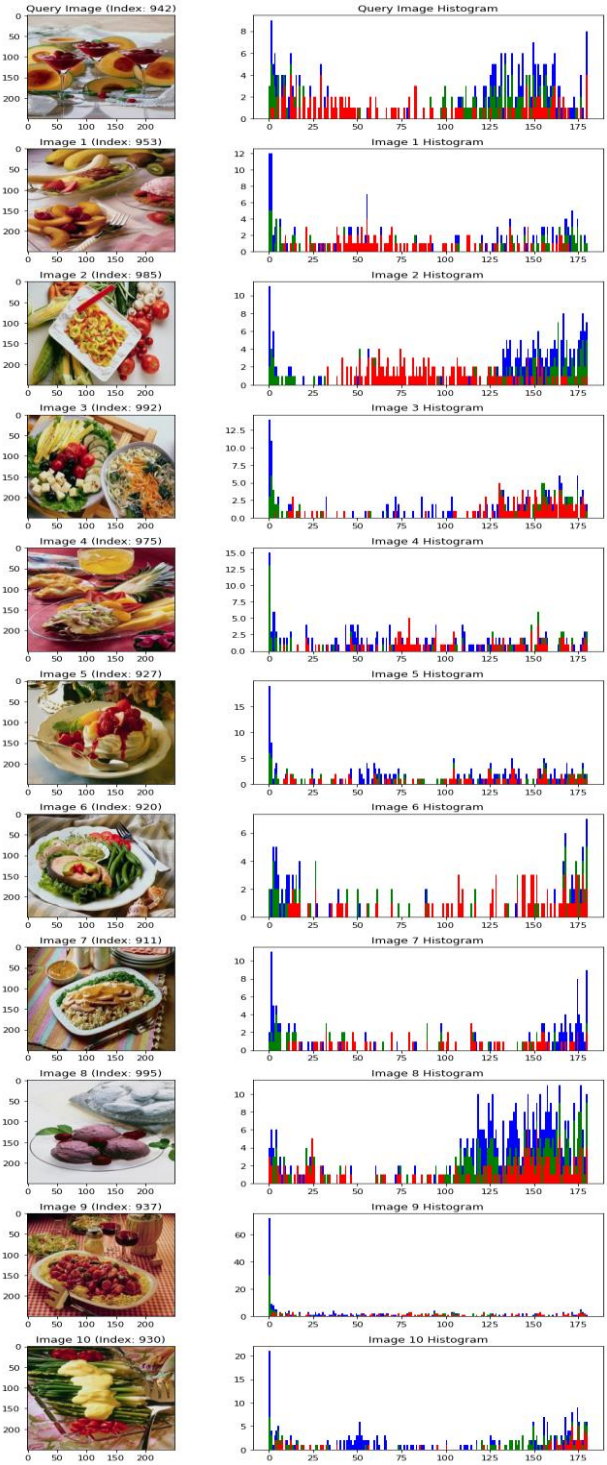
```
Time taken for 120 bins: 79.9514 seconds

--- Average Evaluation Results for 120 Bins for All Queries at Each Threshold
Threshold: 0.0524
Average Precision: 0.6475
Average Recall: 0.0340
Average F1 Score: 0.0535

Threshold: 0.0632
Average Precision: 0.6068
Average Recall: 0.0670
Average F1 Score: 0.0862

Threshold: 0.0704
Average Precision: 0.5537
Average Recall: 0.0860
Average F1 Score: 0.0946

Threshold: 0.0731
Average Precision: 0.5545
Average Recall: 0.0940
Average F1 Score: 0.0982

Threshold: 0.0762
Average Precision: 0.5301
Average Recall: 0.1030
Average F1 Score: 0.1009

Threshold: 0.0811
Average Precision: 0.5080
Average Recall: 0.1220
Average F1 Score: 0.1054

Threshold: 0.0831
Average Precision: 0.5006
Average Recall: 0.1330
Average F1 Score: 0.1101

Threshold: 0.0873
Average Precision: 0.4637
Average Recall: 0.1660
Average F1 Score: 0.1245

Threshold: 0.0933
Average Precision: 0.3660
Average Recall: 0.1990
Average F1 Score: 0.1270

Threshold: 0.1001
Average Precision: 0.3244
Average Recall: 0.2350
Average F1 Score: 0.1293

Threshold: 0.1004
Average Precision: 0.2901
Average Recall: 0.2370
Average F1 Score: 0.1291

Threshold: 0.1021
Average Precision: 0.2914
Average Recall: 0.2450
Average F1 Score: 0.1288
```

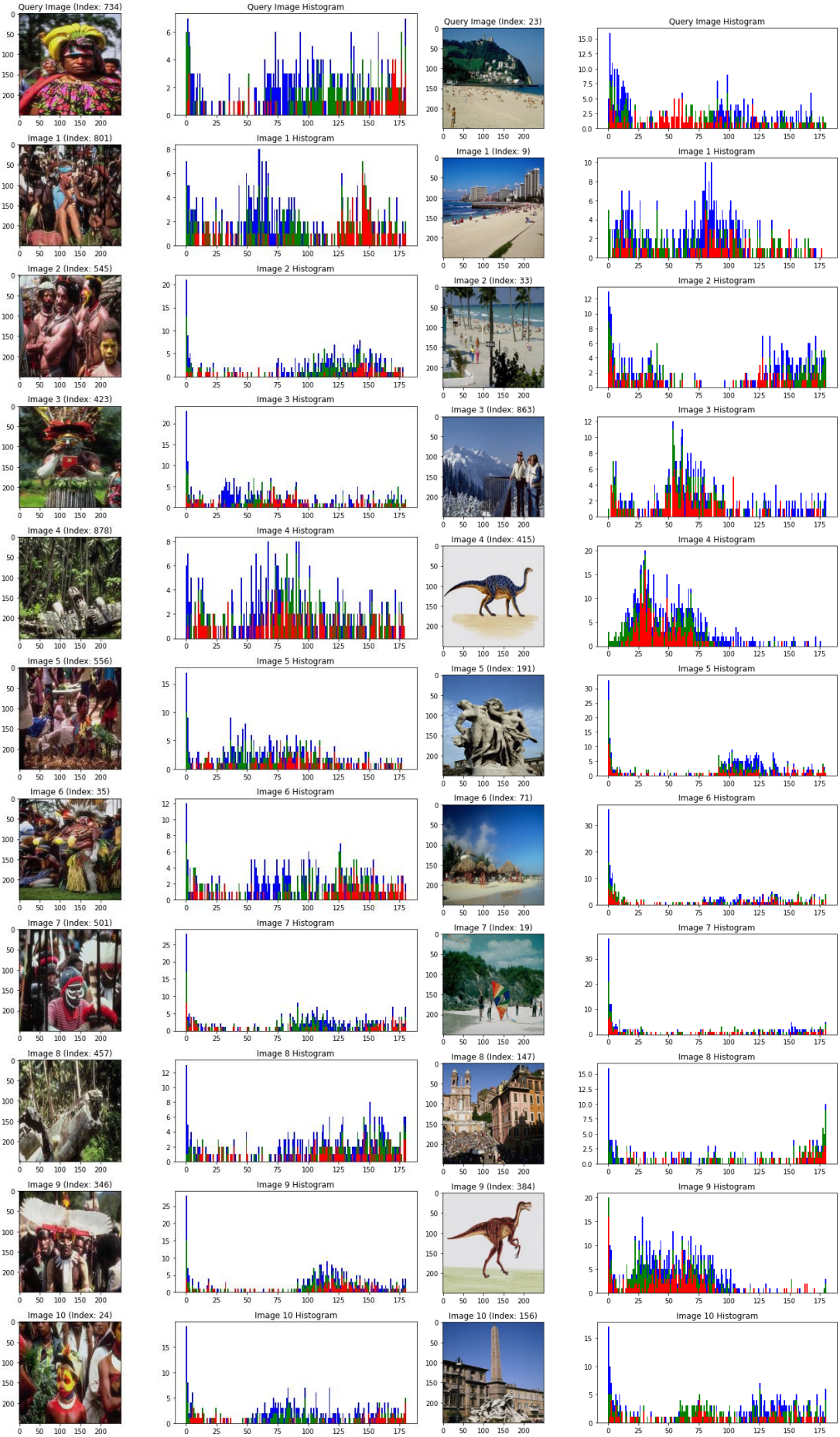


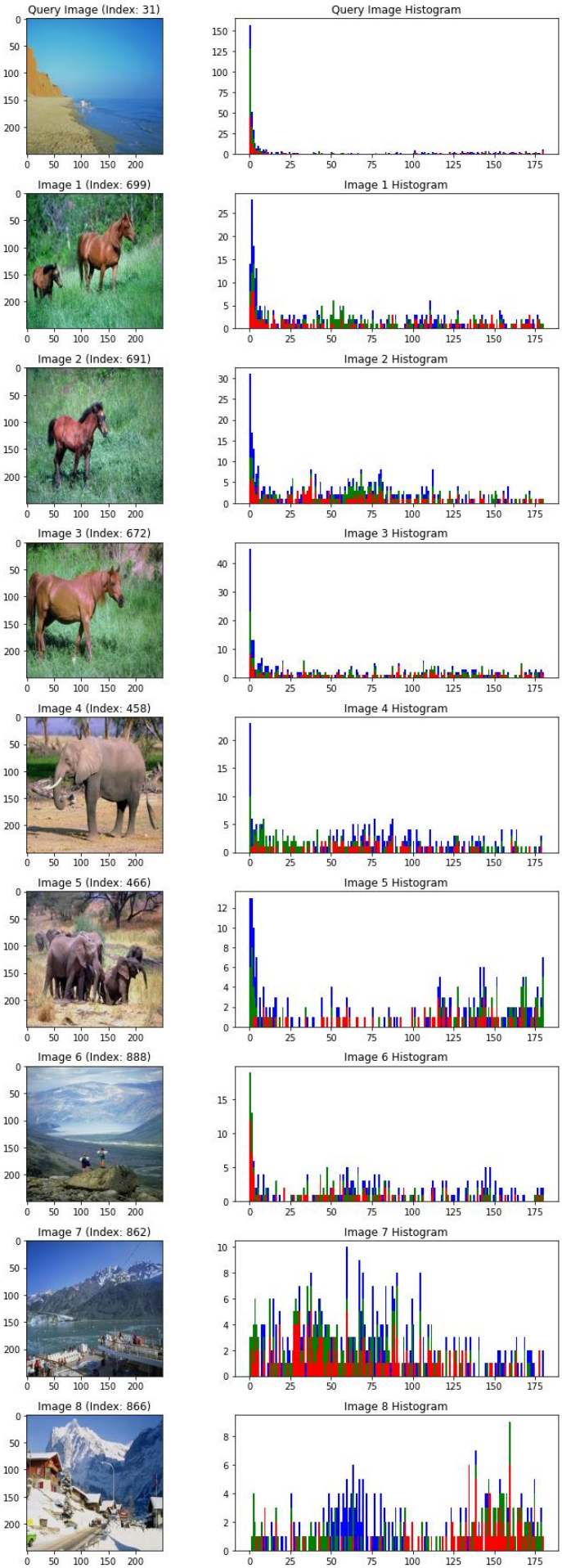
Results for 120 bins:

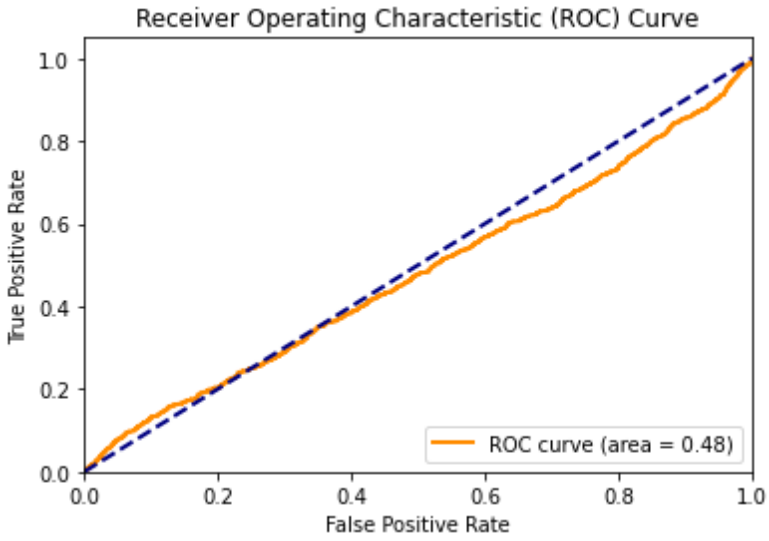
Here I am showing samples, rest of images for this test could be found in folder RandomQuiriesTest1.


Query Image (Index: 384)
Query Image Histogram
Image 1 (Index: 342)
Image 1 Histogram
Image 2 (Index: 391)
Image 2 Histogram
Image 3 (Index: 417)
Image 3 Histogram
Image 4 (Index: 383)
Image 4 Histogram
Image 5 (Index: 376)
Image 5 Histogram
Image 6 (Index: 377)
Image 6 Histogram
Image 7 (Index: 371)
Image 7 Histogram
Image 8 (Index: 415)
Image 8 Histogram
Image 9 (Index: 422)
Image 9 Histogram
Image 10 (Index: 442)
Image 10 Histogram


Query Image (Index: 942)
Query Image Histogram
Image 1 (Index: 953)
Image 1 Histogram
Image 2 (Index: 985)
Image 2 Histogram
Image 3 (Index: 992)
Image 3 Histogram
Image 4 (Index: 975)
Image 4 Histogram
Image 5 (Index: 927)
Image 5 Histogram
Image 6 (Index: 920)
Image 6 Histogram
Image 7 (Index: 911)
Image 7 Histogram
Image 8 (Index: 995)
Image 8 Histogram
Image 9 (Index: 937)
Image 9 Histogram
Image 10 (Index: 930)
Image 10 Histogram


Query Image (Index: 734)
Query Image Histogram
Query Image (Index: 23)
Query Image Histogram
Image 1 (Index: 801)
Image 1 Histogram
Image 1 (Index: 9)
Image 1 Histogram
Image 2 (Index: 545)
Image 2 Histogram
Image 2 (Index: 33)
Image 2 Histogram
Image 3 (Index: 423)
Image 3 Histogram
Image 3 (Index: 863)
Image 3 Histogram
Image 4 (Index: 878)
Image 4 Histogram
Image 4 (Index: 415)
Image 4 Histogram
Image 5 (Index: 556)
Image 5 Histogram
Image 5 (Index: 191)
Image 5 Histogram
Image 6 (Index: 35)
Image 6 Histogram
Image 6 (Index: 71)
Image 6 Histogram
Image 7 (Index: 501)
Image 7 Histogram
Image 7 (Index: 19)
Image 7 Histogram
Image 8 (Index: 457)
Image 8 Histogram
Image 8 (Index: 147)
Image 8 Histogram
Image 9 (Index: 346)
Image 9 Histogram
Image 9 (Index: 384)
Image 9 Histogram
Image 10 (Index: 24)
Image 10 Histogram
Image 10 (Index: 156)
Image 10 Histogram


Query Image (Index: 31)
Query Image Histogram
Image 1 (Index: 699)
Image 1 Histogram
Image 2 (Index: 691)
Image 2 Histogram
Image 3 (Index: 672)
Image 3 Histogram
Image 4 (Index: 458)
Image 4 Histogram
Image 5 (Index: 466)
Image 5 Histogram
Image 6 (Index: 888)
Image 6 Histogram
Image 7 (Index: 862)
Image 7 Histogram
Image 8 (Index: 866)
Image 8 Histogram

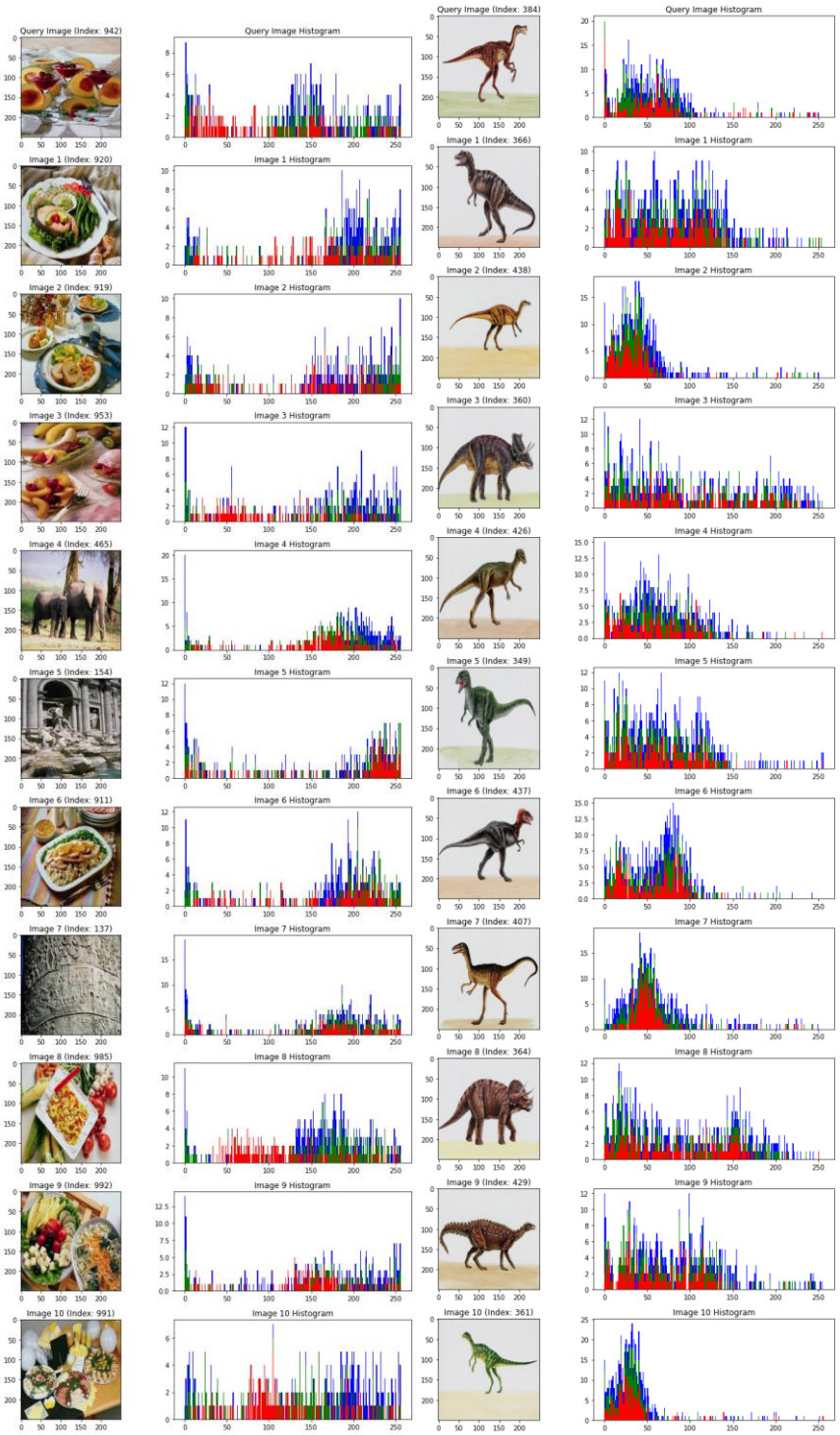


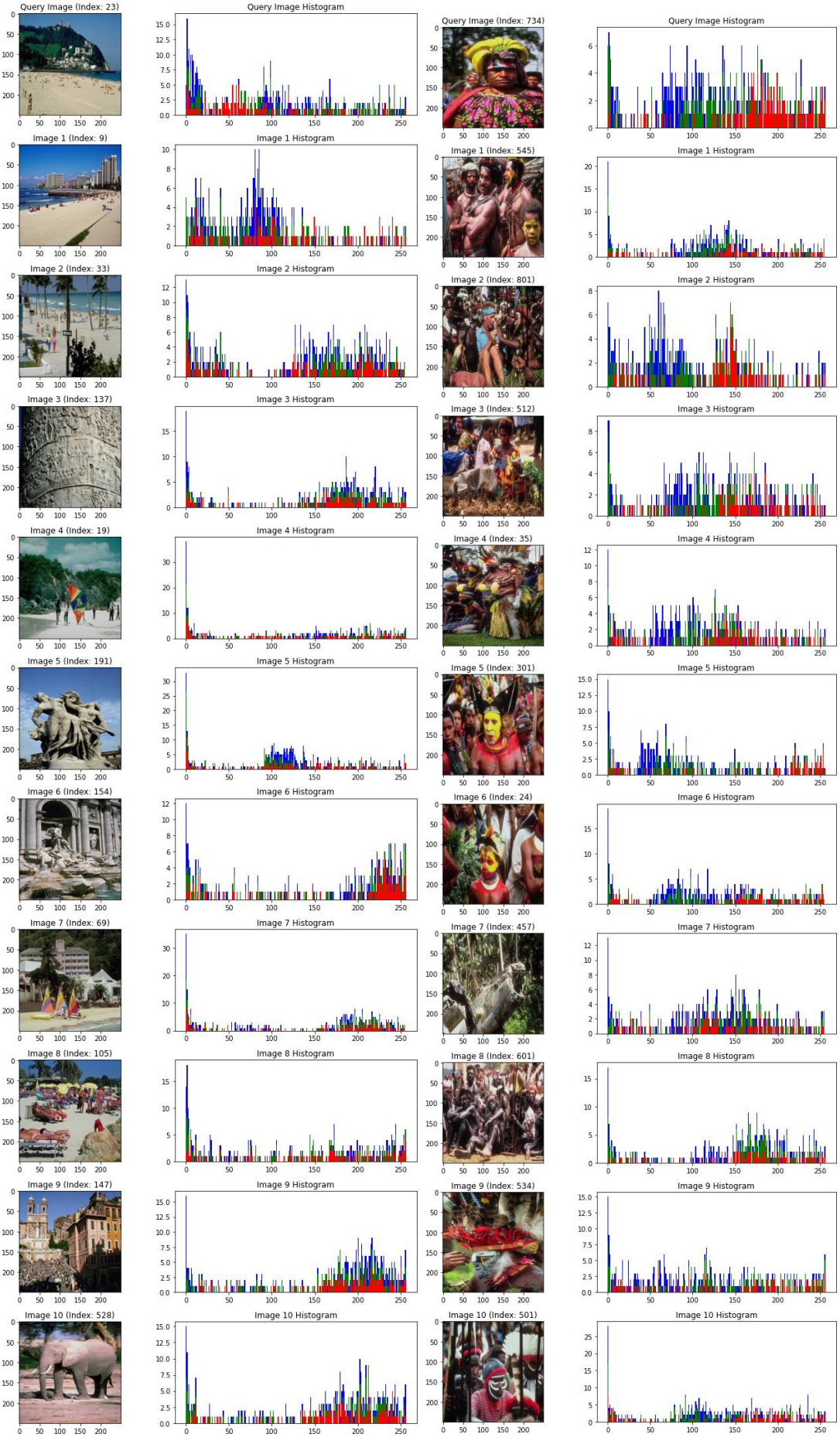

```
Time taken for 180 bins: 100.5498 seconds

--- Average Evaluation Results for 180 Bins for All Queries at Each Threshold -
Threshold: 0.0638
Average Precision: 0.6806
Average Recall: 0.0270
Average F1 Score: 0.0429

Threshold: 0.0765
Average Precision: 0.6280
Average Recall: 0.0570
Average F1 Score: 0.0800

Threshold: 0.0845
Average Precision: 0.5892
Average Recall: 0.0940
Average F1 Score: 0.1118

Threshold: 0.0874
Average Precision: 0.5159
Average Recall: 0.1050
Average F1 Score: 0.1156

Threshold: 0.0906
Average Precision: 0.5145
Average Recall: 0.1200
Average F1 Score: 0.1232

Threshold: 0.0957
Average Precision: 0.4703
Average Recall: 0.1340
Average F1 Score: 0.1211

Threshold: 0.0978
Average Precision: 0.4634
Average Recall: 0.1420
Average F1 Score: 0.1205

Threshold: 0.1023
Average Precision: 0.3704
Average Recall: 0.1570
Average F1 Score: 0.1186

Threshold: 0.1084
Average Precision: 0.3632
Average Recall: 0.1920
Average F1 Score: 0.1293

Threshold: 0.1152
Average Precision: 0.3515
Average Recall: 0.2310
Average F1 Score: 0.1378

Threshold: 0.1156
Average Precision: 0.3499
Average Recall: 0.2330
Average F1 Score: 0.1378

Threshold: 0.1173
Average Precision: 0.3455
Average Recall: 0.2450
Average F1 Score: 0.1391
```

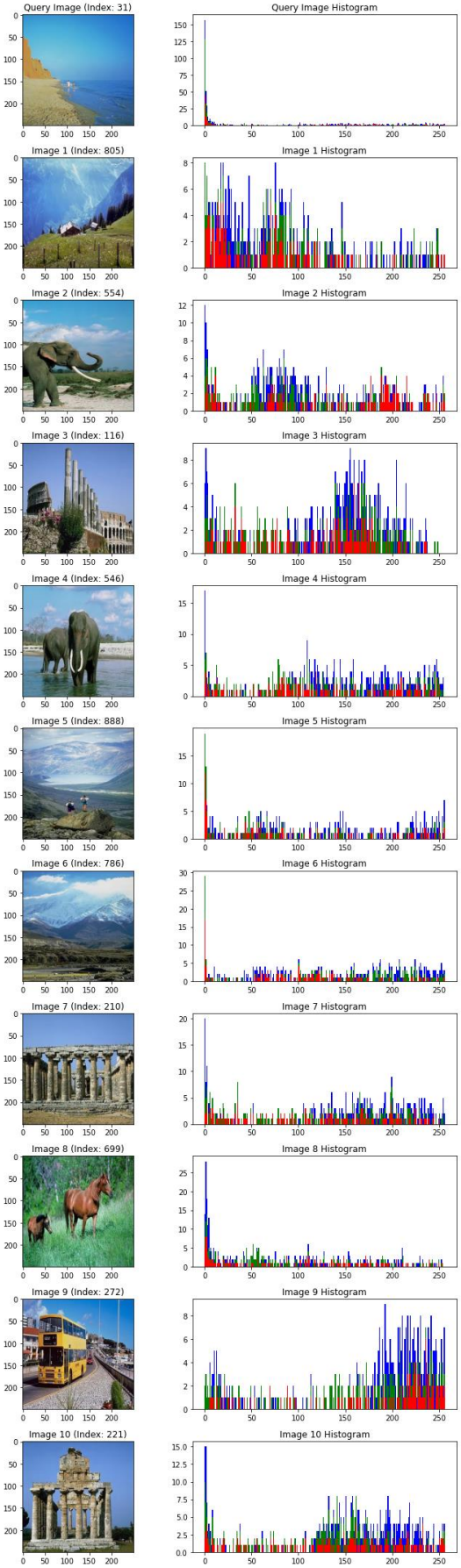
```
Threshold: 0.1249
Average Precision: 0.2643
Average Recall: 0.2880
Average F1 Score: 0.1445

Threshold: 0.1332
Average Precision: 0.2457
Average Recall: 0.3340
Average F1 Score: 0.1473

Threshold: 0.1382
Average Precision: 0.2311
Average Recall: 0.3650
Average F1 Score: 0.1532

Threshold: 0.1424
Average Precision: 0.2161
Average Recall: 0.3810
Average F1 Score: 0.1516


--- Overall Average Evaluation Results for 180 Bins ---
Overall Average Precision: 0.4125
Overall Average Recall: 0.1941
Overall Average F1 Score: 0.1234
```

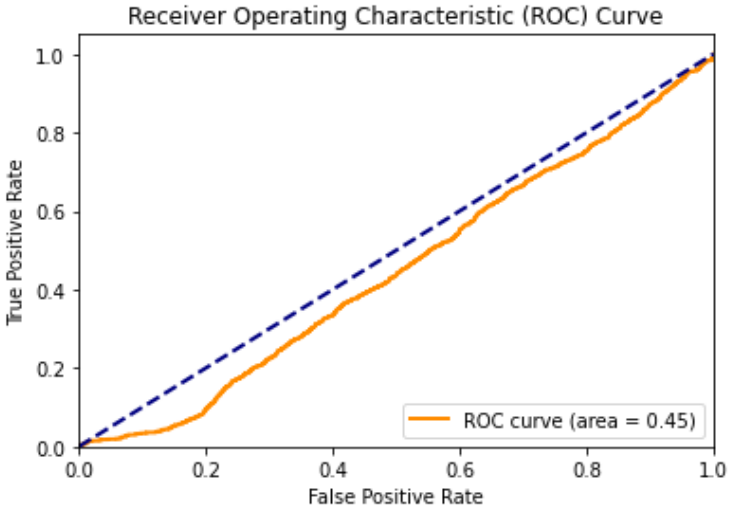
Results for 256 bins:

Here I am showing samples, rest of images for this test could be found in folder RandomQuiriesTest1 .




Query Image (Index: 23)
Query Image Histogram
Query Image (Index: 734)
Query Image Histogram
Image 1 (Index: 9)
Image 1 Histogram
Image 1 (Index: 545)
Image 1 Histogram
Image 2 (Index: 33)
Image 2 Histogram
Image 2 (Index: 801)
Image 2 Histogram
Image 3 (Index: 137)
Image 3 Histogram
Image 3 (Index: 512)
Image 3 Histogram
Image 4 (Index: 19)
Image 4 Histogram
Image 4 (Index: 35)
Image 4 Histogram
Image 5 (Index: 191)
Image 5 Histogram
Image 5 (Index: 301)
Image 5 Histogram
Image 6 (Index: 154)
Image 6 Histogram
Image 6 (Index: 24)
Image 6 Histogram
Image 7 (Index: 69)
Image 7 Histogram
Image 7 (Index: 457)
Image 7 Histogram
Image 8 (Index: 105)
Image 8 Histogram
Image 8 (Index: 601)
Image 8 Histogram
Image 9 (Index: 147)
Image 9 Histogram
Image 9 (Index: 534)
Image 9 Histogram
Image 10 (Index: 528)
Image 10 Histogram
Image 10 (Index: 501)
Image 10 Histogram


Query Image (Index: 31)
Query Image Histogram
Image 1 (Index: 805)
Image 1 Histogram
Image 2 (Index: 554)
Image 2 Histogram
Image 3 (Index: 116)
Image 3 Histogram
Image 4 (Index: 546)
Image 4 Histogram
Image 5 (Index: 888)
Image 5 Histogram
Image 6 (Index: 786)
Image 6 Histogram
Image 7 (Index: 210)
Image 7 Histogram
Image 8 (Index: 699)
Image 8 Histogram
Image 9 (Index: 272)
Image 9 Histogram
Image 10 (Index: 221)
Image 10 Histogram




```
Time taken for 256 bins: 130.9678 seconds

--- Average Evaluation Results for 256 Bins for All Queries at Each Threshold -
Threshold: 0.0527
Average Precision: 0.6735
Average Recall: 0.0250
Average F1 Score: 0.0380

Threshold: 0.0632
Average Precision: 0.6235
Average Recall: 0.0450
Average F1 Score: 0.0548

Threshold: 0.0698
Average Precision: 0.5601
Average Recall: 0.0790
Average F1 Score: 0.0837

Threshold: 0.0723
Average Precision: 0.5357
Average Recall: 0.0890
Average F1 Score: 0.0892

Threshold: 0.0750
Average Precision: 0.5197
Average Recall: 0.1030
Average F1 Score: 0.0920

Threshold: 0.0793
Average Precision: 0.5081
Average Recall: 0.1310
Average F1 Score: 0.1010

Threshold: 0.0811
Average Precision: 0.5197
Average Recall: 0.1460
Average F1 Score: 0.1054

Threshold: 0.0847
Average Precision: 0.5062
Average Recall: 0.1650
Average F1 Score: 0.1061

Threshold: 0.0897
Average Precision: 0.4862
Average Recall: 0.1960
Average F1 Score: 0.1106

Threshold: 0.0954
Average Precision: 0.4105
Average Recall: 0.2260
Average F1 Score: 0.1134

Threshold: 0.0956
Average Precision: 0.4106
Average Recall: 0.2280
Average F1 Score: 0.1138

Threshold: 0.0970
Average Precision: 0.3958
Average Recall: 0.2390
```

```
Threshold: 0.0970
Average Precision: 0.3958
Average Recall: 0.2390
Average F1 Score: 0.1158

Threshold: 0.1033
Average Precision: 0.3412
Average Recall: 0.2710
Average F1 Score: 0.1186

Threshold: 0.1106
Average Precision: 0.3004
Average Recall: 0.3010
Average F1 Score: 0.1217

Threshold: 0.1150
Average Precision: 0.2921
Average Recall: 0.3270
Average F1 Score: 0.1274

Threshold: 0.1187
Average Precision: 0.2715
Average Recall: 0.3470
Average F1 Score: 0.1340


--- Overall Average Evaluation Results for 256 Bins -
Overall Average Precision: 0.4597
Overall Average Recall: 0.1824
Overall Average F1 Score: 0.1016

Completed all experiments.
```

Discussing the results above:

- **About the ROC curve and AUC**

120 Bins: The ROC curve has an AUC of 0.50, which suggests that the classifier does no better than random chance in distinguishing between the two classes.

180 Bins: The AUC here is slightly less than before, at 0.48, which indicates a potential decrease in the classifier's performance as the number of bins increases.

256 Bins: The AUC is 0.45, which further suggests a decrease in performance with an increase in the number of bins.

- **About the measures**

As shown on this test, as the number of bins increases from 120 to 256, the overall average precision and recall seem to fluctuate. There isn't a clear trend showing improvement or decline with the increase in bins. However, the precision appears to increase about 0.01which is not worth to mention, and the overall F1 score appears to slightly decrease.

For 120 Bins the system has average precision and recall rates around 0.40 and an F1 score around 0.11. For 180 Bins precision and recall rates are similar to those with 120 bins, very little increase in precision, the F1 score slightly improves. For 256 Bins the average precision and recall do not improve significantly, very little increase in precision, and the F1 score decreases slightly compared to 180 bins. The time taken for computations increases with the number of bins, which is expected due to the increase in computational complexity.

The results suggest that increasing the number of histogram bins for this color-based retrieval system does not necessarily improve performance. It may slightly degrade the classifier's ability to distinguish between the positive and negative classes, as indicated by the ROC curves' AUC values. Additionally, it seems to increase the computational load without a corresponding increase in accuracy. The performance degrading with more bins has many possible reasons.

dataset issues could be one of the reasons for the poor performance. If the dataset lacks variability in colors between different classes, then a color-based retrieval system might struggle to differentiate between them. There might be significant overlap in color distributions across different classes, making it hard for the histogram method to distinguish between them effectively. Also, f the dataset is too small or does not have enough representative samples for each class, the system may not be able to learn the distinguishing features properly. Besides,

the choice of color space (RGB) could affect performance. Some color spaces might encapsulate the differences between classes better than it. Also, If the dataset is imbalanced, with some classes having significantly more samples than others, it can skew the training process and affect the system's ability to generalize. For some datasets, color may not be a relevant feature for class discrimination. For instance, in datasets where the shape, texture, or context is more important, color histograms might not provide useful information for classification.

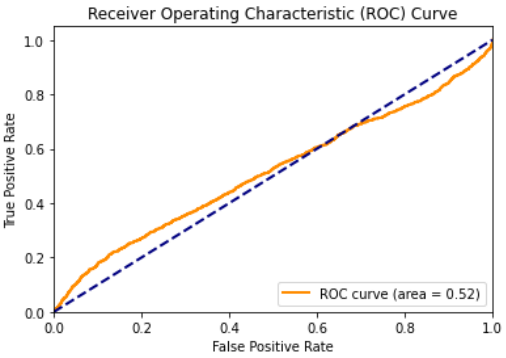
Results Using these thresholds for the 10 queries: (Test 2)

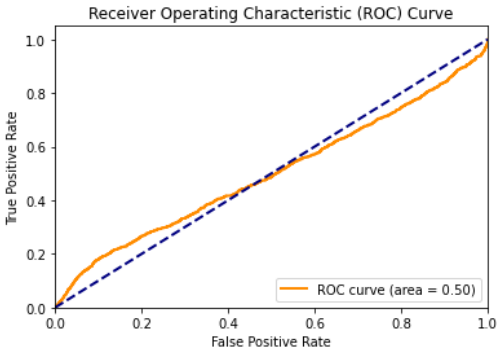
quantiles = [0.0111, 0.03,0.05,0.059,0.07, 0.09,0.099,.119,0.15,0.188, 0.19,0.2,0.246,0.3,0.333,0.36, 0.015, 0.02, 0.025, 0.035, 0.04, 0.045, 0.055, 0.06, 0.065, 0.075, 0.08, 0.085, 0.095, 0.1001, 0.105, 0.11, 0.12, 0.125, 0.13, 0.135, 0.14, 0.145, 0.155, 0.16, 0.165, 0.17, 0.175, 0.18, 0.185, 0.195]

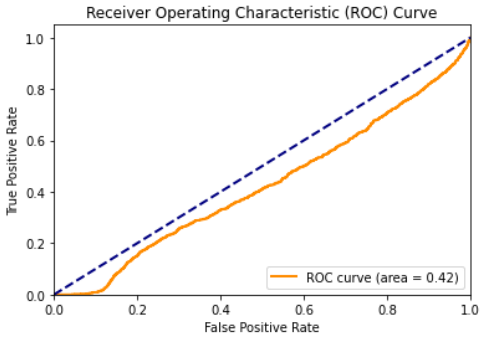
Got the following results:

All images could be found in RandomQuiriesTest3

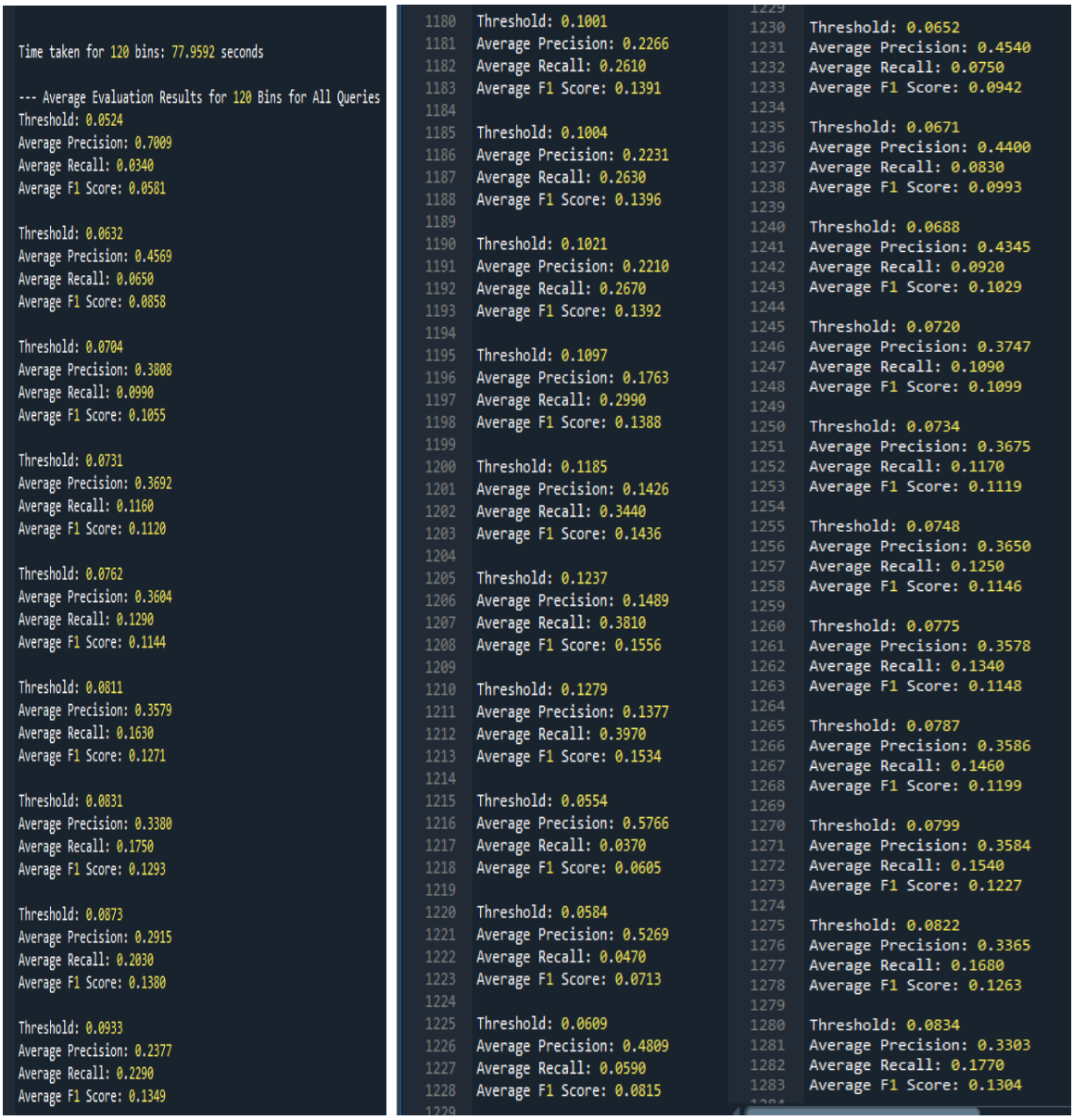
For 120 bins, 180 bins, and 256 bins, respectively.







The complete Consol output for this test could be found in Test3.log:



```
Time taken for 120 bins: 77.9592 seconds

--- Average Evaluation Results for 120 Bins for All Queries
Threshold: 0.0524
Average Precision: 0.7009
Average Recall: 0.0340
Average F1 Score: 0.0581

Threshold: 0.0632
Average Precision: 0.4569
Average Recall: 0.0650
Average F1 Score: 0.0858

Threshold: 0.0704
Average Precision: 0.3808
Average Recall: 0.0990
Average F1 Score: 0.1055

Threshold: 0.0731
Average Precision: 0.3692
Average Recall: 0.1160
Average F1 Score: 0.1120

Threshold: 0.0762
Average Precision: 0.3604
Average Recall: 0.1290
Average F1 Score: 0.1144

Threshold: 0.0811
Average Precision: 0.3579
Average Recall: 0.1630
Average F1 Score: 0.1271

Threshold: 0.0831
Average Precision: 0.3380
Average Recall: 0.1750
Average F1 Score: 0.1293

Threshold: 0.0873
Average Precision: 0.2915
Average Recall: 0.2030
Average F1 Score: 0.1380

Threshold: 0.0933
Average Precision: 0.2377
Average Recall: 0.2290
Average F1 Score: 0.1349

Threshold: 0.1001
Average Precision: 0.2266
Average Recall: 0.2610
Average F1 Score: 0.1391

Threshold: 0.1004
Average Precision: 0.2231
Average Recall: 0.2630
Average F1 Score: 0.1396

Threshold: 0.1021
Average Precision: 0.2210
Average Recall: 0.2670
Average F1 Score: 0.1392

Threshold: 0.1097
Average Precision: 0.1763
Average Recall: 0.2990
Average F1 Score: 0.1388

Threshold: 0.1185
Average Precision: 0.1426
Average Recall: 0.3440
Average F1 Score: 0.1436

Threshold: 0.1237
Average Precision: 0.1489
Average Recall: 0.3810
Average F1 Score: 0.1556

Threshold: 0.1279
Average Precision: 0.1377
Average Recall: 0.3970
Average F1 Score: 0.1534

Threshold: 0.0554
Average Precision: 0.5766
Average Recall: 0.0370
Average F1 Score: 0.0605

Threshold: 0.0584
Average Precision: 0.5269
Average Recall: 0.0470
Average F1 Score: 0.0713

Threshold: 0.0609
Average Precision: 0.4809
Average Recall: 0.0590
Average F1 Score: 0.0815

Threshold: 0.0652
Average Precision: 0.4540
Average Recall: 0.0750
Average F1 Score: 0.0942

Threshold: 0.0671
Average Precision: 0.4400
Average Recall: 0.0830
Average F1 Score: 0.0993

Threshold: 0.0688
Average Precision: 0.4345
Average Recall: 0.0920
Average F1 Score: 0.1029

Threshold: 0.0720
Average Precision: 0.3747
Average Recall: 0.1090
Average F1 Score: 0.1099

Threshold: 0.0734
Average Precision: 0.3675
Average Recall: 0.1170
Average F1 Score: 0.1119

Threshold: 0.0748
Average Precision: 0.3650
Average Recall: 0.1250
Average F1 Score: 0.1146

Threshold: 0.0775
Average Precision: 0.3578
Average Recall: 0.1340
Average F1 Score: 0.1148

Threshold: 0.0787
Average Precision: 0.3586
Average Recall: 0.1460
Average F1 Score: 0.1199

Threshold: 0.0799
Average Precision: 0.3584
Average Recall: 0.1540
Average F1 Score: 0.1227

Threshold: 0.0822
Average Precision: 0.3365
Average Recall: 0.1680
Average F1 Score: 0.1263

Threshold: 0.0834
Average Precision: 0.3303
Average Recall: 0.1770
Average F1 Score: 0.1304
```

```
Threshold: 0.0924
Average Precision: 0.2537
Average Recall: 0.2270
Average F1 Score: 0.1365

Threshold: 0.0942
Average Precision: 0.2288
Average Recall: 0.2350
Average F1 Score: 0.1360

Threshold: 0.0952
Average Precision: 0.2318
Average Recall: 0.2370
Average F1 Score: 0.1356

Threshold: 0.0961
Average Precision: 0.2348
Average Recall: 0.2410
Average F1 Score: 0.1362

Threshold: 0.0969
Average Precision: 0.2297
Average Recall: 0.2440
Average F1 Score: 0.1353

Threshold: 0.0978
Average Precision: 0.2285
Average Recall: 0.2490
Average F1 Score: 0.1364

Threshold: 0.0987
Average Precision: 0.2282
Average Recall: 0.2550
Average F1 Score: 0.1395

Threshold: 0.0995
Average Precision: 0.2246
Average Recall: 0.2570
Average F1 Score: 0.1379

Threshold: 0.1013
Average Precision: 0.2236
Average Recall: 0.2650
Average F1 Score: 0.1404


--- Overall Average Evaluation Results for 120 Bins ---
Overall Average Precision: 0.3189
Overall Average Recall: 0.1870
Overall Average F1 Score: 0.1231
```

```
Time taken for 180 bins: 103.2280 seconds
```

```


--- Overall Average Evaluation Results for 180 Bins ---
Overall Average Precision: 0.3450
Overall Average Recall: 0.1753
Overall Average F1 Score: 0.1144

```

```
Time taken for 256 bins: 127.4501 seconds
```

```

--- Overall Average Evaluation Results for 256 Bins ---
Overall Average Precision: 0.3685
Overall Average Recall: 0.2018
Overall Average F1 Score: 0.1403

```

Discussing the results above:

- **ROC and AUC**

For 120 bins, the AUC is 0.52, which is slightly better than random guessing but still indicates poor discriminative ability.

For 180 bins, the AUC drops to 0.50, which suggests no better than random chance.

For 256 bins, the AUC is 0.42, which is even worse than random chance, indicating that the model is performing poorly at this bin size.

- **Evaluation Metrics:**

With 120 bins, the overall average precision is around 0.3189, recall is 0.1870, and F1 score is 0.1231.

With 180 bins, the overall average precision decreases slightly to 0.3450, recall is 0.1753, and F1 score is 0.1144.

With 256 bins, there's an increase in overall average precision to 0.3685, recall also increases to 0.2018, and F1 score goes up to 0.1403.

when looking at the overall average precision, recall, and F1 score, there is a slight improvement as the number of bins increases to 256. This could indicate that for this particular dataset and retrieval task, having more bins allows for a more detailed color histogram, which may capture more nuances in the color distributions of the images.

However, an AUC below 0.5 for the 256 bins model suggests that despite the apparent improvements in precision and recall, the model may be making decisions that are systematically incorrect, or the method used to construct the ROC curve may have some issues. The time taken for computations increases with the number of bins, which is expected.

While the precision and recall seem to improve with more bins, the ROC AUC values suggest that the system is not effectively discriminating between the classes, especially at

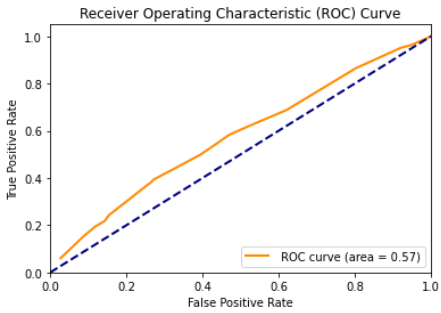
256 bins. These conflicting signals imply that while the system might be retrieving more relevant items, it is also making many incorrect classifications. This scenario may warrant further investigation into the dataset, and sure the feature extraction method itself, since it is very basic simple model.

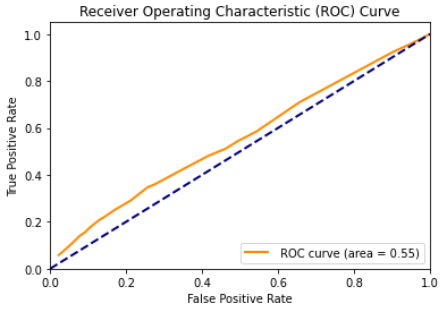
Results Using these thresholds for the 10 queries: (Test 3)

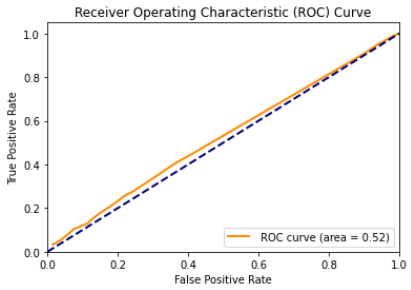
For the following thresholds:

quantiles = [0.0111,0.00111 , 0.03,0.05,0.059,0.07, 0.09,0.099,.119,0.15,0.188, 0.19,0.2,0.246,0.3,0.333,0.36, 0.4,0.5,0.8,.88,.85,.66,.9,0.77,0.67,.73,.95, .999,1]

AUC values: Decreasing from 0.57 for 120 bins , to .55 for 180 bins, to 0.52 for 256 bins, as shown below.







The overall average measures:

```
--- Overall Average Evaluation Results for 120 Bins ---
Overall Average Precision: 0.2379
Overall Average Recall: 0.4960
Overall Average F1 Score: 0.1441
```

```
--- Overall Average Evaluation Results for 180 Bins ---
Overall Average Precision: 0.2100
Overall Average Recall: 0.5389
Overall Average F1 Score: 0.1648
```

```
--- Overall Average Evaluation Results for 256 Bins ---
Overall Average Precision: 0.2617
Overall Average Recall: 0.5390
Overall Average F1 Score: 0.1631
```

AUC of 0.57 for Bins 120: An AUC value of 0.57 is generally considered slightly better than random guessing. However, it's not a very strong performance. In practical terms, this indicates that the model's ability to distinguish between classes is only marginally better than chance.

AUC of 0.55 for Bins 180: Here, the data is divided into 180 bins, and the model's AUC drops slightly to 0.55. This decrease, albeit small, suggests that increasing the bins does not improve the model's performance; in fact, it slightly worsens it.

AUC of 0.52 for Bins 256: With the data divided into 256 bins, the AUC further decreases to 0.52. This is a very modest improvement over random guessing and suggests that the model is not effectively learning the distinctions between classes at this level of data segmentation.

Results Using these thresholds for the 10 queries: (Test 4)

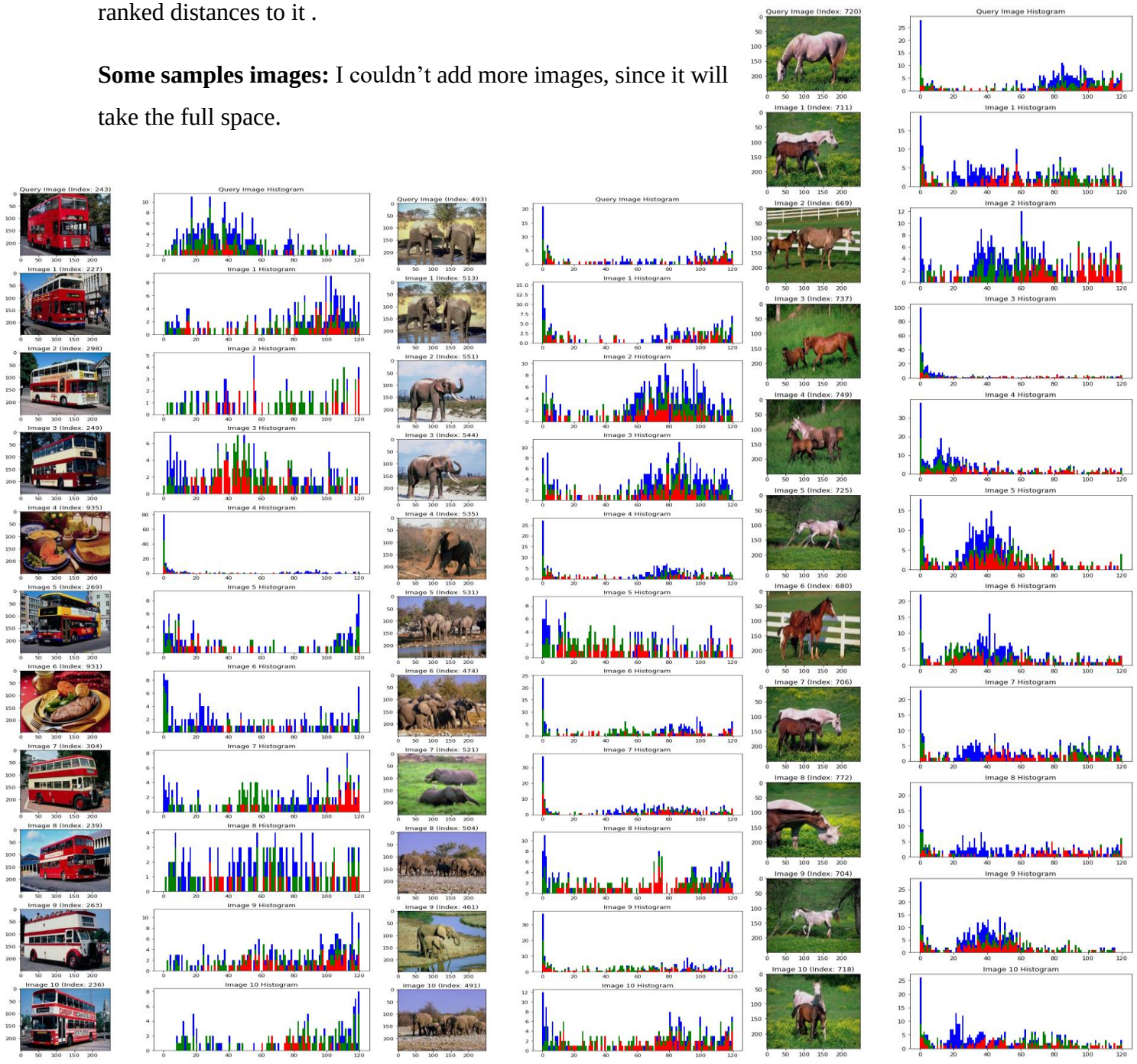
I would mention that at some point, for another test, Test4, I found that AUC increase with increasing bins but still very low increasing (From 0.39 , to 0.41, to 0.51), Which indicate that increasing number of bins has a good impact in performance.

You can find the run results of this test on Test4.log

Testing for specific 10 queries:

Here 10 specific queries chosen , each was tested on specific thresholds values suits the ranked distances to it .

Some samples images: I couldn't add more images, since it will take the full space.



Here I am showing overall results of test5.log, in test5.log and Task1and2-specific query.log you can find the testing results with different thresholds for these specific 10 queries, both gave near results to each other's.

```
Time taken for 120 bins: 81.5891 seconds
```

```
--- Overall Average Evaluation Results for 120 Bins ---
Overall Average Precision: 0.2537
Overall Average Recall: 0.4427
Overall Average F1 Score: 0.2372
```

```
Time taken for 180 bins: 101.6776 seconds
```

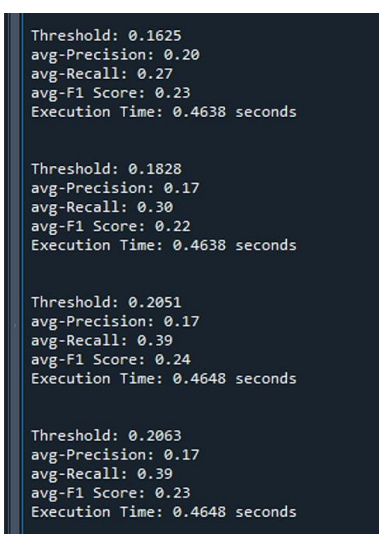
```
--- Overall Average Evaluation Results for 180 Bins ---
Overall Average Precision: 0.4138
Overall Average Recall: 0.1114
Overall Average F1 Score: 0.1120
```

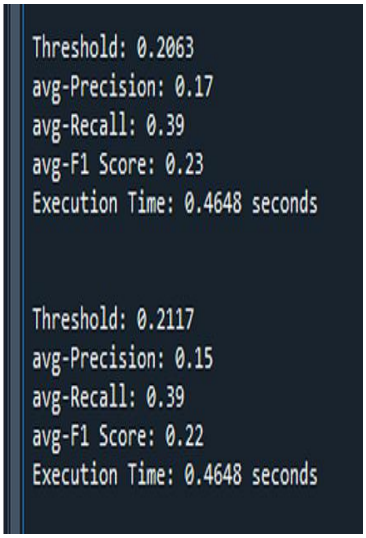
```
Time taken for 256 bins: 128.7877 seconds
```

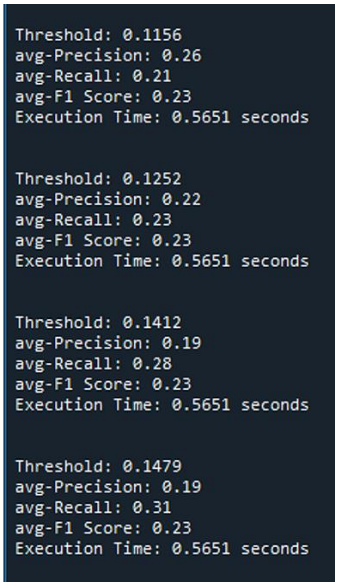
```
--- Overall Average Evaluation Results for 256 Bins ---
Overall Average Precision: 0.4920
Overall Average Recall: 0.0868
Overall Average F1 Score: 0.0882
```

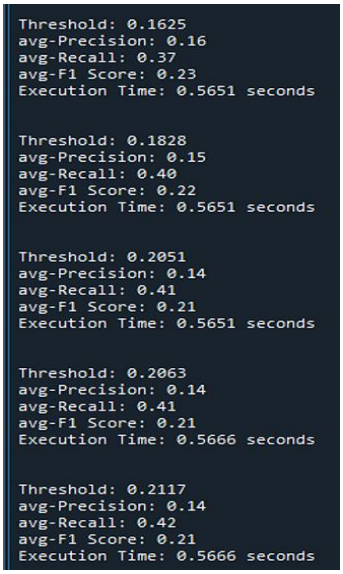
Here, because I chosen thresholds per query , I got better results for 120 bins as shown , but as increasing number of bins , precision increase but recall decrease , and overall performance decrease (f1 score decreased).

Some good results on some thresholds:









Conclusion for this part:

In good datasets higher bin counts may lead to better discrimination between relevant and irrelevant images for the task and potentially improve the AUC and ROC performance. In this work, the evaluation of the model's performance with varying bin sizes and thresholds reveals nuanced findings. The ROC and AUC analysis indicate that, overall, the model struggles to effectively discriminate between classes. While the precision, recall, and F1 score show a slight improvement as the number of bins increases to 256, the conflicting signals from the AUC values suggest potential issues in constructing the ROC curve. The reasons of this decrease in AUC are explained before during discussion results.

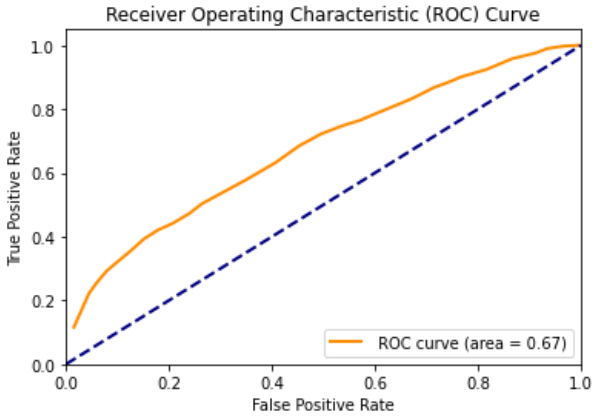
The results from different tests, such as Test 4, demonstrate varying impacts on performance with an increasing number of bins. In some cases, the AUC increases marginally, indicating a potential positive influence on the model's discriminatory ability. However, this improvement is modest and may not be consistent across different datasets or retrieval tasks.

The choice of thresholds per query in Test 5 leads to improved results with 120 bins, but as the number of bins increases, precision improves while recall decreases, resulting in an overall decrease in performance (F1 score). This suggests a delicate balance between granularity and computation time, with 120 bins providing a medium level of granularity, 180 bins offering higher resolution, and 256 bins providing the highest level of granularity.

3.2 Task 3: CBIR System Using Color Moments

Subsection 3.2.1: Experiment with Equal Weights

ROC and AUC:



Some measures result at many thresholds for different queries:

```
Average values for all thresholds for Query Index: 959
avg-Precision: 0.29
avg-Recall: 0.26
avg-F1 Score: 0.23
Query Index: 440
```

```
At Threshold: 0.1818, Precision: 0.40, Recall: 0.41, F1 Sc
Average values for all thresholds for Query Index: 440
avg-Precision: 0.31
avg-Recall: 0.28
avg-F1 Score: 0.25
Query Index: 979
```

```
Average values for all thresholds for Query Index: 979
avg-Precision: 0.31
avg-Recall: 0.29
avg-F1 Score: 0.26
Query Index: 373
```

```
At Threshold: 0.1818, Precision: 0.56, Recall: 0.53, F1 Sc
Average values for all thresholds for Query Index: 373
avg-Precision: 0.34
avg-Recall: 0.30
avg-F1 Score: 0.29
Query Index: 472
```

```
Average values for all thresholds for Query Index: 472
avg-Precision: 0.34
avg-Recall: 0.28
avg-F1 Score: 0.27
```

```
Threshold: 0.1818
Average Precision: 0.2632
Average Recall: 0.2000
Average F1 Score: 0.2273

Threshold: 0.1907
Average Precision: 0.2796
Average Recall: 0.2600
Average F1 Score: 0.2694

Threshold: 0.2029
Average Precision: 0.2881
Average Recall: 0.3400
Average F1 Score: 0.3119

Threshold: 0.2155
Average Precision: 0.2676
Average Recall: 0.3800
Average F1 Score: 0.3140

Threshold: 0.2060
Average Precision: 0.2960
Average Recall: 0.3700
Average F1 Score: 0.3289

Threshold: 0.2187
Average Precision: 0.2550
Average Recall: 0.3800
Average F1 Score: 0.3052
```

```
--- Overall Average Evaluation Results ---
Overall Average Precision: 0.2953
Overall Average Recall: 0.1414
Overall Average F1 Score: 0.1604
```

```
Total Execution Time: 12.0261 seconds
```

Full run results could be found on Task3part1.log , the results above show better f1scores compared with the ones we got on the task2 when we used the same queries and same thresholds.

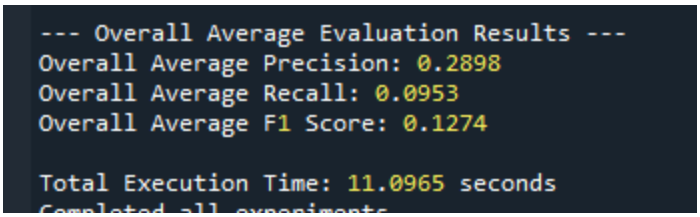
They got a ROC AUC score of 0.67. An AUC of 0.67 suggests that the model has moderate discriminatory ability. As adjusting the threshold can impact precision, recall, and the F1 score, the images above show good thresholds since they give good trade off.

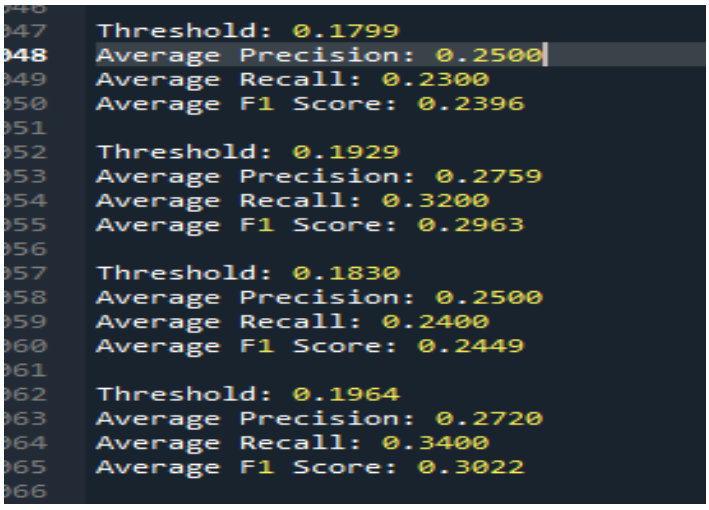
Subsection 3.2.2: Experiment with Different Weights

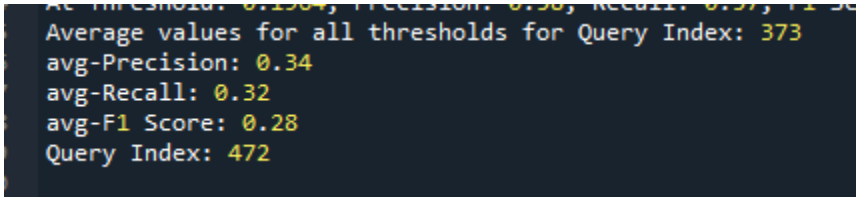
1. *Emphasizing Mean (Average Color)*
   - Weights: **[3, 1, 1] * 3** (3 times the weight for the mean of each color channel)

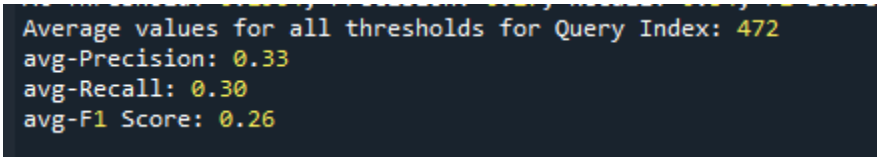
This implies that the mean of each color channel is given three times the weight compared to the other color moments (variance and skewness). Such a weighting scheme suggests a hypothesis that the average color of an image is more significant for retrieval purposes than its color distribution and asymmetry.

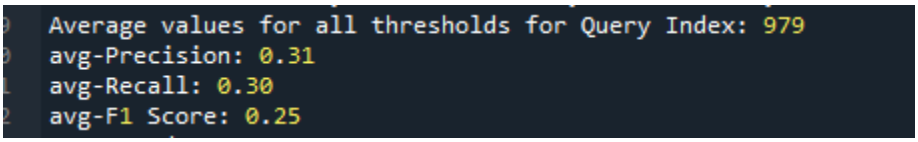
Auc measure was 0.66





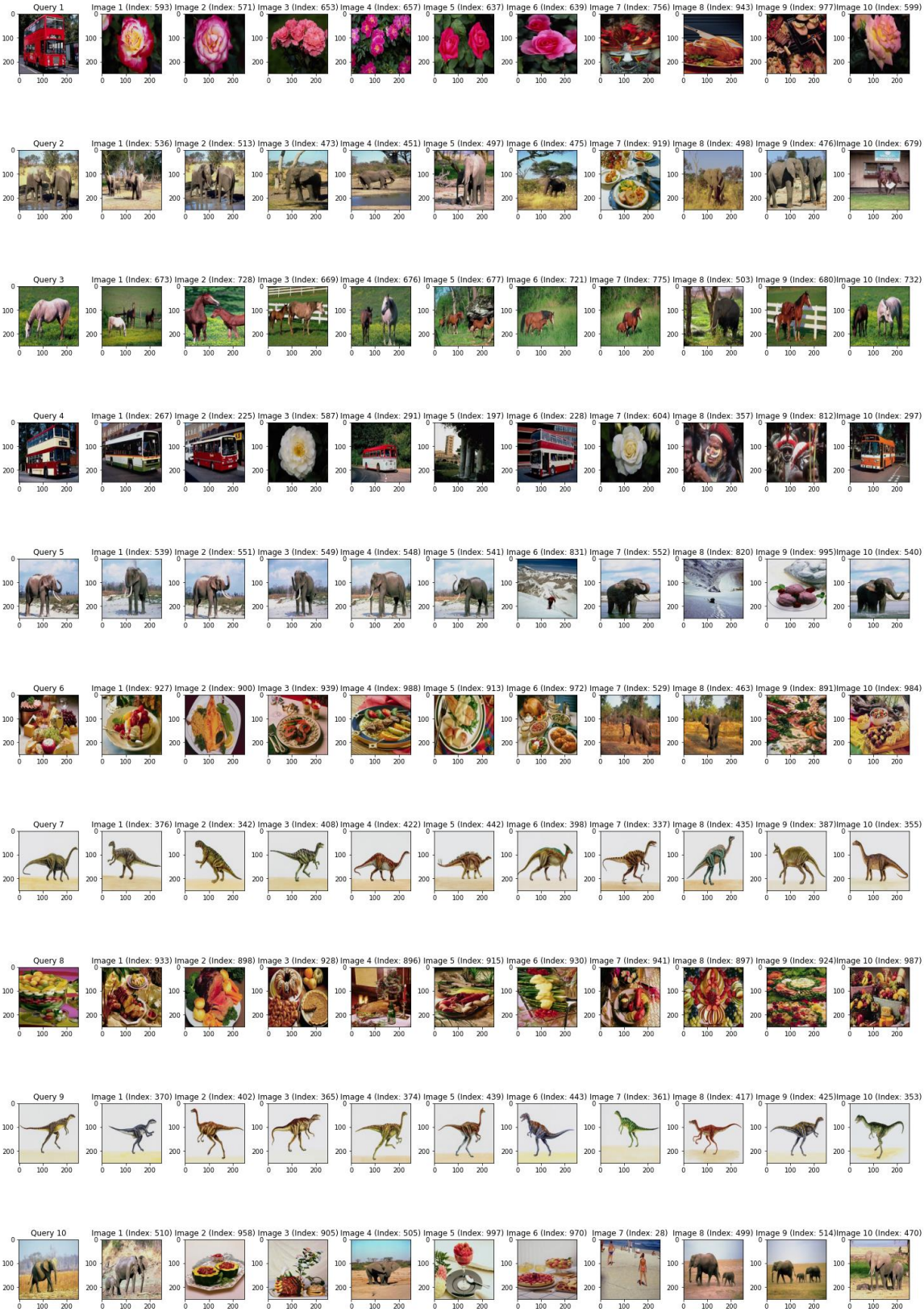






- Increasing the weight of mean more : weights = **[10, 1, 1]*3**

Here, the mean is given even more prominence, suggesting that the average color is thought to be far more critical compared to other weights.


Query 1
Image 1 (Index: 593) Image 2 (Index: 571) Image 3 (Index: 653) Image 4 (Index: 657) Image 5 (Index: 637) Image 6 (Index: 639) Image 7 (Index: 756) Image 8 (Index: 943) Image 9 (Index: 977) Image 10 (Index: 599)
Query 2
Image 1 (Index: 536) Image 2 (Index: 513) Image 3 (Index: 473) Image 4 (Index: 451) Image 5 (Index: 497) Image 6 (Index: 475) Image 7 (Index: 919) Image 8 (Index: 498) Image 9 (Index: 476) Image 10 (Index: 679)
Query 3
Image 1 (Index: 673) Image 2 (Index: 728) Image 3 (Index: 669) Image 4 (Index: 676) Image 5 (Index: 677) Image 6 (Index: 721) Image 7 (Index: 775) Image 8 (Index: 503) Image 9 (Index: 680) Image 10 (Index: 732)
Query 4
Image 1 (Index: 267) Image 2 (Index: 225) Image 3 (Index: 587) Image 4 (Index: 291) Image 5 (Index: 197) Image 6 (Index: 228) Image 7 (Index: 604) Image 8 (Index: 357) Image 9 (Index: 812) Image 10 (Index: 297)
Query 5
Image 1 (Index: 539) Image 2 (Index: 551) Image 3 (Index: 549) Image 4 (Index: 548) Image 5 (Index: 541) Image 6 (Index: 831) Image 7 (Index: 552) Image 8 (Index: 820) Image 9 (Index: 995) Image 10 (Index: 540)
Query 6
Image 1 (Index: 927) Image 2 (Index: 900) Image 3 (Index: 939) Image 4 (Index: 988) Image 5 (Index: 913) Image 6 (Index: 972) Image 7 (Index: 529) Image 8 (Index: 463) Image 9 (Index: 891) Image 10 (Index: 984)
Query 7
Image 1 (Index: 376) Image 2 (Index: 342) Image 3 (Index: 408) Image 4 (Index: 422) Image 5 (Index: 442) Image 6 (Index: 398) Image 7 (Index: 337) Image 8 (Index: 435) Image 9 (Index: 387) Image 10 (Index: 355)
Query 8
Image 1 (Index: 933) Image 2 (Index: 898) Image 3 (Index: 928) Image 4 (Index: 896) Image 5 (Index: 915) Image 6 (Index: 930) Image 7 (Index: 941) Image 8 (Index: 897) Image 9 (Index: 924) Image 10 (Index: 987)
Query 9
Image 1 (Index: 370) Image 2 (Index: 402) Image 3 (Index: 365) Image 4 (Index: 374) Image 5 (Index: 439) Image 6 (Index: 443) Image 7 (Index: 361) Image 8 (Index: 417) Image 9 (Index: 425) Image 10 (Index: 353)
Query 10
Image 1 (Index: 510) Image 2 (Index: 958) Image 3 (Index: 905) Image 4 (Index: 505) Image 5 (Index: 997) Image 6 (Index: 970) Image 7 (Index: 28) Image 8 (Index: 499) Image 9 (Index: 514) Image 10 (Index: 470)

Auc measure was 0.66, and the other measures as shown :

```
--- Overall Average Evaluation Results ---
Overall Average Precision: 0.2806
Overall Average Recall: 0.0910
Overall Average F1 Score: 0.1219

Total Execution Time: 10.0233 seconds
```

```
Threshold: 0.1771
Average Precision: 0.2500
Average Recall: 0.2200
Average F1 Score: 0.2340

Threshold: 0.1902
Average Precision: 0.2696
Average Recall: 0.3100
Average F1 Score: 0.2884

Threshold: 0.1803
Average Precision: 0.2581
Average Recall: 0.2400
Average F1 Score: 0.2487

Threshold: 0.1938
Average Precision: 0.2705
Average Recall: 0.3300
Average F1 Score: 0.2973
```

```
Average values for all thresholds for Query Index: 373
avg-Precision: 0.33
avg-Recall: 0.31
avg-F1 Score: 0.28
```

```
Average values for all thresholds for Query Index: 979
avg-Precision: 0.31
avg-Recall: 0.30
avg-F1 Score: 0.25
```

Comparing Results:

By increasing the weight of the mean color, one might anticipate an improvement in the precision, recall, and F1 score if the mean color is indeed a more critical factor for the retrieval process, as the system would become more attuned to the average color differences between images. However, overemphasizing a single aspect can lead to the

neglect of other significant features like texture or shape, which are partly reflected by the variance and skewness of colors. This could result in suboptimal performance. In this case, there is no decline in performance, but rather a lack of enhancement; the outcomes remain consistent. For instance, with equal weights, the AUC was 0.67, and upon emphasizing the mean, it stands at 0.66. As for the F1 score, which was initially 0.1604, it slightly decreased to 0.1274 after tripling the weight of the mean, and with a tenfold increase in the mean's weight, there was virtually no change, dropping to just 0.1219. Therefore, in this scenario, giving additional weight to the mean color alone does not yield a significant benefit.

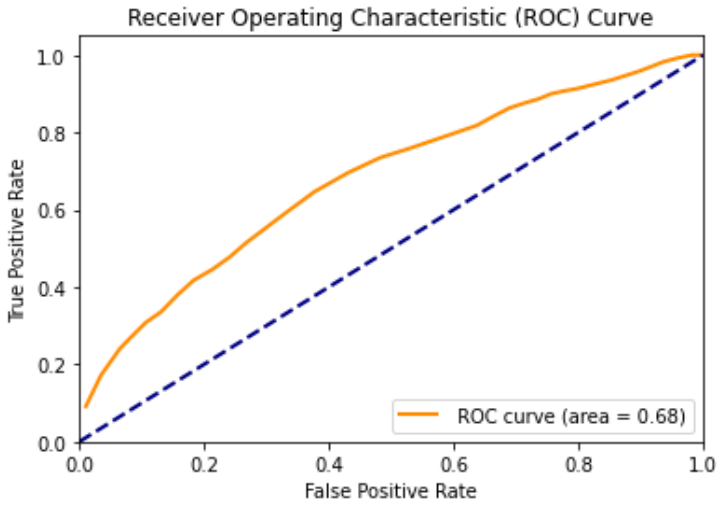
The AUC (Area Under the Curve) for both the [3, 1, 1] * 3 and [10, 1, 1] * 3 weighted scenarios is 0.66, which is the same. This indicates that there is no significant difference in the discriminative power of the classifier when changing the weights from a 3-times emphasis to a 10-times emphasis on the mean color. This could be interpreted in several ways:

- The mean color is not the only critical feature: It's possible that the mean color alone isn't sufficiently discriminative, and by focusing too much on it, the system is not improving.
- Diminishing Returns: There might be a point beyond which giving more weight to the mean color doesn't contribute to a better separation of classes because the system needs a balanced consideration of all color moments.
- Dataset Characteristics: The specific dataset used for evaluation might not have significant variations in mean color across different classes, making it a less effective feature for discrimination.
- System Limitations: The CBIR system itself might have limitations, such as in the way it uses the weighted features, that prevent it from benefiting from the changed weights.

2. *Emphasizing Contrast (Standard Deviation)*

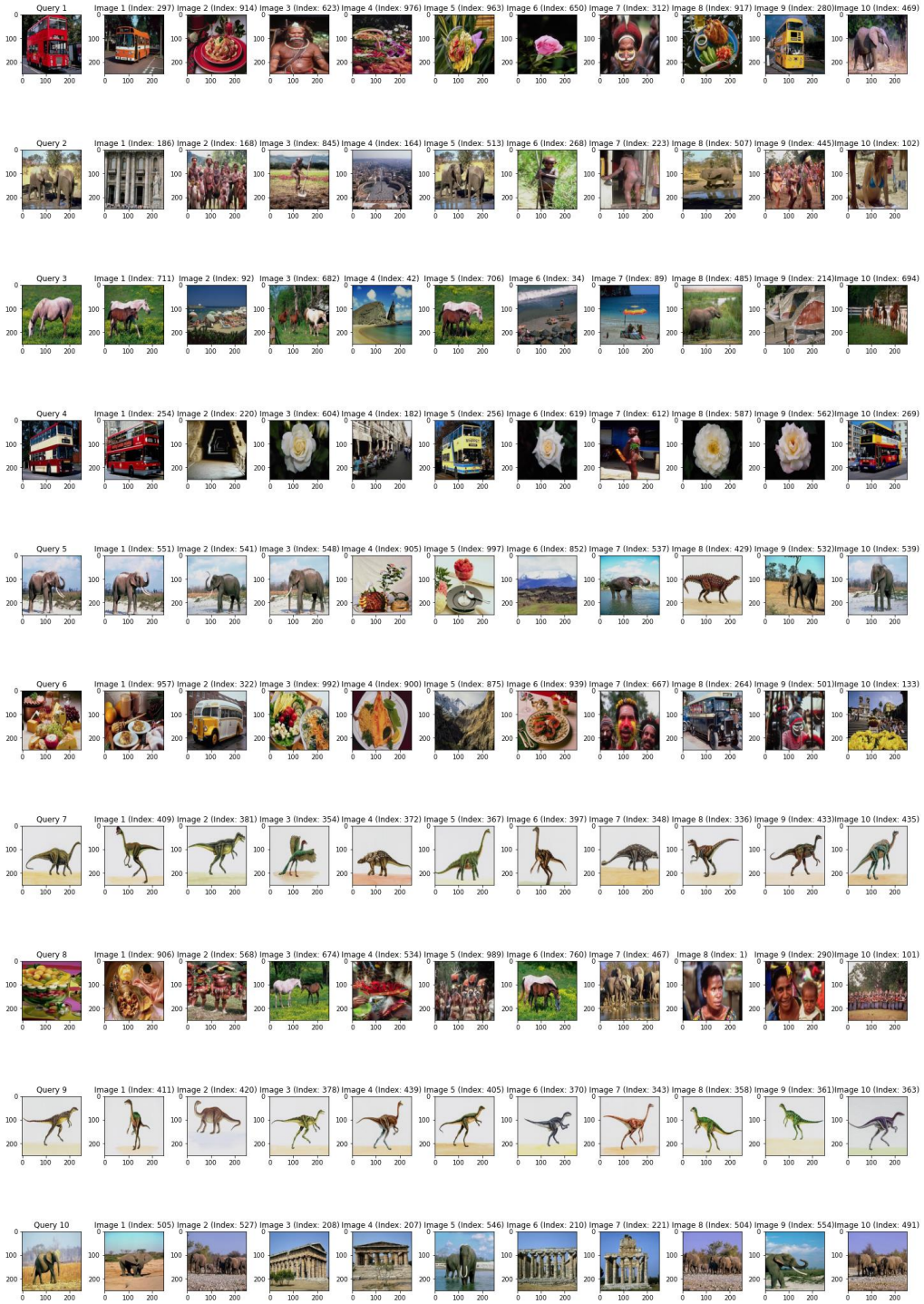
- Weights: **[1, 3, 1] * 3** (3 times the weight for the standard deviation of each color channel)

Auc measure was 0.68, and the other measures as shown:




```
Threshold: 0.2248
Average Precision: 0.3171
Average Recall: 0.3900
Average F1 Score: 0.3498

Threshold: 0.2325
Average Precision: 0.3134
Average Recall: 0.4200
Average F1 Score: 0.3590

Threshold: 0.2403
Average Precision: 0.3025
Average Recall: 0.4900
Average F1 Score: 0.3740

Threshold: 0.2479
Average Precision: 0.3027
Average Recall: 0.5600
Average F1 Score: 0.3930

Threshold: 0.2595
Average Precision: 0.2791
Average Recall: 0.6000
Average F1 Score: 0.3810

Threshold: 0.2750
Average Precision: 0.2679
Average Recall: 0.7100
Average F1 Score: 0.3890

Threshold: 0.2908
Average Precision: 0.2542
Average Recall: 0.7600
Average F1 Score: 0.3810

Threshold: 0.2789
Average Precision: 0.2667
Average Recall: 0.7200
Average F1 Score: 0.3892

Threshold: 0.2948
Average Precision: 0.2516
Average Recall: 0.7800
Average F1 Score: 0.3805


--- Overall Average Evaluation Results ---
Overall Average Precision: 0.3055
Overall Average Recall: 0.2102
Overall Average F1 Score: 0.1877

Total Execution Time: 10.3223 seconds
```

- Weights: **[1, 10, 1] * 3** (3 times the weight for the standard deviation of each color channel), **<u>f1 score increased from the value when had equal weight</u>**

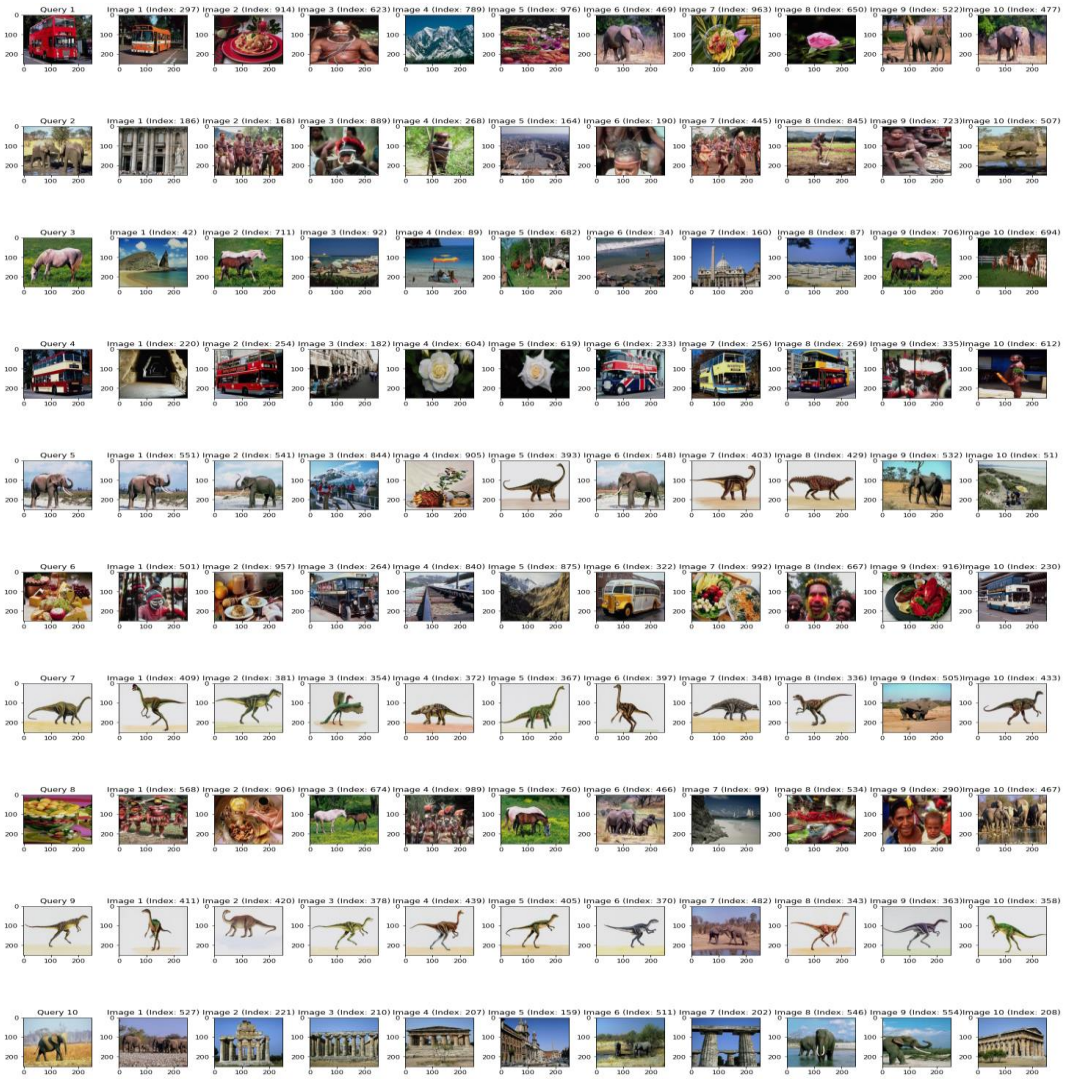


```
--- Overall Average Evaluation Results ---
Overall Average Precision: 0.2190
Overall Average Recall: 0.3058
Overall Average F1 Score: 0.2183

Total Execution Time: 10.3502 seconds
Completed all experiments.
```

AUC was 0.64, f1 score increased more.

- Weights: [1, 20, 1] * 3 (3 times the weight for the standard deviation of each color channel)

AUC was 0.62, f1 score remained as for weight 10 .



```
--- Overall Average Evaluation Results ---
Overall Average Precision: 0.1916
Overall Average Recall: 0.2965
Overall Average F1 Score: 0.2073
```

Comparing Results:

When the Content-Based Image Retrieval (CBIR) system was configured with equal weights across all color channels, the resulting Area Under the Curve (AUC) was 0.67, with an F1 score of 0.1604. Adjusting the system to emphasize contrast by increasing the weight of the standard deviation of each color channel to three times (with weights [1, 3, 1] * 3) resulted in a slight improvement in the AUC to 0.68. This suggests that contrast may play a more critical role in the retrieval process than previously assumed.

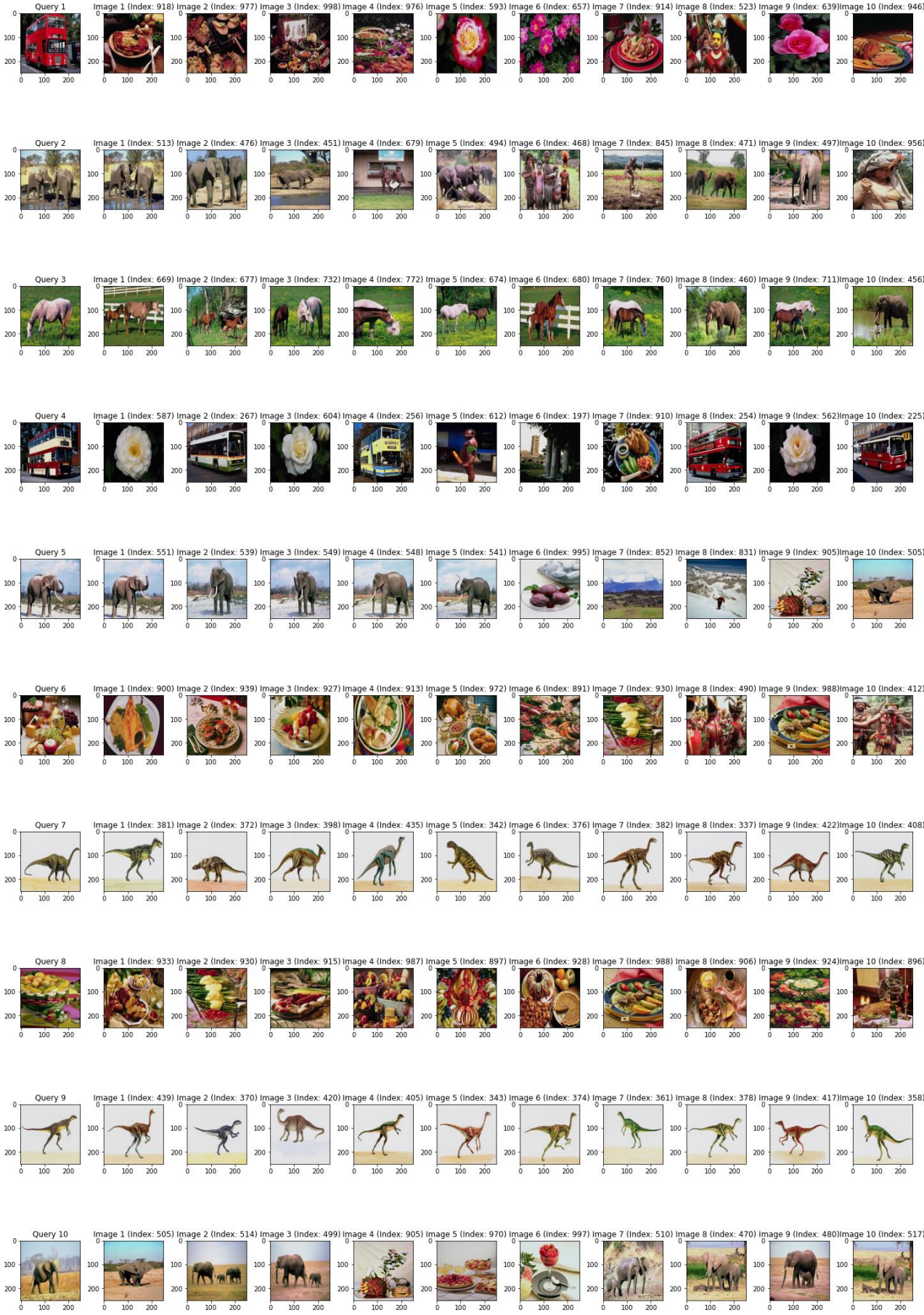
Further emphasizing contrast by increasing the weight to ten times the other factors (with weights [1, 10, 1] * 3) led to a decrease in AUC to 0.64. However, it's notable that the F1 score increased compared to the scenario with equal weights, indicating that while the model's overall discriminative power as measured by AUC decreased, its harmonic balance of precision and recall improved.

Pushing the emphasis on contrast even further to twenty times the other factors (with weights [1, 20, 1] * 3) saw the AUC decrease further to 0.62. Despite this continued decline in AUC, the F1 score remained at the level it was when the weight was ten, indicating a plateau in the benefits gained from further increasing the weight of contrast in terms of F1 score.

3. Emphasizing Color Symmetry (Skewness)

- Weights: **[1, 1, 3] * 3** (3 times the weight for skewness of each color channel)

Auc was 0.67, f1 score 0.146

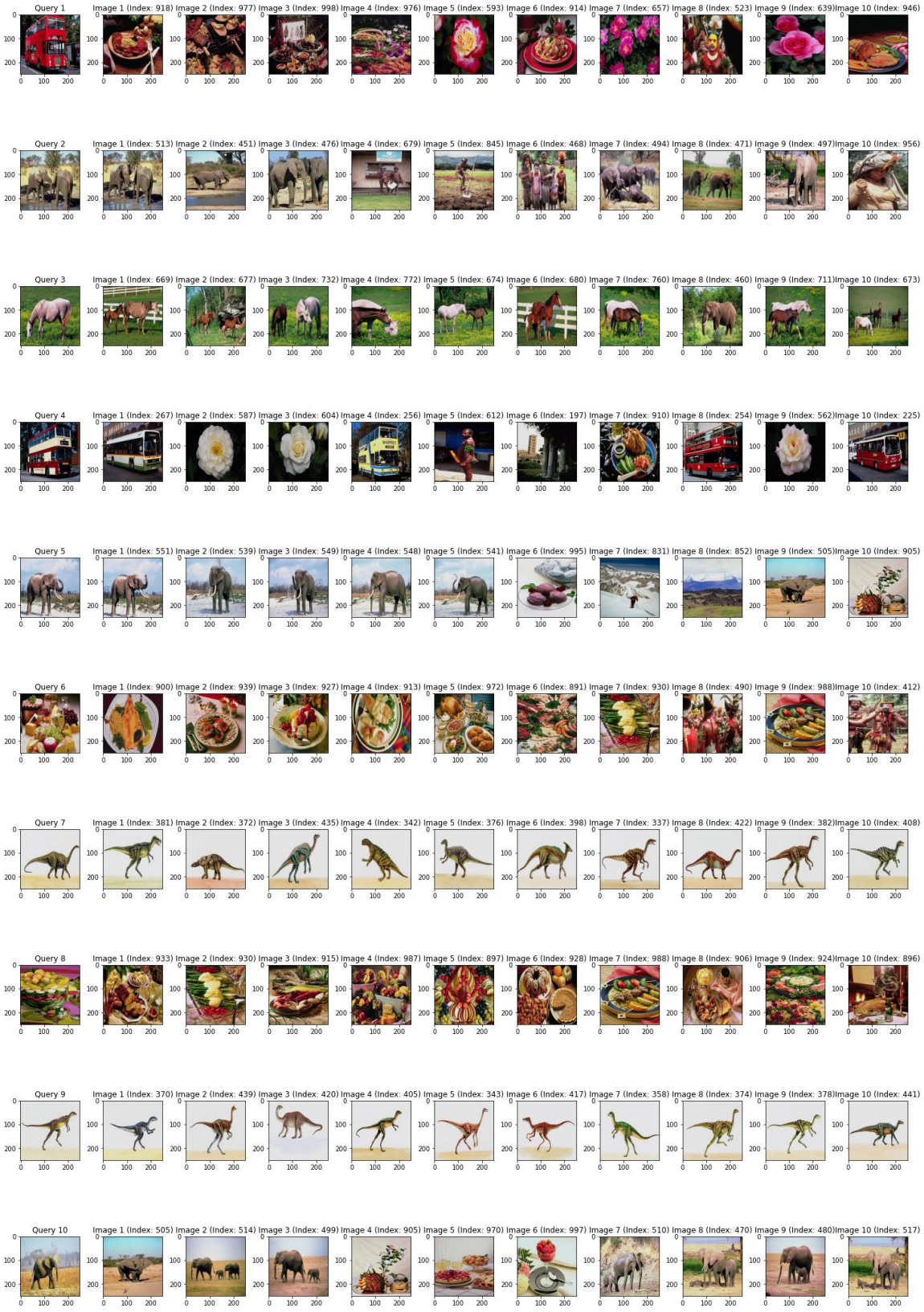

Query 1
Image 1 (Index: 918)
Image 2 (Index: 977)
Image 3 (Index: 998)
Image 4 (Index: 976)
Image 5 (Index: 593)
Image 6 (Index: 657)
Image 7 (Index: 914)
Image 8 (Index: 523)
Image 9 (Index: 639)
Image 10 (Index: 946)
Query 2
Image 1 (Index: 513)
Image 2 (Index: 476)
Image 3 (Index: 451)
Image 4 (Index: 679)
Image 5 (Index: 494)
Image 6 (Index: 468)
Image 7 (Index: 845)
Image 8 (Index: 471)
Image 9 (Index: 497)
Image 10 (Index: 956)
Query 3
Image 1 (Index: 669)
Image 2 (Index: 677)
Image 3 (Index: 732)
Image 4 (Index: 772)
Image 5 (Index: 674)
Image 6 (Index: 680)
Image 7 (Index: 760)
Image 8 (Index: 460)
Image 9 (Index: 711)
Image 10 (Index: 456)
Query 4
Image 1 (Index: 587)
Image 2 (Index: 267)
Image 3 (Index: 604)
Image 4 (Index: 256)
Image 5 (Index: 612)
Image 6 (Index: 197)
Image 7 (Index: 910)
Image 8 (Index: 254)
Image 9 (Index: 562)
Image 10 (Index: 225)
Query 5
Image 1 (Index: 551)
Image 2 (Index: 539)
Image 3 (Index: 549)
Image 4 (Index: 548)
Image 5 (Index: 541)
Image 6 (Index: 995)
Image 7 (Index: 852)
Image 8 (Index: 831)
Image 9 (Index: 905)
Image 10 (Index: 505)
Query 6
Image 1 (Index: 900)
Image 2 (Index: 939)
Image 3 (Index: 927)
Image 4 (Index: 913)
Image 5 (Index: 972)
Image 6 (Index: 891)
Image 7 (Index: 930)
Image 8 (Index: 490)
Image 9 (Index: 988)
Image 10 (Index: 412)
Query 7
Image 1 (Index: 381)
Image 2 (Index: 372)
Image 3 (Index: 398)
Image 4 (Index: 435)
Image 5 (Index: 342)
Image 6 (Index: 376)
Image 7 (Index: 382)
Image 8 (Index: 337)
Image 9 (Index: 422)
Image 10 (Index: 408)
Query 8
Image 1 (Index: 933)
Image 2 (Index: 930)
Image 3 (Index: 915)
Image 4 (Index: 987)
Image 5 (Index: 897)
Image 6 (Index: 928)
Image 7 (Index: 988)
Image 8 (Index: 906)
Image 9 (Index: 924)
Image 10 (Index: 896)
Query 9
Image 1 (Index: 439)
Image 2 (Index: 370)
Image 3 (Index: 420)
Image 4 (Index: 405)
Image 5 (Index: 343)
Image 6 (Index: 374)
Image 7 (Index: 361)
Image 8 (Index: 378)
Image 9 (Index: 417)
Image 10 (Index: 358)
Query 10
Image 1 (Index: 505)
Image 2 (Index: 514)
Image 3 (Index: 499)
Image 4 (Index: 905)
Image 5 (Index: 970)
Image 6 (Index: 997)
Image 7 (Index: 510)
Image 8 (Index: 470)
Image 9 (Index: 480)
Image 10 (Index: 517)

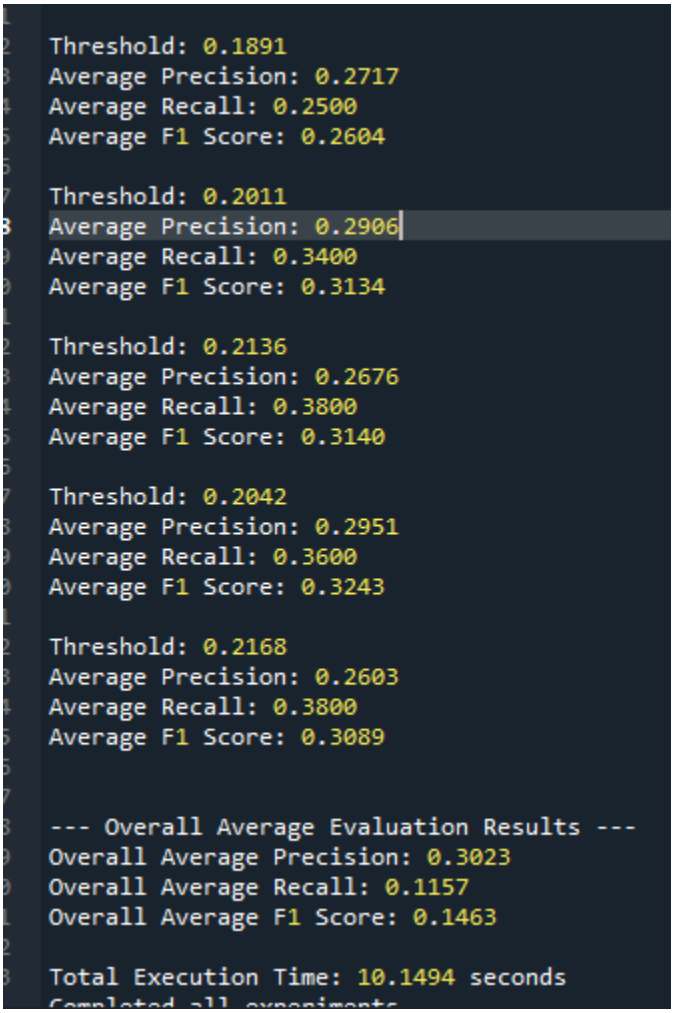
```
67
68  --- Overall Average Evaluation Results ---
69  Overall Average Precision: 0.2997
70  Overall Average Recall: 0.1157
71  Overall Average F1 Score: 0.1460
72
73  Total Execution Time: 10.7500 seconds
```

```
42  Threshold: 0.1903
43  Average Precision: 0.2717
44  Average Recall: 0.2500
45  Average F1 Score: 0.2604
46
47  Threshold: 0.2025
48  Average Precision: 0.2881
49  Average Recall: 0.3400
50  Average F1 Score: 0.3119
51
52  Threshold: 0.2151
53  Average Precision: 0.2676
54  Average Recall: 0.3800
55  Average F1 Score: 0.3140
56
57  Threshold: 0.2056
58  Average Precision: 0.2927
59  Average Recall: 0.3600
60  Average F1 Score: 0.3229
61
62  Threshold: 0.2183
63  Average Precision: 0.2550
64  Average Recall: 0.3800
65  Average F1 Score: 0.3052
66
```

- Weights: **[1, 1, 6] * 3** (3 times the weight for skewness of each color channel)

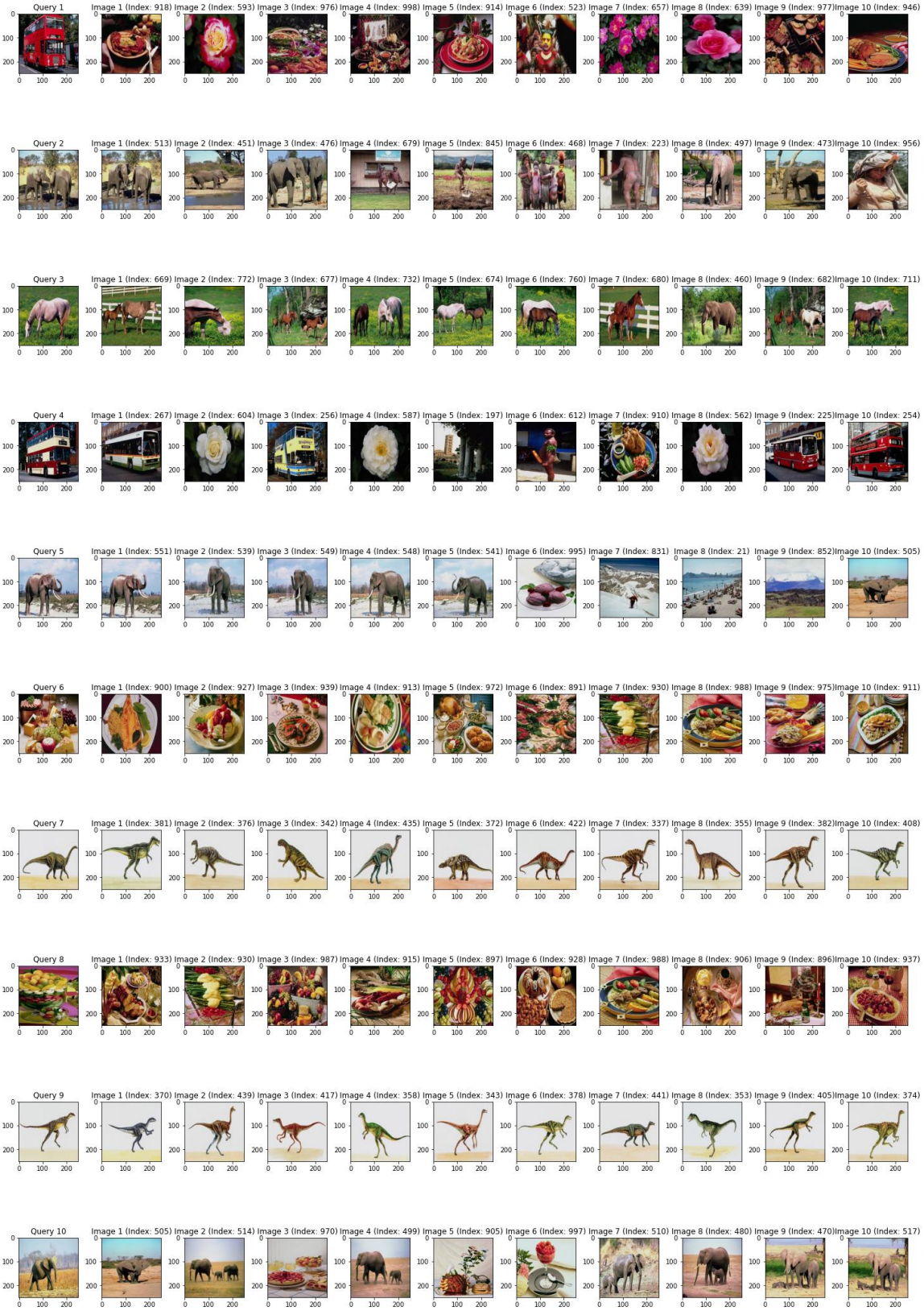
  AUC was 0.67, f1 score 0.1463

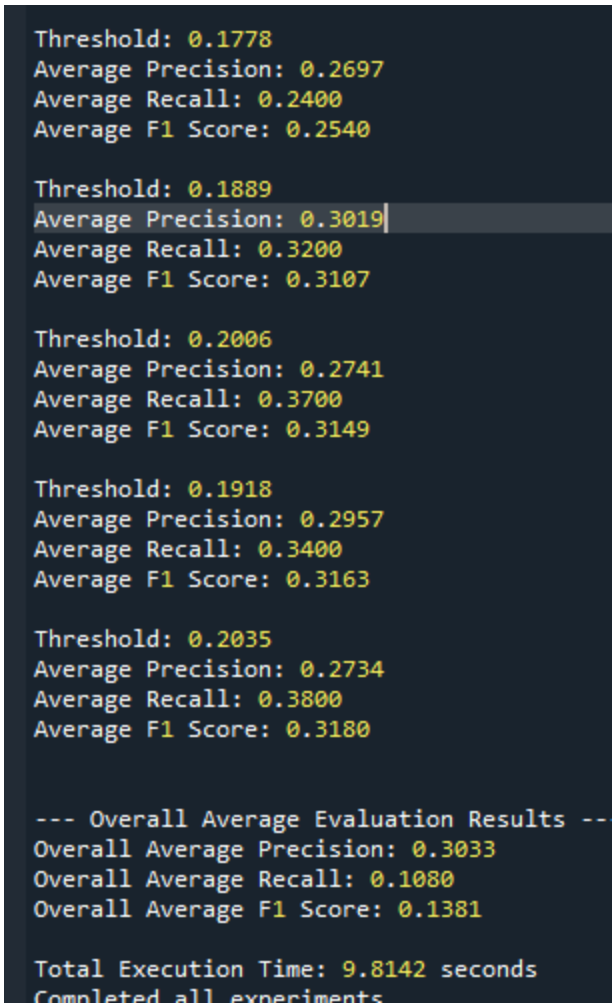

Query 1
Image 1 (Index: 918) Image 2 (Index: 977) Image 3 (Index: 998) Image 4 (Index: 976) Image 5 (Index: 593) Image 6 (Index: 914) Image 7 (Index: 657) Image 8 (Index: 523) Image 9 (Index: 639) Image 10 (Index: 946)
Query 2
Image 1 (Index: 513) Image 2 (Index: 451) Image 3 (Index: 476) Image 4 (Index: 679) Image 5 (Index: 845) Image 6 (Index: 468) Image 7 (Index: 494) Image 8 (Index: 471) Image 9 (Index: 497) Image 10 (Index: 956)
Query 3
Image 1 (Index: 669) Image 2 (Index: 677) Image 3 (Index: 732) Image 4 (Index: 772) Image 5 (Index: 674) Image 6 (Index: 680) Image 7 (Index: 760) Image 8 (Index: 460) Image 9 (Index: 711) Image 10 (Index: 673)
Query 4
Image 1 (Index: 267) Image 2 (Index: 587) Image 3 (Index: 604) Image 4 (Index: 256) Image 5 (Index: 612) Image 6 (Index: 197) Image 7 (Index: 910) Image 8 (Index: 254) Image 9 (Index: 562) Image 10 (Index: 225)
Query 5
Image 1 (Index: 551) Image 2 (Index: 539) Image 3 (Index: 549) Image 4 (Index: 548) Image 5 (Index: 541) Image 6 (Index: 995) Image 7 (Index: 831) Image 8 (Index: 852) Image 9 (Index: 505) Image 10 (Index: 905)
Query 6
Image 1 (Index: 900) Image 2 (Index: 939) Image 3 (Index: 927) Image 4 (Index: 913) Image 5 (Index: 972) Image 6 (Index: 891) Image 7 (Index: 930) Image 8 (Index: 490) Image 9 (Index: 988) Image 10 (Index: 412)
Query 7
Image 1 (Index: 381) Image 2 (Index: 372) Image 3 (Index: 435) Image 4 (Index: 342) Image 5 (Index: 376) Image 6 (Index: 398) Image 7 (Index: 337) Image 8 (Index: 422) Image 9 (Index: 382) Image 10 (Index: 408)
Query 8
Image 1 (Index: 933) Image 2 (Index: 930) Image 3 (Index: 915) Image 4 (Index: 987) Image 5 (Index: 897) Image 6 (Index: 928) Image 7 (Index: 988) Image 8 (Index: 906) Image 9 (Index: 924) Image 10 (Index: 896)
Query 9
Image 1 (Index: 370) Image 2 (Index: 439) Image 3 (Index: 420) Image 4 (Index: 405) Image 5 (Index: 343) Image 6 (Index: 417) Image 7 (Index: 358) Image 8 (Index: 374) Image 9 (Index: 378) Image 10 (Index: 441)
Query 10
Image 1 (Index: 505) Image 2 (Index: 514) Image 3 (Index: 499) Image 4 (Index: 905) Image 5 (Index: 970) Image 6 (Index: 997) Image 7 (Index: 510) Image 8 (Index: 470) Image 9 (Index: 480) Image 10 (Index: 517)



- Weights: **[1, 1, 15] * 3** (3 times the weight for skewness of each color channel)

  AUC was 0.67 , f1 score 0.1381


Query 1
Image 1 (Index: 918) Image 2 (Index: 593) Image 3 (Index: 976) Image 4 (Index: 998) Image 5 (Index: 914) Image 6 (Index: 523) Image 7 (Index: 657) Image 8 (Index: 639) Image 9 (Index: 977) Image 10 (Index: 946)
Query 2
Image 1 (Index: 513) Image 2 (Index: 451) Image 3 (Index: 476) Image 4 (Index: 679) Image 5 (Index: 845) Image 6 (Index: 468) Image 7 (Index: 223) Image 8 (Index: 497) Image 9 (Index: 473) Image 10 (Index: 956)
Query 3
Image 1 (Index: 669) Image 2 (Index: 772) Image 3 (Index: 677) Image 4 (Index: 732) Image 5 (Index: 674) Image 6 (Index: 760) Image 7 (Index: 680) Image 8 (Index: 460) Image 9 (Index: 682) Image 10 (Index: 711)
Query 4
Image 1 (Index: 267) Image 2 (Index: 604) Image 3 (Index: 256) Image 4 (Index: 587) Image 5 (Index: 197) Image 6 (Index: 612) Image 7 (Index: 910) Image 8 (Index: 562) Image 9 (Index: 225) Image 10 (Index: 254)
Query 5
Image 1 (Index: 551) Image 2 (Index: 539) Image 3 (Index: 549) Image 4 (Index: 548) Image 5 (Index: 541) Image 6 (Index: 995) Image 7 (Index: 831) Image 8 (Index: 21) Image 9 (Index: 852) Image 10 (Index: 505)
Query 6
Image 1 (Index: 900) Image 2 (Index: 927) Image 3 (Index: 939) Image 4 (Index: 913) Image 5 (Index: 972) Image 6 (Index: 891) Image 7 (Index: 930) Image 8 (Index: 988) Image 9 (Index: 975) Image 10 (Index: 911)
Query 7
Image 1 (Index: 381) Image 2 (Index: 376) Image 3 (Index: 342) Image 4 (Index: 435) Image 5 (Index: 372) Image 6 (Index: 422) Image 7 (Index: 337) Image 8 (Index: 355) Image 9 (Index: 382) Image 10 (Index: 408)
Query 8
Image 1 (Index: 933) Image 2 (Index: 930) Image 3 (Index: 987) Image 4 (Index: 915) Image 5 (Index: 897) Image 6 (Index: 928) Image 7 (Index: 988) Image 8 (Index: 906) Image 9 (Index: 896) Image 10 (Index: 937)
Query 9
Image 1 (Index: 370) Image 2 (Index: 439) Image 3 (Index: 417) Image 4 (Index: 358) Image 5 (Index: 343) Image 6 (Index: 378) Image 7 (Index: 441) Image 8 (Index: 353) Image 9 (Index: 405) Image 10 (Index: 374)
Query 10
Image 1 (Index: 505) Image 2 (Index: 514) Image 3 (Index: 970) Image 4 (Index: 499) Image 5 (Index: 905) Image 6 (Index: 997) Image 7 (Index: 510) Image 8 (Index: 480) Image 9 (Index: 470) Image 10 (Index: 517)



Comparing Results:

When the Content-Based Image Retrieval (CBIR) system utilized equal weights for all color channels, it achieved an Area Under the Curve (AUC) of 0.67 and an F1 score of 0.1604. Subsequently, the system was adjusted to emphasize color symmetry by tripling the weight for the skewness of each color channel (with weights [1, 1, 3] * 3). This alteration maintained the AUC at 0.67, but the F1 score saw a slight decrease to 0.146.

Increasing the emphasis on color symmetry further, by assigning six times the weight to skewness (with weights [1, 1, 6] * 3), resulted in the AUC remaining unchanged at 0.67, and the F1 score experienced a negligible increase to 0.1463. This minute change suggests that the impact of skewness on the retrieval performance was minimal.

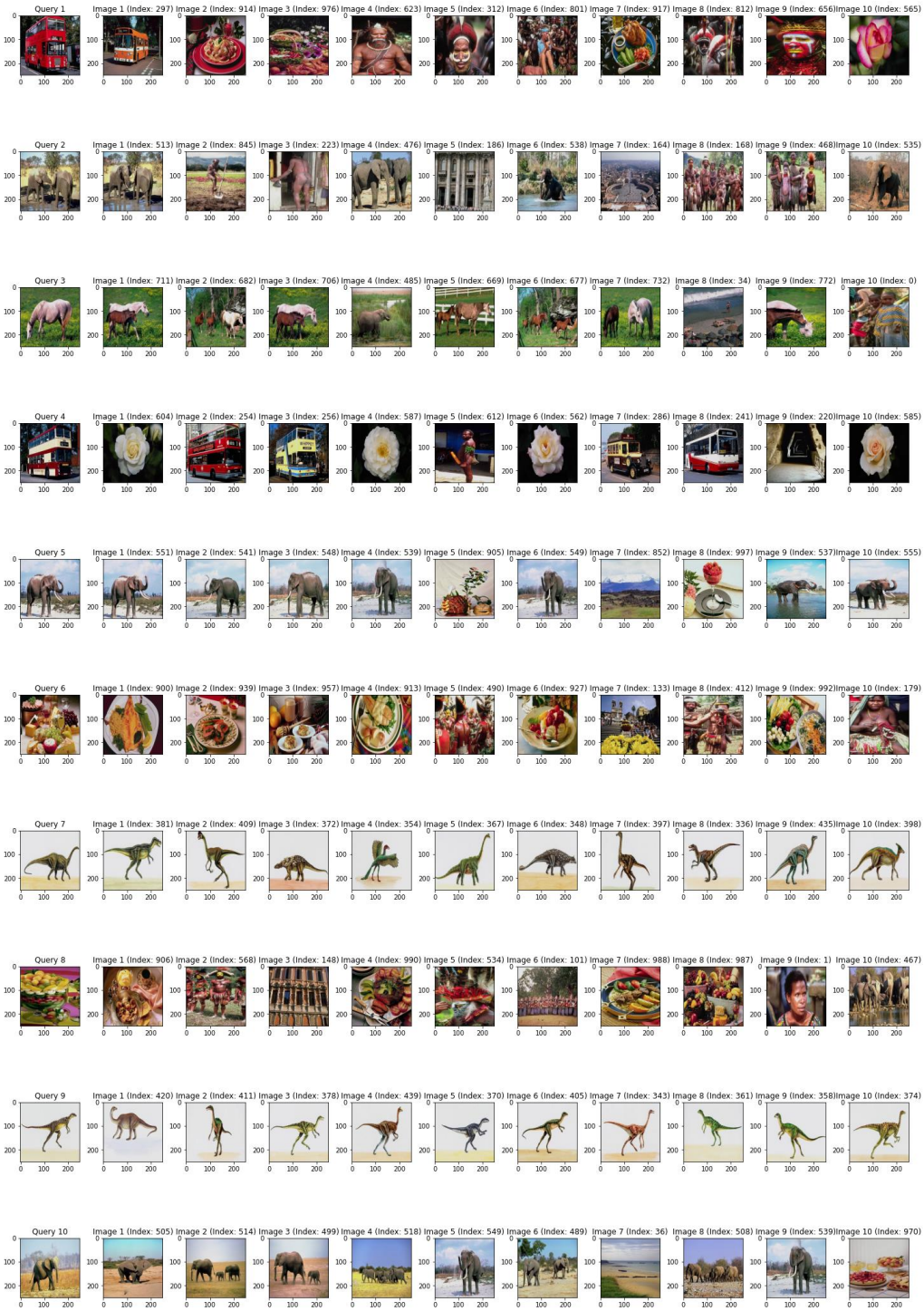
An even more pronounced weight adjustment was tested by assigning fifteen times the weight to skewness (with weights [1, 1, 15] * 3). Despite this significant emphasis, the AUC persisted at 0.67, while the F1 score dipped slightly to 0.1381.

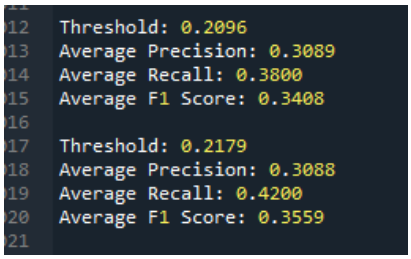
These experiments indicate that while skewness is an aspect of color distribution in images, its influence on the retrieval performance of the CBIR system is limited. Even with increased weighting, the system's ability to discriminate between relevant and non-relevant images, as measured by the AUC, remained constant, and the precision-recall balance, reflected by the F1 score, showed a slight overall decline. The results suggest there is a threshold beyond which additional emphasis on color symmetry does not translate into improved performance, highlighting the necessity for a balanced approach to feature weighting in image retrieval systems.

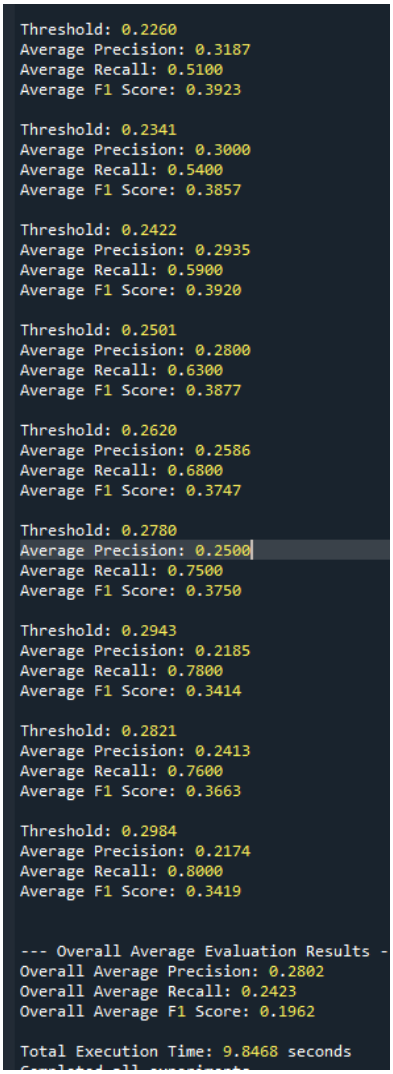
4. *Custom Scenario*

- **weights = [2, 8, 2] *3**, giving more weight for standard deviation since it gave the best improvement.

<u>Auc was 0.67 , 0.1962 , at some thresholds it reached 0.34 and 0.42</u>

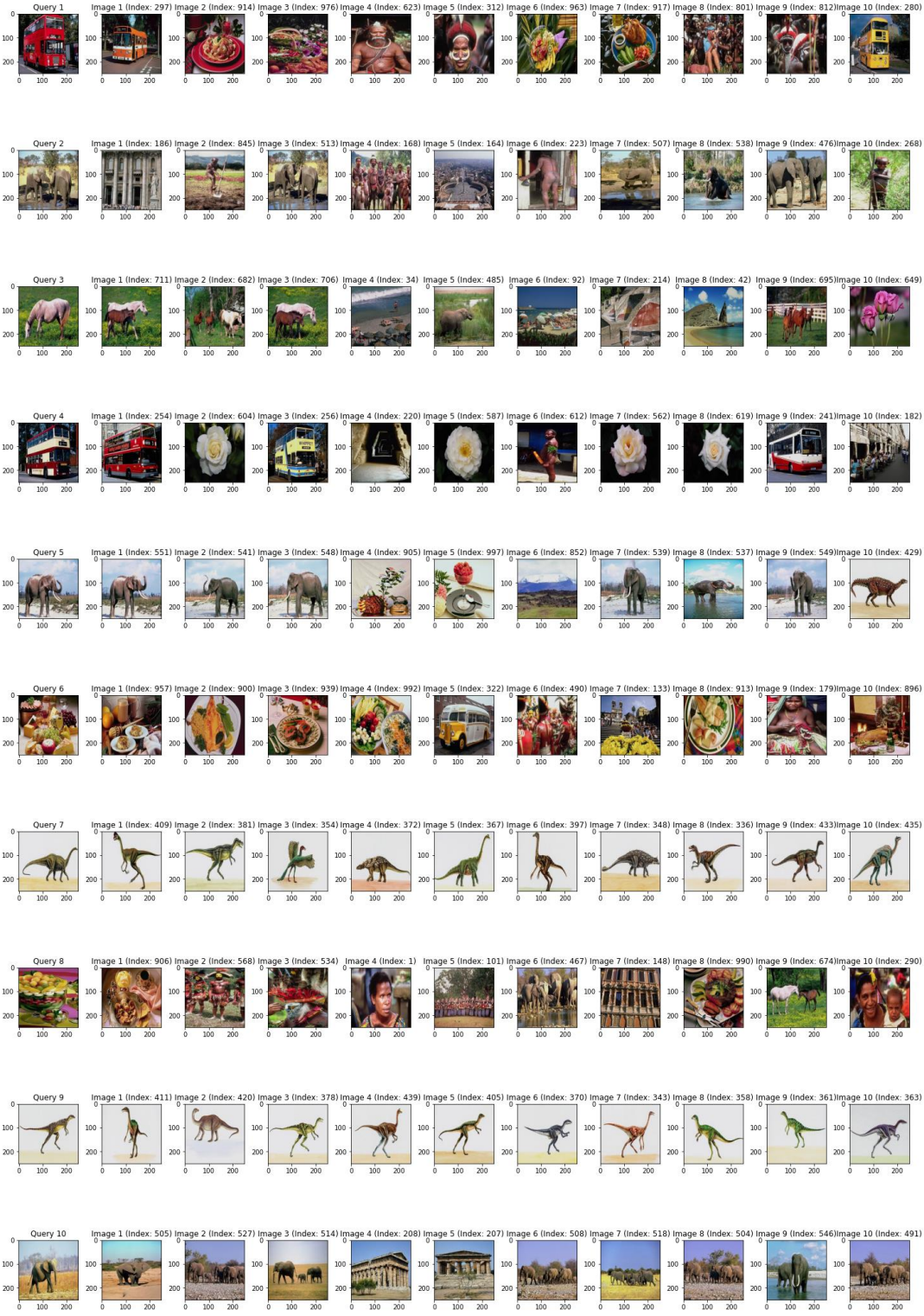






- **weights = [3, 20, 2]*3**

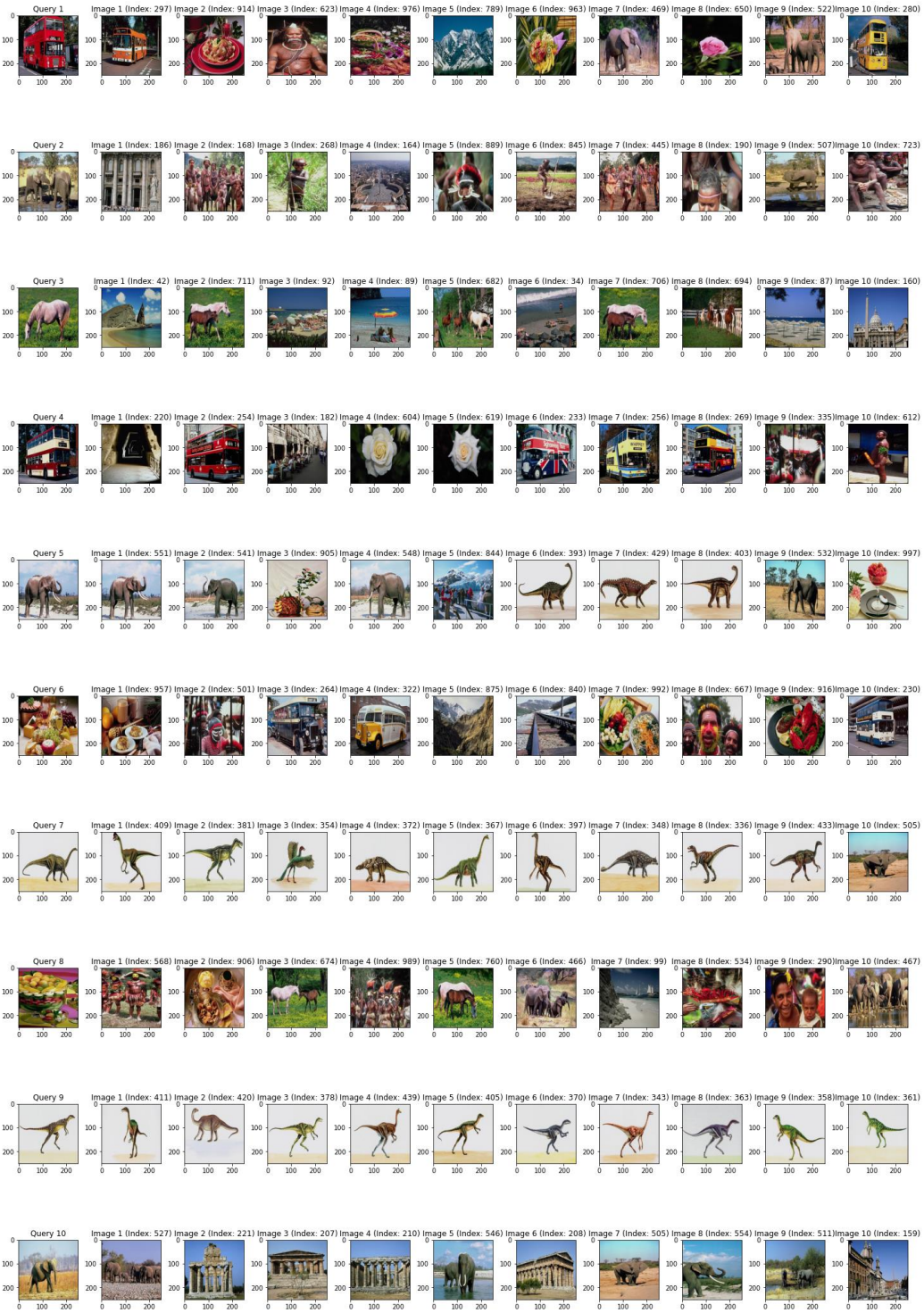
  Auc gave 0.66 , f1 score was 0.2103


Query 1 Image 1 (Index: 297) Image 2 (Index: 914) Image 3 (Index: 976) Image 4 (Index: 623) Image 5 (Index: 312) Image 6 (Index: 963) Image 7 (Index: 917) Image 8 (Index: 801) Image 9 (Index: 812) Image 10 (Index: 280)
Query 2 Image 1 (Index: 186) Image 2 (Index: 845) Image 3 (Index: 513) Image 4 (Index: 168) Image 5 (Index: 164) Image 6 (Index: 223) Image 7 (Index: 507) Image 8 (Index: 538) Image 9 (Index: 476) Image 10 (Index: 268)
Query 3 Image 1 (Index: 711) Image 2 (Index: 682) Image 3 (Index: 706) Image 4 (Index: 34) Image 5 (Index: 485) Image 6 (Index: 92) Image 7 (Index: 214) Image 8 (Index: 42) Image 9 (Index: 695) Image 10 (Index: 649)
Query 4 Image 1 (Index: 254) Image 2 (Index: 604) Image 3 (Index: 256) Image 4 (Index: 220) Image 5 (Index: 587) Image 6 (Index: 612) Image 7 (Index: 562) Image 8 (Index: 619) Image 9 (Index: 241) Image 10 (Index: 182)
Query 5 Image 1 (Index: 551) Image 2 (Index: 541) Image 3 (Index: 548) Image 4 (Index: 905) Image 5 (Index: 997) Image 6 (Index: 852) Image 7 (Index: 539) Image 8 (Index: 537) Image 9 (Index: 549) Image 10 (Index: 429)
Query 6 Image 1 (Index: 957) Image 2 (Index: 900) Image 3 (Index: 939) Image 4 (Index: 992) Image 5 (Index: 322) Image 6 (Index: 490) Image 7 (Index: 133) Image 8 (Index: 913) Image 9 (Index: 179) Image 10 (Index: 896)
Query 7 Image 1 (Index: 409) Image 2 (Index: 381) Image 3 (Index: 354) Image 4 (Index: 372) Image 5 (Index: 367) Image 6 (Index: 397) Image 7 (Index: 348) Image 8 (Index: 336) Image 9 (Index: 433) Image 10 (Index: 435)
Query 8 Image 1 (Index: 906) Image 2 (Index: 568) Image 3 (Index: 534) Image 4 (Index: 1) Image 5 (Index: 101) Image 6 (Index: 467) Image 7 (Index: 148) Image 8 (Index: 990) Image 9 (Index: 674) Image 10 (Index: 290)
Query 9 Image 1 (Index: 411) Image 2 (Index: 420) Image 3 (Index: 378) Image 4 (Index: 439) Image 5 (Index: 405) Image 6 (Index: 370) Image 7 (Index: 343) Image 8 (Index: 358) Image 9 (Index: 361) Image 10 (Index: 363)
Query 10 Image 1 (Index: 505) Image 2 (Index: 527) Image 3 (Index: 514) Image 4 (Index: 208) Image 5 (Index: 207) Image 6 (Index: 508) Image 7 (Index: 518) Image 8 (Index: 504) Image 9 (Index: 546) Image 10 (Index: 491)

```
Threshold: 0.2133
Average Precision: 0.2605
Average Recall: 0.5600
Average F1 Score: 0.3556

Threshold: 0.2212
Average Precision: 0.2625
Average Recall: 0.6300
Average F1 Score: 0.3706

Threshold: 0.2291
Average Precision: 0.2538
Average Recall: 0.6600
Average F1 Score: 0.3667

Threshold: 0.2368
Average Precision: 0.2464
Average Recall: 0.6900
Average F1 Score: 0.3632

Threshold: 0.2483
Average Precision: 0.2271
Average Recall: 0.7200
Average F1 Score: 0.3453

Threshold: 0.2636
Average Precision: 0.2094
Average Recall: 0.7600
Average F1 Score: 0.3283

Threshold: 0.2785
Average Precision: 0.1971
Average Recall: 0.8100
Average F1 Score: 0.3170

Threshold: 0.2673
Average Precision: 0.2042
Average Recall: 0.7700
Average F1 Score: 0.3229

Threshold: 0.2823
Average Precision: 0.1920
Average Recall: 0.8200
Average F1 Score: 0.3112


--- Overall Average Evaluation Results ---
Overall Average Precision: 0.2353
Overall Average Recall: 0.2880
Overall Average F1 Score: 0.2103

Total Execution Time: 10.5262 seconds
Completed all experiments
```

- **weights = [3, 50, 2]*3 , increasing more!**

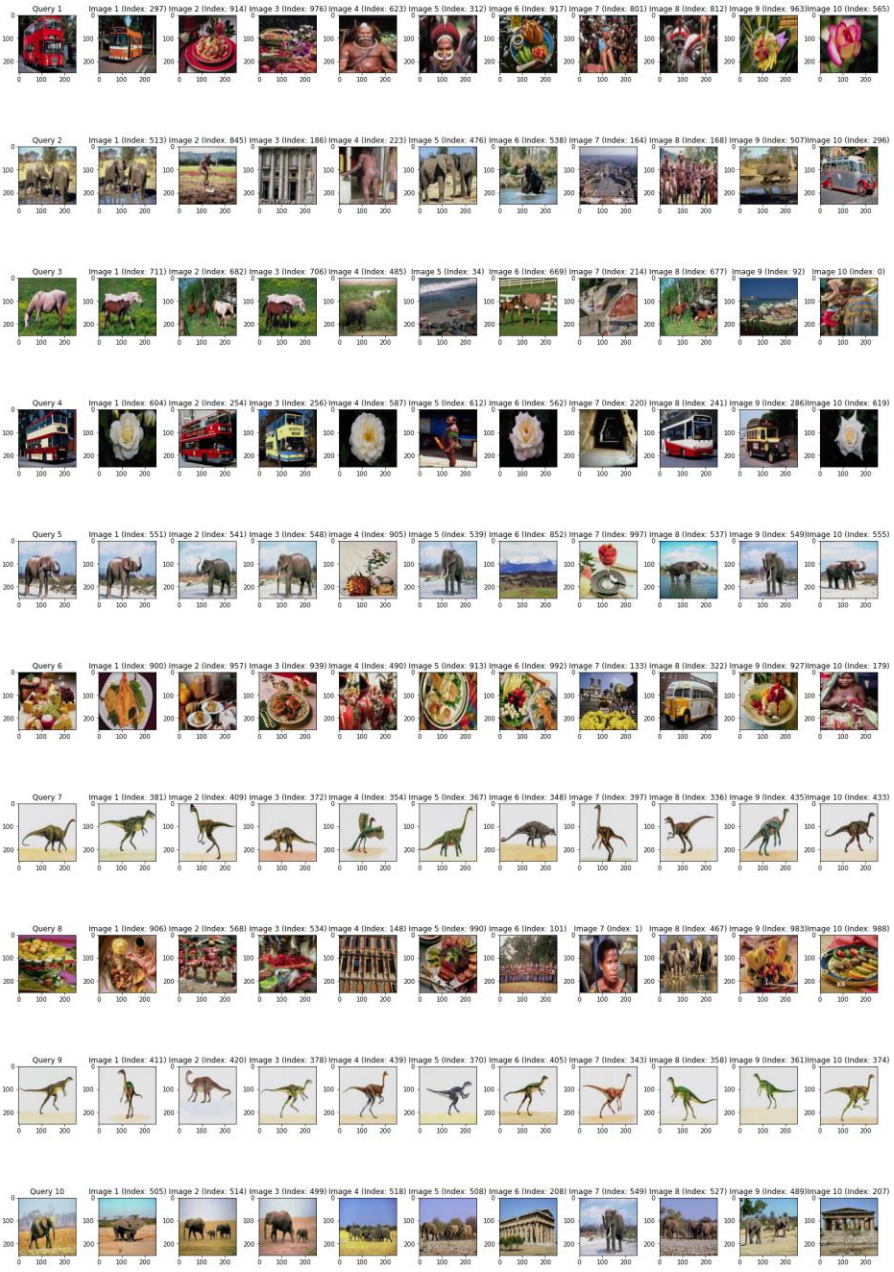
<u>Auc was 62 , f1 score was 0.2107</u>



```
--- Overall Average Evaluation Results ---
Overall Average Precision: 0.1970
Overall Average Recall: 0.3000
Overall Average F1 Score: 0.2107

Total Execution Time: 10.2847 seconds
```

- **weights = [10, 50, 20]*3**

<u>Auc was 67, f1 score was 0.1999</u>



```
Threshold: 0.2217
Average Precision: 0.2865
Average Recall: 0.5300
Average F1 Score: 0.3719

Threshold: 0.2297
Average Precision: 0.2823
Average Recall: 0.5900
Average F1 Score: 0.3819

Threshold: 0.2378
Average Precision: 0.2826
Average Recall: 0.6500
Average F1 Score: 0.3939

Threshold: 0.2458
Average Precision: 0.2588
Average Recall: 0.6600
Average F1 Score: 0.3718

Threshold: 0.2578
Average Precision: 0.2509
Average Recall: 0.7200
Average F1 Score: 0.3721

Threshold: 0.2736
Average Precision: 0.2269
Average Recall: 0.7600
Average F1 Score: 0.3494

Threshold: 0.2896
Average Precision: 0.2109
Average Recall: 0.8100
Average F1 Score: 0.3347

Threshold: 0.2776
Average Precision: 0.2197
Average Recall: 0.7600
Average F1 Score: 0.3408

Threshold: 0.2936
Average Precision: 0.2071
Average Recall: 0.8200
Average F1 Score: 0.3306


--- Overall Average Evaluation Results ---
Overall Average Precision: 0.2560
Overall Average Recall: 0.2633
Overall Average F1 Score: 0.1999

Total Execution Time: 9.9895 seconds
```

Results comparing:

In the first custom scenario, weights were adjusted to [2, 8, 2] * 3, giving more weight to the standard deviation, which previously showed promising improvement. The AUC remained steady

at 0.67, but the F1 score improved to 0.1962, indicating a better balance between precision and recall. Notably, at certain thresholds, the F1 score reached higher peaks of 0.34 and 0.42, suggesting that at specific operational points, the system demonstrated considerably improved performance.

The next configuration, with weights [3, 20, 2] * 3, yielded a slight decrease in the AUC to 0.66 but an enhanced F1 score of 0.2103. This further increase in the weight of standard deviation continued to show an overall positive trend in retrieval quality as assessed by the F1 score.

Pushing the standard deviation weight even more to [3, 50, 2] * 3, the AUC dropped to 0.62, but the F1 score saw a marginal improvement to 0.2107. This suggests that an extreme emphasis on standard deviation may start to adversely affect the system's discriminatory power (as AUC decreases) while still marginally improving the F1 score.

The final scenario tested a significant alteration in the weights to [10, 50, 20] * 3. Here, the AUC returned to 0.67, and the F1 score was 0.1999. This indicates that a substantial increase in the weights of all color channels provided a balance that brought the AUC back to the level of the baseline, with an improved F1 score, although not the highest observed.

conclusion of this part:

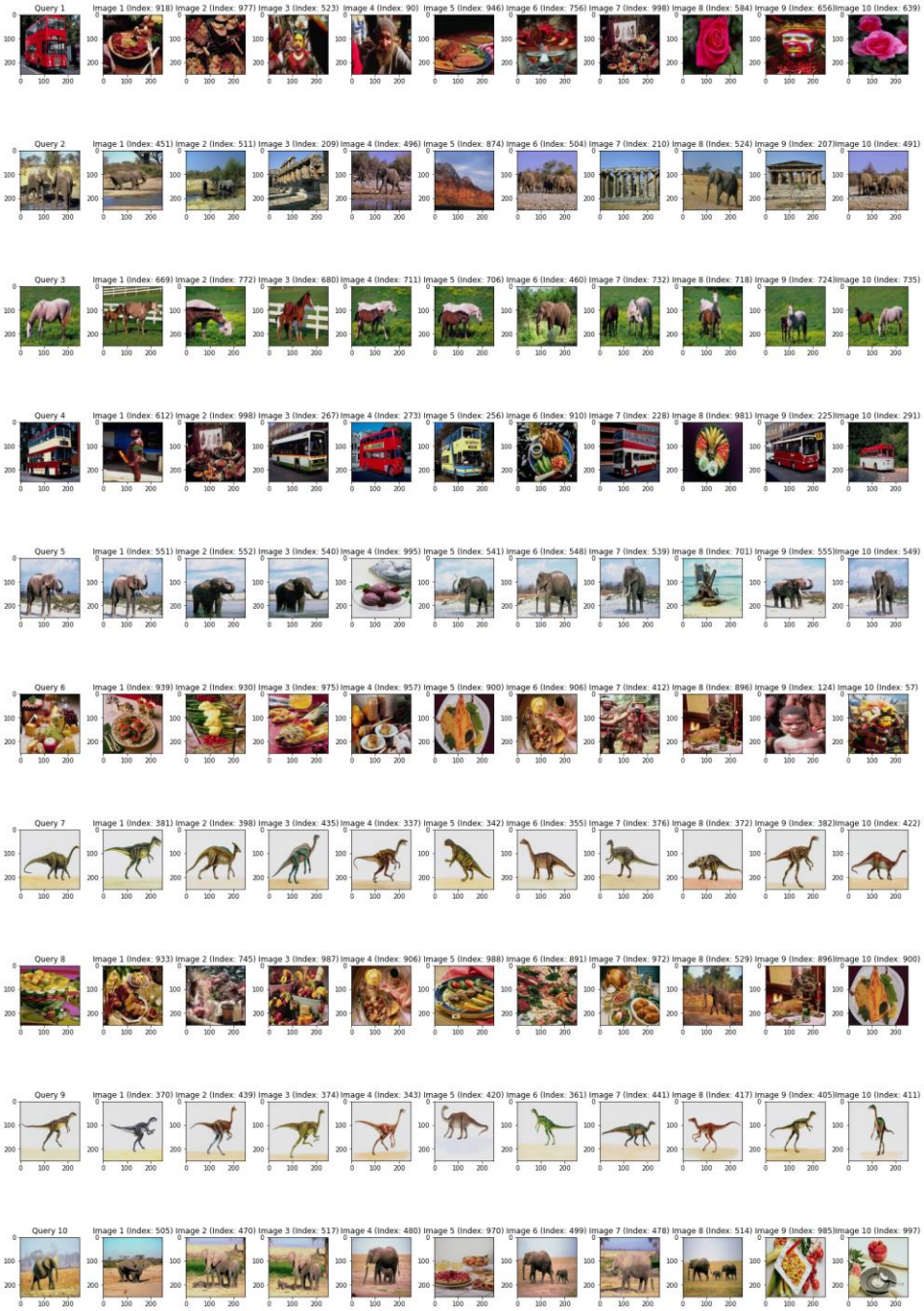
These scenarios reveal a complex relationship between the weights of the color channels and the performance of the CBIR system. While certain weight distributions can enhance the F1 score, suggesting an improved precision-recall balance, they may not necessarily increase the AUC. Also, while standard deviation seems to have a positive impact on the precision and recall balance of the system, it does not unilaterally improve all aspects of the system's performance. This suggests that while standard deviation is an important feature, the system also relies on a balanced weighting of other features to perform optimally across all metrics.

Subsection 3.2.3: Additional Moments

1) Balanced Approach

- Weights: **[1, 1, 1, 1, 1, 1] * 3** (Equal weight for all color moments)

<u>Auc was 0.68 , f1 score 0.2327</u>



```
At Threshold: 0.3526, Precision: 0.15, Recall: 0.84, F1 Score: 0.2
Average values for all thresholds for Query Index: 979
avg-Precision: 0.29
avg-Recall: 0.31
avg-F1 Score: 0.23
```

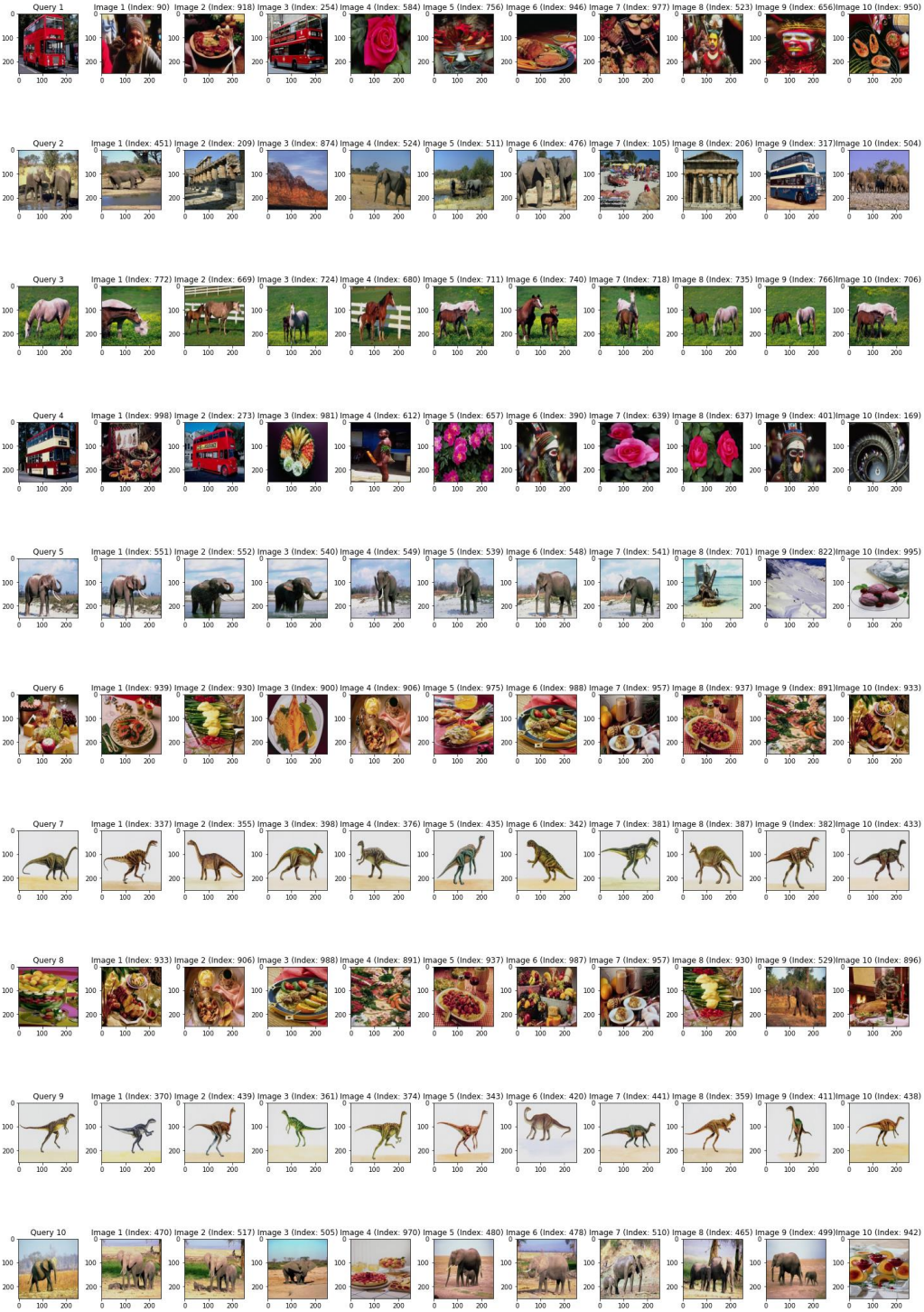
```
Average values for all thresholds for Query Index: 373
avg-Precision: 0.32
avg-Recall: 0.34
avg-F1 Score: 0.27
Query Index: 472
```

```
Threshold: 0.2443
Average Precision: 0.2933
Average Recall: 0.6100
Average F1 Score: 0.3961

Threshold: 0.2543
Average Precision: 0.2889
Average Recall: 0.6500
Average F1 Score: 0.4000

Threshold: 0.2643
Average Precision: 0.2936
Average Recall: 0.6900
Average F1 Score: 0.4119

Threshold: 0.2742
Average Precision: 0.2893
Average Recall: 0.7000
Average F1 Score: 0.4094

Threshold: 0.2891
Average Precision: 0.2812
Average Recall: 0.7200
Average F1 Score: 0.4045

Threshold: 0.3086
Average Precision: 0.2637
Average Recall: 0.7200
Average F1 Score: 0.3861

Threshold: 0.3279
Average Precision: 0.2559
Average Recall: 0.7600
Average F1 Score: 0.3829

Threshold: 0.3136
Average Precision: 0.2616
Average Recall: 0.7300
Average F1 Score: 0.3852

Threshold: 0.3326
Average Precision: 0.2476
Average Recall: 0.7600
Average F1 Score: 0.3735


--- Overall Average Evaluation Results ---
Overall Average Precision: 0.3132
Overall Average Recall: 0.2765
Overall Average F1 Score: 0.2327

Total Execution Time: 10.7751 seconds
Completed all experiments
```

2) *Emphasizing Median Color Value*

- Weights: **[1, 1, 1, 3, 1, 1] * 3** (3 times the weight for the median of each color channel)

Auc was 0.66, f1 score 0.1850



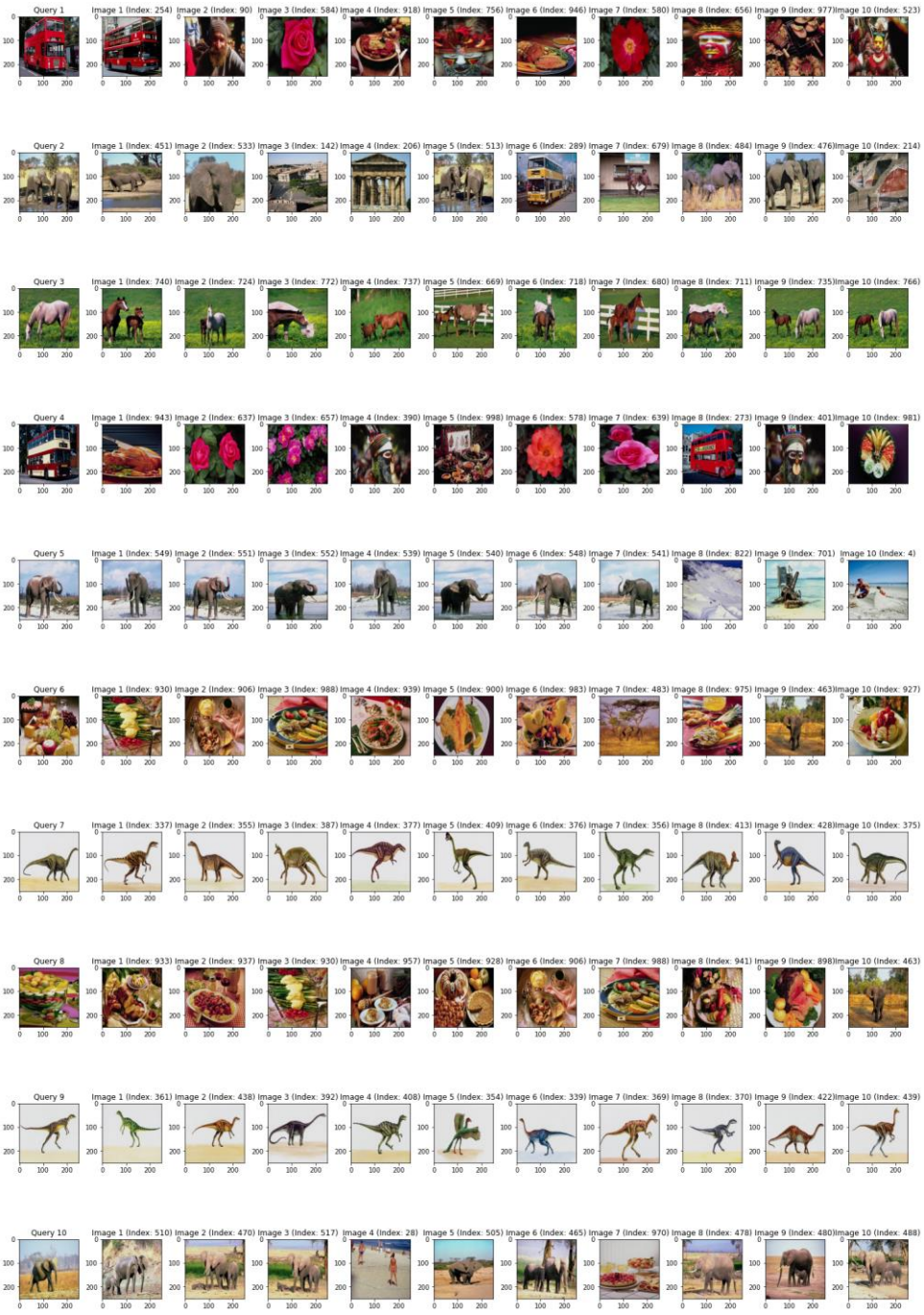
```
Threshold: 0.1981
Average Precision: 0.2660
Average Recall: 0.2500
Average F1 Score: 0.2577

Threshold: 0.2049
Average Precision: 0.2475
Average Recall: 0.2500
Average F1 Score: 0.2488

Threshold: 0.2117
Average Precision: 0.2455
Average Recall: 0.2700
Average F1 Score: 0.2571

Threshold: 0.2218
Average Precision: 0.2377
Average Recall: 0.2900
Average F1 Score: 0.2613

Threshold: 0.2356
Average Precision: 0.2083
Average Recall: 0.3000
Average F1 Score: 0.2459

Threshold: 0.2497
Average Precision: 0.1902
Average Recall: 0.3100
Average F1 Score: 0.2357

Threshold: 0.2391
Average Precision: 0.2081
Average Recall: 0.3100
Average F1 Score: 0.2490

Threshold: 0.2532
Average Precision: 0.1845
Average Recall: 0.3100
Average F1 Score: 0.2313


--- Overall Average Evaluation Results ---
Overall Average Precision: 0.3408
Overall Average Recall: 0.1505
Overall Average F1 Score: 0.1850

Total Execution Time: 9.5296 seconds
Completed all experiments
```

```
Average values for all thresholds for Query Index: 373
avg-Precision: 0.34
avg-Recall: 0.33
avg-F1 Score: 0.28
```

```
Average values for all thresholds for Query Index: 979
avg-Precision: 0.31
avg-Recall: 0.30
avg-F1 Score: 0.24
```

- Weights: **[1, 1, 1, 10, 1, 1] * 3** (3 times the weight for the median of each color channel)

Auc was 0.65, f1 score 0.1643



```
Threshold: 0.1545
Average Precision: 0.3137
Average Recall: 0.1600
Average F1 Score: 0.2119

Threshold: 0.1612
Average Precision: 0.3220
Average Recall: 0.1900
Average F1 Score: 0.2390

Threshold: 0.1679
Average Precision: 0.2923
Average Recall: 0.1900
Average F1 Score: 0.2303

Threshold: 0.1747
Average Precision: 0.2632
Average Recall: 0.2000
Average F1 Score: 0.2273

Threshold: 0.1850
Average Precision: 0.2553
Average Recall: 0.2400
Average F1 Score: 0.2474

Threshold: 0.1987
Average Precision: 0.2250
Average Recall: 0.2700
Average F1 Score: 0.2455

Threshold: 0.2132
Average Precision: 0.2101
Average Recall: 0.2900
Average F1 Score: 0.2437

Threshold: 0.2023
Average Precision: 0.2295
Average Recall: 0.2800
Average F1 Score: 0.2523

Threshold: 0.2169
Average Precision: 0.2083
Average Recall: 0.3000
Average F1 Score: 0.2459


--- Overall Average Evaluation Results ---
Overall Average Precision: 0.3812
Overall Average Recall: 0.1217
Overall Average F1 Score: 0.1643

Total Execution Time: 10.1377 seconds
```

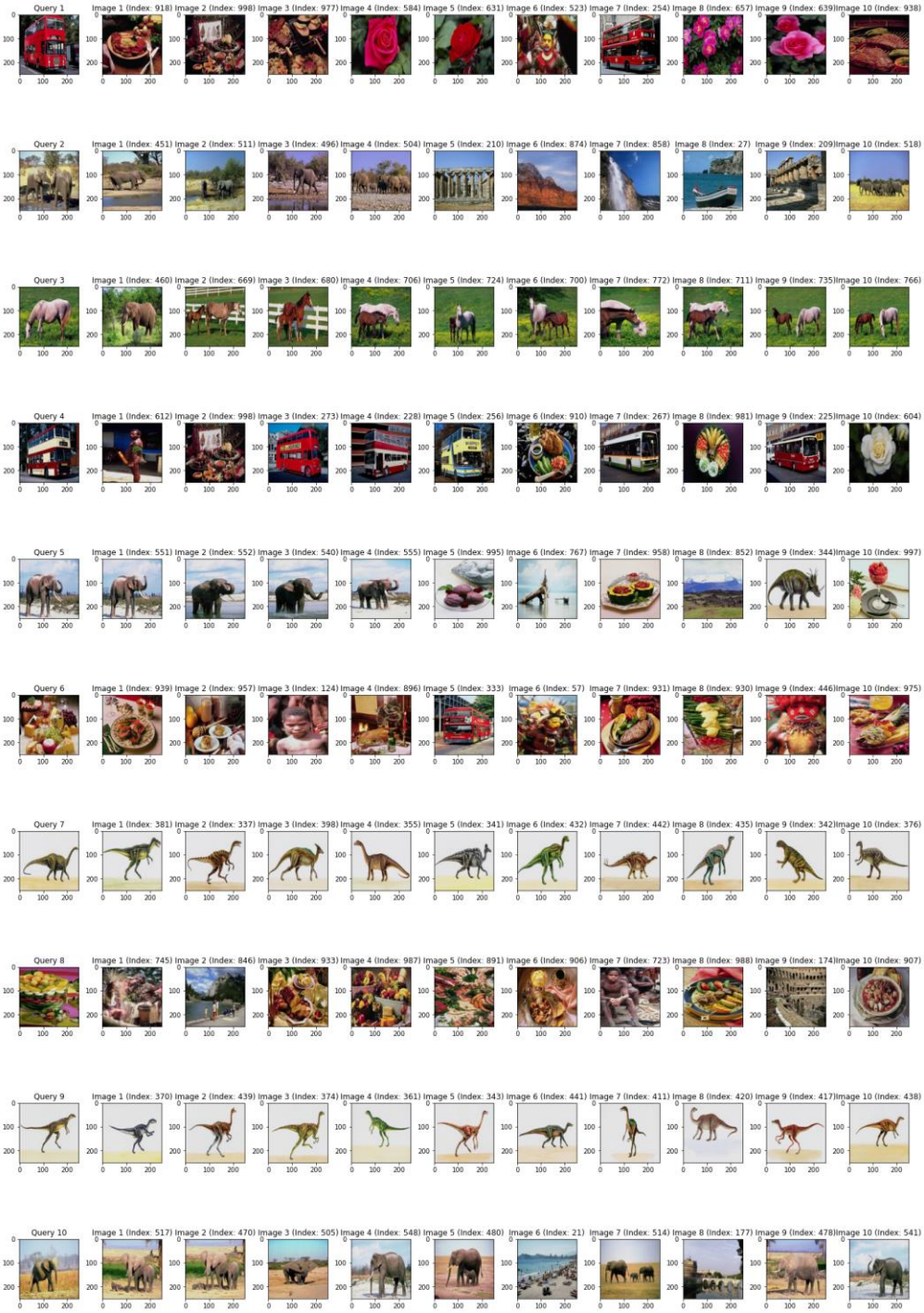
```
Average values for all thresholds for Query Index: 373
avg-Precision: 0.37
avg-Recall: 0.32
avg-F1 Score: 0.28
```

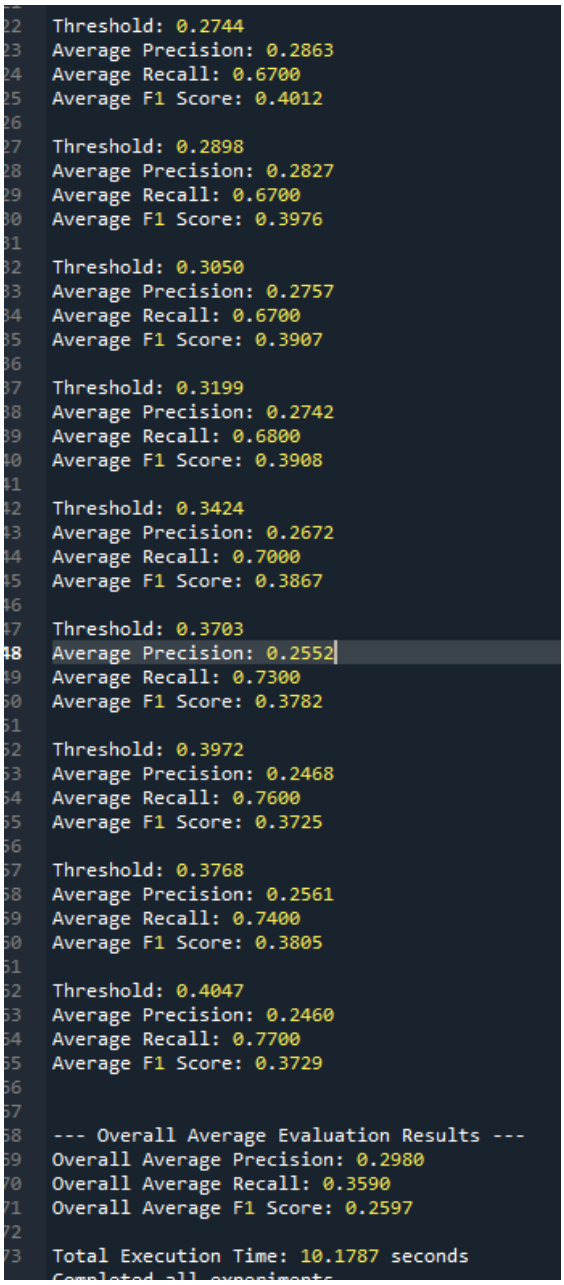
```
At Threshold: 0.2169, Precision: 0.14, Recall: 0.70, F1 Score:
Average values for all thresholds for Query Index: 979
avg-Precision: 0.35
avg-Recall: 0.29
avg-F1 Score: 0.24
```

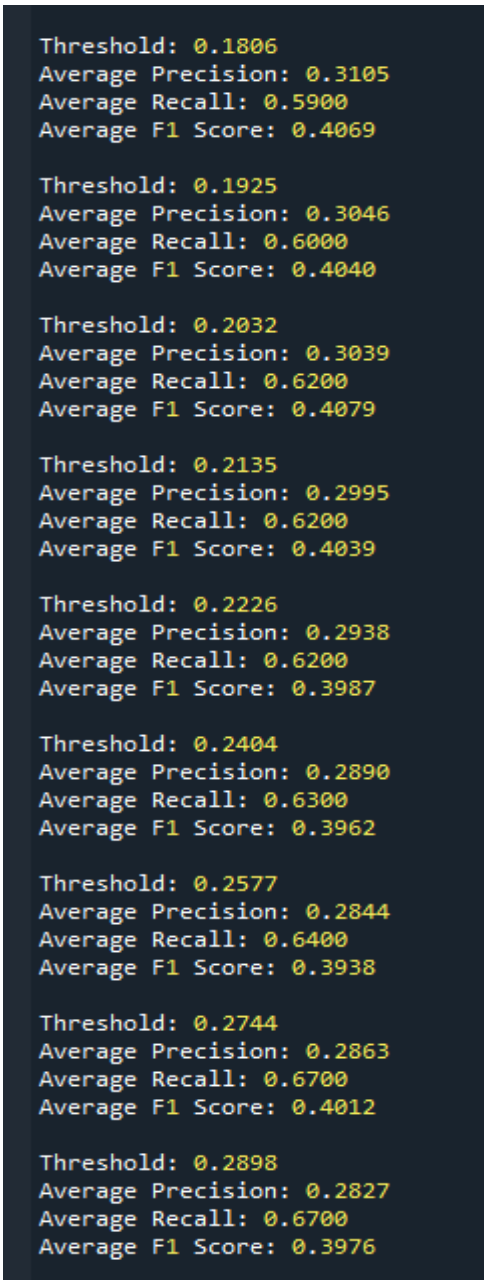
3) Emphasizing Most Common Color (Mode)

- Weights: **[1, 1, 1, 1, 3, 1] * 3** (3 times the weight for the mode of each color channel)

Auc was 67, f1 score 0.2597

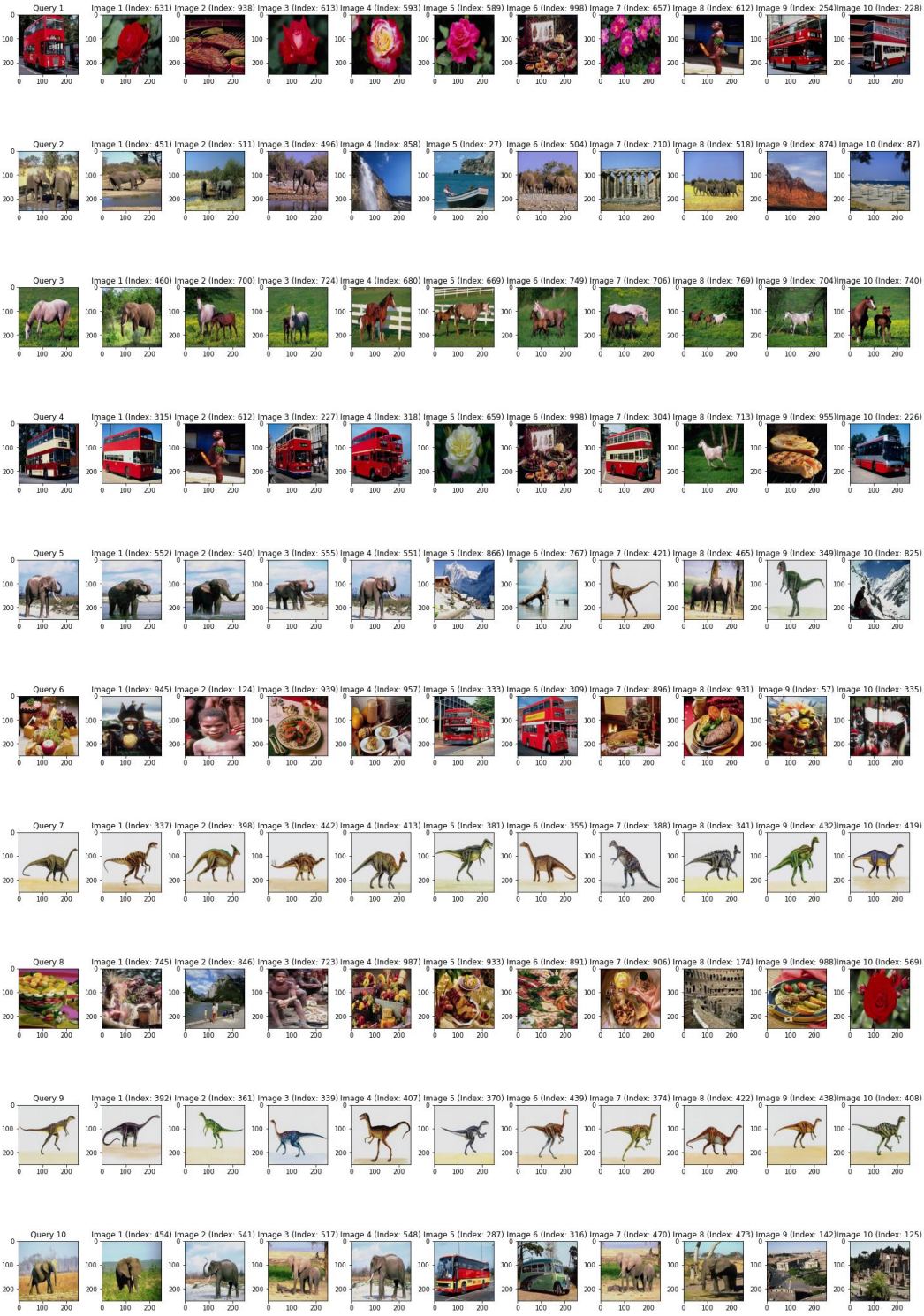


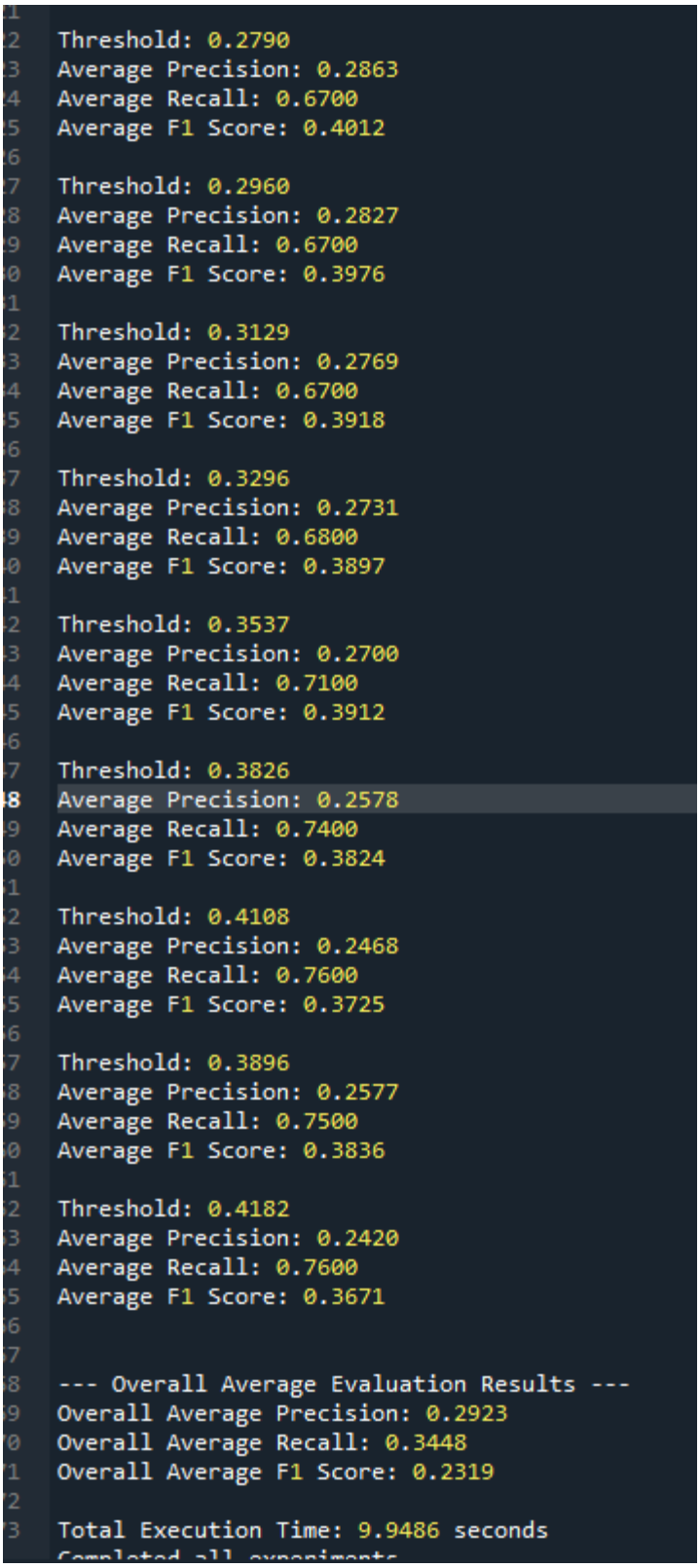

Threshold: 0.2744
Average Precision: 0.2863
Average Recall: 0.6700
Average F1 Score: 0.4012
Threshold: 0.2898
Average Precision: 0.2827
Average Recall: 0.6700
Average F1 Score: 0.3976
Threshold: 0.3050
Average Precision: 0.2757
Average Recall: 0.6700
Average F1 Score: 0.3907
Threshold: 0.3199
Average Precision: 0.2742
Average Recall: 0.6800
Average F1 Score: 0.3908
Threshold: 0.3424
Average Precision: 0.2672
Average Recall: 0.7000
Average F1 Score: 0.3867
Threshold: 0.3703
Average Precision: 0.2552
Average Recall: 0.7300
Average F1 Score: 0.3782
Threshold: 0.3972
Average Precision: 0.2468
Average Recall: 0.7600
Average F1 Score: 0.3725
Threshold: 0.3768
Average Precision: 0.2561
Average Recall: 0.7400
Average F1 Score: 0.3805
Threshold: 0.4047
Average Precision: 0.2460
Average Recall: 0.7700
Average F1 Score: 0.3729
--- Overall Average Evaluation Results ---
Overall Average Precision: 0.2980
Overall Average Recall: 0.3590
Overall Average F1 Score: 0.2597
Total Execution Time: 10.1787 seconds


Threshold: 0.1806
Average Precision: 0.3105
Average Recall: 0.5900
Average F1 Score: 0.4069
Threshold: 0.1925
Average Precision: 0.3046
Average Recall: 0.6000
Average F1 Score: 0.4040
Threshold: 0.2032
Average Precision: 0.3039
Average Recall: 0.6200
Average F1 Score: 0.4079
Threshold: 0.2135
Average Precision: 0.2995
Average Recall: 0.6200
Average F1 Score: 0.4039
Threshold: 0.2226
Average Precision: 0.2938
Average Recall: 0.6200
Average F1 Score: 0.3987
Threshold: 0.2404
Average Precision: 0.2890
Average Recall: 0.6300
Average F1 Score: 0.3962
Threshold: 0.2577
Average Precision: 0.2844
Average Recall: 0.6400
Average F1 Score: 0.3938
Threshold: 0.2744
Average Precision: 0.2863
Average Recall: 0.6700
Average F1 Score: 0.4012
Threshold: 0.2898
Average Precision: 0.2827
Average Recall: 0.6700
Average F1 Score: 0.3976

- Weights: **[1, 1, 1, 1, 15, 1] * 3** (3 times the weight for the mode of each color channel)

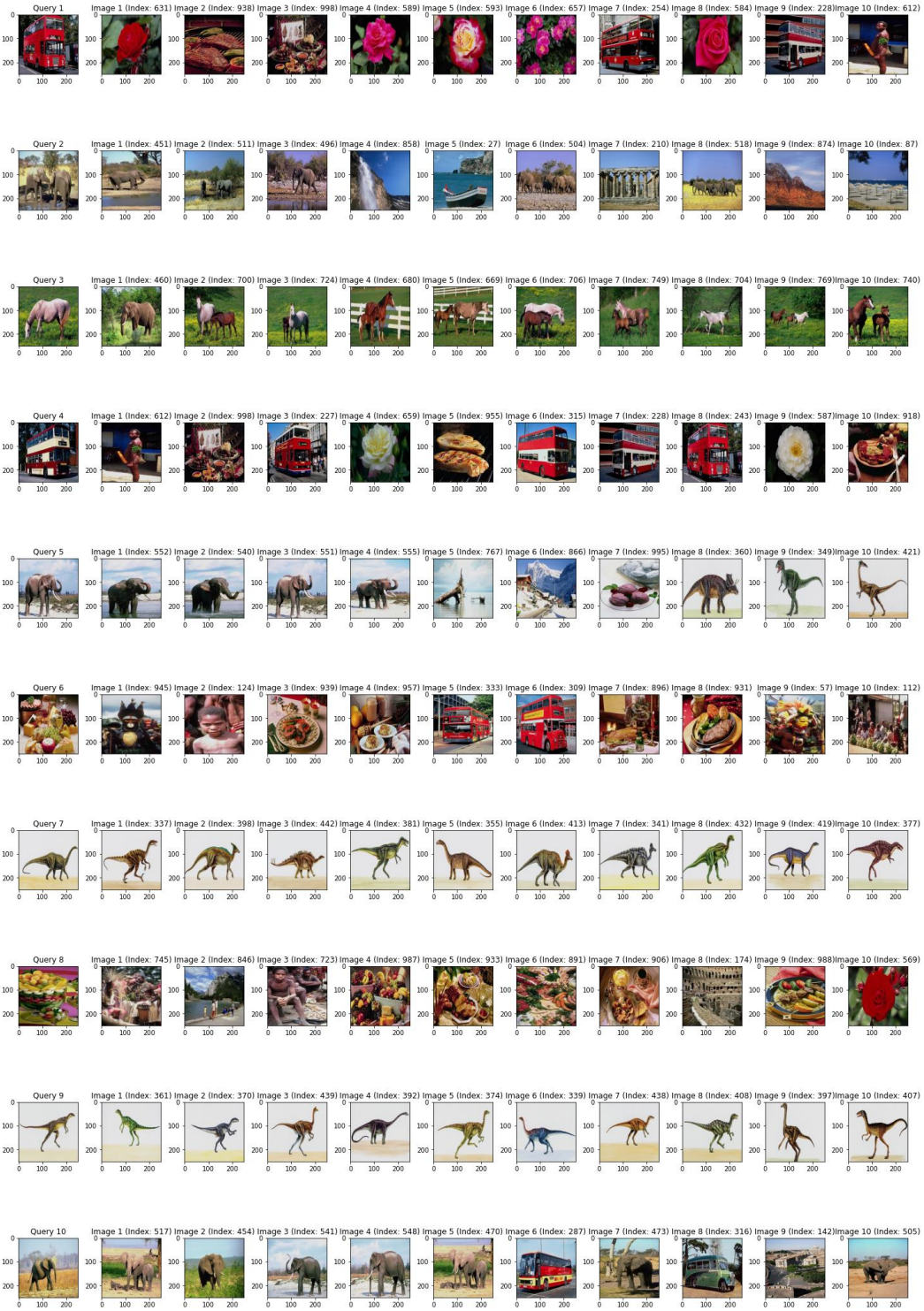
Auc was 66, f1 score 0.2319




Threshold: 0.2790
Average Precision: 0.2863
Average Recall: 0.6700
Average F1 Score: 0.4012
Threshold: 0.2960
Average Precision: 0.2827
Average Recall: 0.6700
Average F1 Score: 0.3976
Threshold: 0.3129
Average Precision: 0.2769
Average Recall: 0.6700
Average F1 Score: 0.3918
Threshold: 0.3296
Average Precision: 0.2731
Average Recall: 0.6800
Average F1 Score: 0.3897
Threshold: 0.3537
Average Precision: 0.2700
Average Recall: 0.7100
Average F1 Score: 0.3912
Threshold: 0.3826
Average Precision: 0.2578
Average Recall: 0.7400
Average F1 Score: 0.3824
Threshold: 0.4108
Average Precision: 0.2468
Average Recall: 0.7600
Average F1 Score: 0.3725
Threshold: 0.3896
Average Precision: 0.2577
Average Recall: 0.7500
Average F1 Score: 0.3836
Threshold: 0.4182
Average Precision: 0.2420
Average Recall: 0.7600
Average F1 Score: 0.3671
--- Overall Average Evaluation Results ---
Overall Average Precision: 0.2923
Overall Average Recall: 0.3448
Overall Average F1 Score: 0.2319
Total Execution Time: 9.9486 seconds

- Weights: **[1, 1, 1, 1, 10, 1] * 3** (3 times the weight for the mode of each color channel)

Auc was 66, f1 score 0.2331

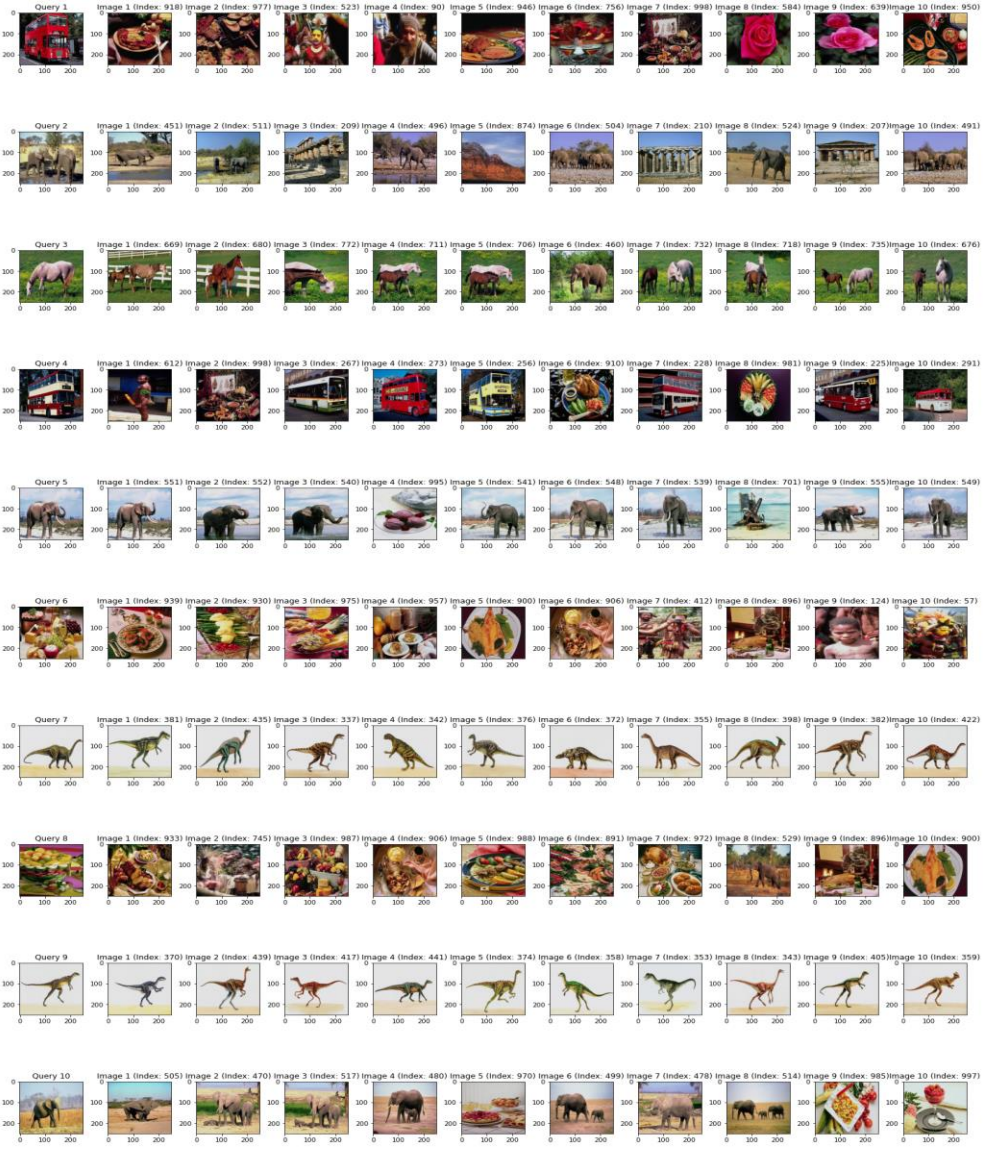


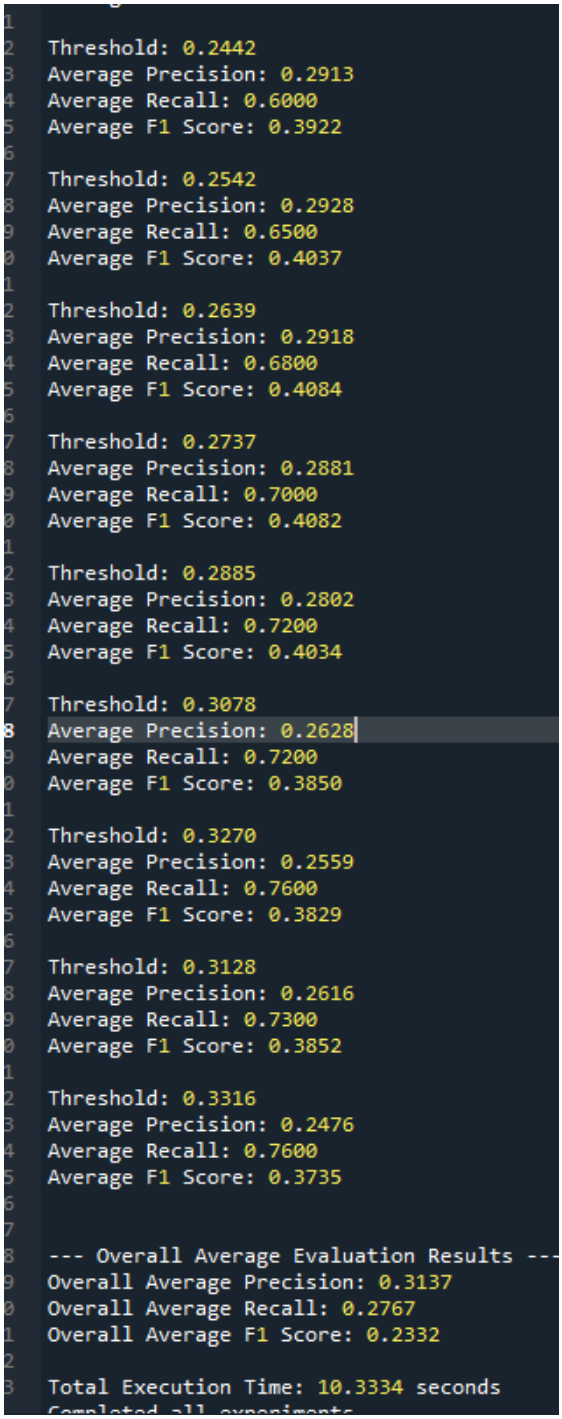
```
--- Overall Average Evaluation Results ---
Overall Average Precision: 0.2838
Overall Average Recall: 0.3435
Overall Average F1 Score: 0.2331

Total Execution Time: 9.9612 seconds
```

4) Emphasizing Color Distribution Shape (Kurtosis)

- Weights: **[1, 1, 1, 1, 1, 3] * 3** (3 times the weight for the kurtosis of each color channel)

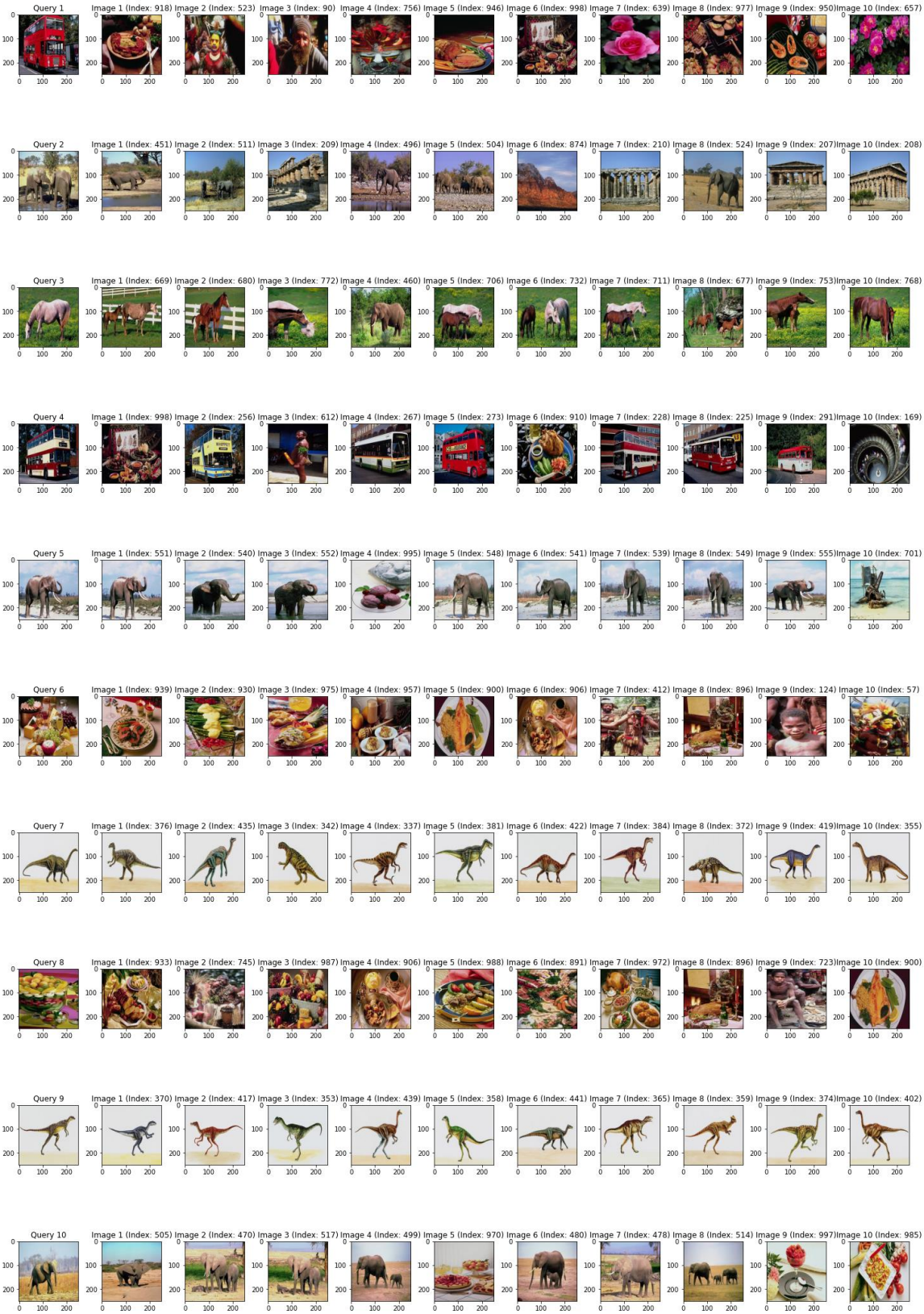
AUC was 0.68 , f1 score 0.2332

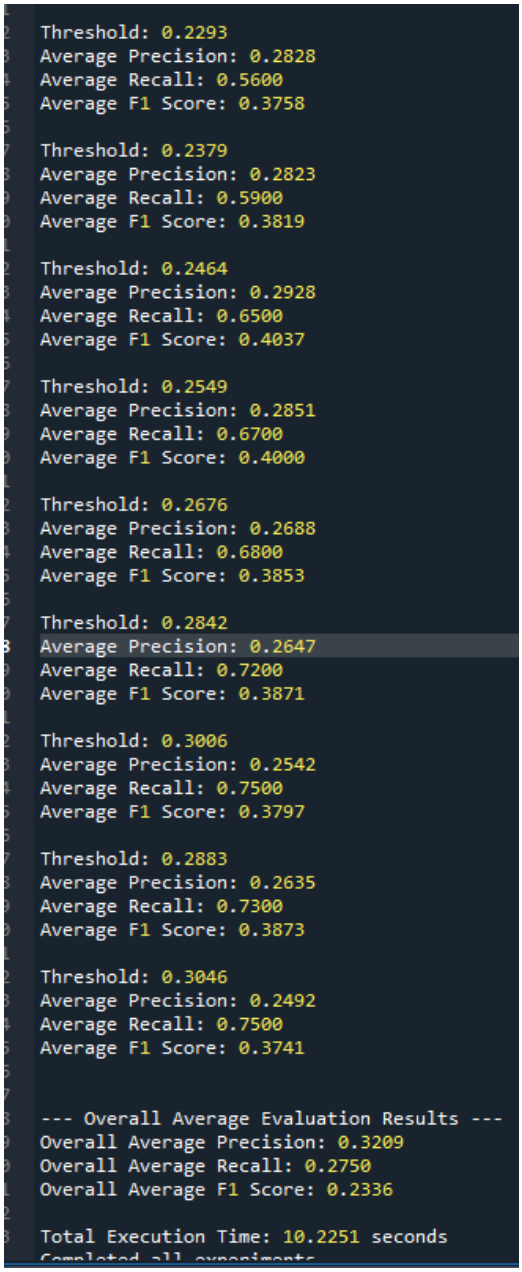




- Weights: **[1, 1, 1, 1, 1, 10] * 3** (3 times the weight for the kurtosis of each color channel)

AUC was 0.68 , f1 score 0.2336

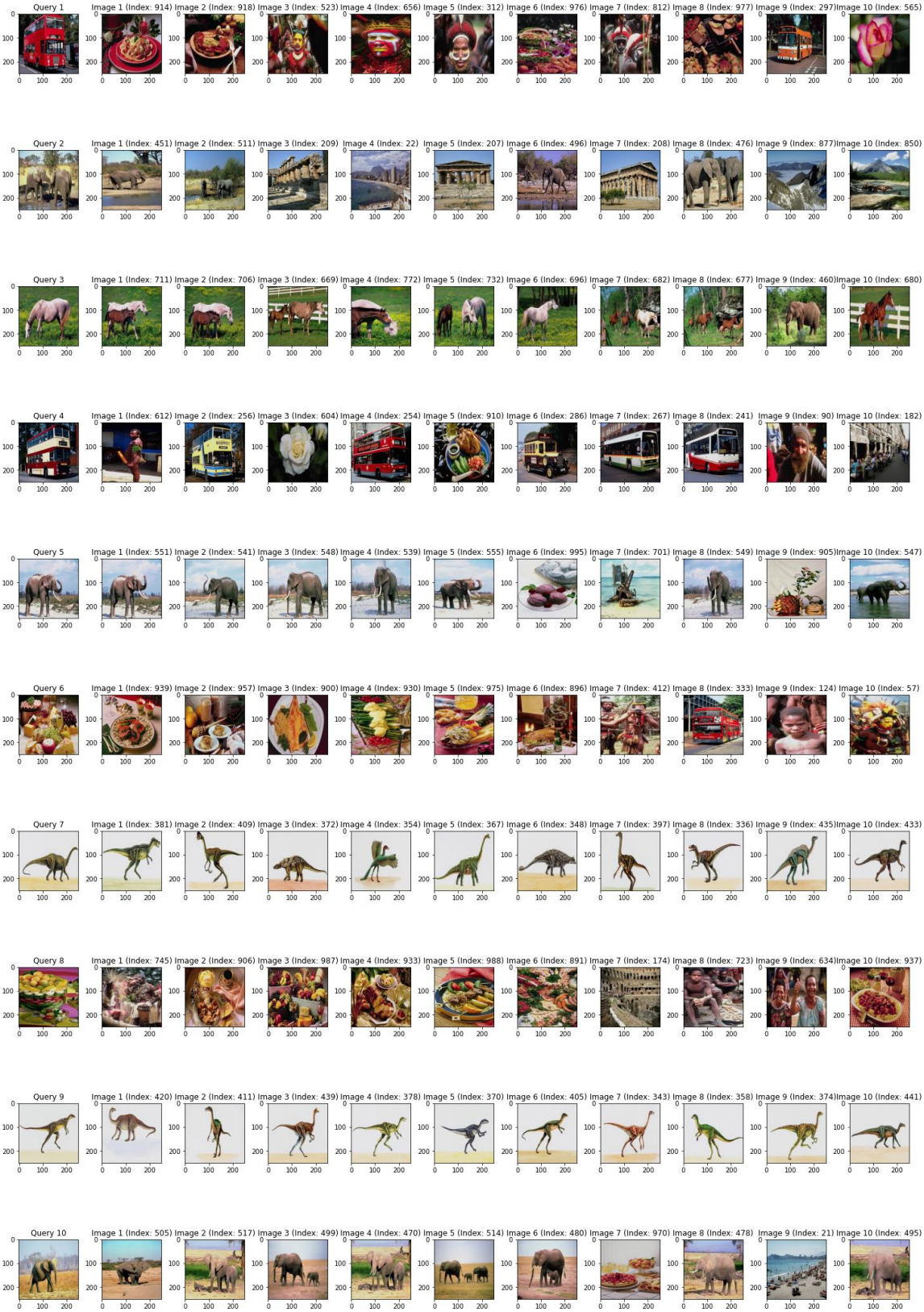

Query 1
Image 1 (Index: 918)
Image 2 (Index: 523)
Image 3 (Index: 90)
Image 4 (Index: 756)
Image 5 (Index: 946)
Image 6 (Index: 998)
Image 7 (Index: 639)
Image 8 (Index: 977)
Image 9 (Index: 950)
Image 10 (Index: 657)
Query 2
Image 1 (Index: 451)
Image 2 (Index: 511)
Image 3 (Index: 209)
Image 4 (Index: 496)
Image 5 (Index: 504)
Image 6 (Index: 874)
Image 7 (Index: 210)
Image 8 (Index: 524)
Image 9 (Index: 207)
Image 10 (Index: 208)
Query 3
Image 1 (Index: 669)
Image 2 (Index: 680)
Image 3 (Index: 772)
Image 4 (Index: 460)
Image 5 (Index: 706)
Image 6 (Index: 732)
Image 7 (Index: 711)
Image 8 (Index: 677)
Image 9 (Index: 753)
Image 10 (Index: 768)
Query 4
Image 1 (Index: 998)
Image 2 (Index: 256)
Image 3 (Index: 612)
Image 4 (Index: 267)
Image 5 (Index: 273)
Image 6 (Index: 910)
Image 7 (Index: 228)
Image 8 (Index: 225)
Image 9 (Index: 291)
Image 10 (Index: 169)
Query 5
Image 1 (Index: 551)
Image 2 (Index: 540)
Image 3 (Index: 552)
Image 4 (Index: 995)
Image 5 (Index: 548)
Image 6 (Index: 541)
Image 7 (Index: 539)
Image 8 (Index: 549)
Image 9 (Index: 555)
Image 10 (Index: 701)
Query 6
Image 1 (Index: 939)
Image 2 (Index: 930)
Image 3 (Index: 975)
Image 4 (Index: 957)
Image 5 (Index: 900)
Image 6 (Index: 906)
Image 7 (Index: 412)
Image 8 (Index: 896)
Image 9 (Index: 124)
Image 10 (Index: 57)
Query 7
Image 1 (Index: 376)
Image 2 (Index: 435)
Image 3 (Index: 342)
Image 4 (Index: 337)
Image 5 (Index: 381)
Image 6 (Index: 422)
Image 7 (Index: 384)
Image 8 (Index: 372)
Image 9 (Index: 419)
Image 10 (Index: 355)
Query 8
Image 1 (Index: 933)
Image 2 (Index: 745)
Image 3 (Index: 987)
Image 4 (Index: 906)
Image 5 (Index: 988)
Image 6 (Index: 891)
Image 7 (Index: 972)
Image 8 (Index: 896)
Image 9 (Index: 723)
Image 10 (Index: 900)
Query 9
Image 1 (Index: 370)
Image 2 (Index: 417)
Image 3 (Index: 353)
Image 4 (Index: 439)
Image 5 (Index: 358)
Image 6 (Index: 441)
Image 7 (Index: 365)
Image 8 (Index: 359)
Image 9 (Index: 374)
Image 10 (Index: 402)
Query 10
Image 1 (Index: 505)
Image 2 (Index: 470)
Image 3 (Index: 517)
Image 4 (Index: 499)
Image 5 (Index: 970)
Image 6 (Index: 480)
Image 7 (Index: 478)
Image 8 (Index: 514)
Image 9 (Index: 997)
Image 10 (Index: 985)



5) Additional scenario

- **weights=[2, 10, 2, 3, 3, 6]*3**

AUC was 0.69, f1 score 0.2897


Query 1
Image 1 (Index: 914)
Image 2 (Index: 918)
Image 3 (Index: 523)
Image 4 (Index: 656)
Image 5 (Index: 312)
Image 6 (Index: 976)
Image 7 (Index: 812)
Image 8 (Index: 977)
Image 9 (Index: 297)
Image 10 (Index: 565)
Query 2
Image 1 (Index: 451)
Image 2 (Index: 511)
Image 3 (Index: 209)
Image 4 (Index: 22)
Image 5 (Index: 207)
Image 6 (Index: 496)
Image 7 (Index: 208)
Image 8 (Index: 476)
Image 9 (Index: 877)
Image 10 (Index: 850)
Query 3
Image 1 (Index: 711)
Image 2 (Index: 706)
Image 3 (Index: 669)
Image 4 (Index: 772)
Image 5 (Index: 732)
Image 6 (Index: 696)
Image 7 (Index: 682)
Image 8 (Index: 677)
Image 9 (Index: 460)
Image 10 (Index: 680)
Query 4
Image 1 (Index: 612)
Image 2 (Index: 256)
Image 3 (Index: 604)
Image 4 (Index: 254)
Image 5 (Index: 910)
Image 6 (Index: 286)
Image 7 (Index: 267)
Image 8 (Index: 241)
Image 9 (Index: 90)
Image 10 (Index: 182)
Query 5
Image 1 (Index: 551)
Image 2 (Index: 541)
Image 3 (Index: 548)
Image 4 (Index: 539)
Image 5 (Index: 555)
Image 6 (Index: 995)
Image 7 (Index: 701)
Image 8 (Index: 549)
Image 9 (Index: 905)
Image 10 (Index: 547)
Query 6
Image 1 (Index: 939)
Image 2 (Index: 957)
Image 3 (Index: 900)
Image 4 (Index: 930)
Image 5 (Index: 975)
Image 6 (Index: 896)
Image 7 (Index: 412)
Image 8 (Index: 333)
Image 9 (Index: 124)
Image 10 (Index: 57)
Query 7
Image 1 (Index: 381)
Image 2 (Index: 409)
Image 3 (Index: 372)
Image 4 (Index: 354)
Image 5 (Index: 367)
Image 6 (Index: 348)
Image 7 (Index: 397)
Image 8 (Index: 336)
Image 9 (Index: 435)
Image 10 (Index: 433)
Query 8
Image 1 (Index: 745)
Image 2 (Index: 906)
Image 3 (Index: 987)
Image 4 (Index: 933)
Image 5 (Index: 988)
Image 6 (Index: 891)
Image 7 (Index: 174)
Image 8 (Index: 723)
Image 9 (Index: 634)
Image 10 (Index: 937)
Query 9
Image 1 (Index: 420)
Image 2 (Index: 411)
Image 3 (Index: 439)
Image 4 (Index: 378)
Image 5 (Index: 370)
Image 6 (Index: 405)
Image 7 (Index: 343)
Image 8 (Index: 358)
Image 9 (Index: 374)
Image 10 (Index: 441)
Query 10
Image 1 (Index: 505)
Image 2 (Index: 517)
Image 3 (Index: 499)
Image 4 (Index: 470)
Image 5 (Index: 514)
Image 6 (Index: 480)
Image 7 (Index: 970)
Image 8 (Index: 478)
Image 9 (Index: 21)
Image 10 (Index: 495)

```
--- Overall Average Evaluation Results ---
Overall Average Precision: 0.3268
Overall Average Recall: 0.3768
Overall Average F1 Score: 0.2897

Total Execution Time: 10.6207 seconds
```

```
Threshold: 0.2720
Average Precision: 0.2982
Average Recall: 0.6800
Average F1 Score: 0.4146

Threshold: 0.2819
Average Precision: 0.2975
Average Recall: 0.7200
Average F1 Score: 0.4211

Threshold: 0.2916
Average Precision: 0.2903
Average Recall: 0.7200
Average F1 Score: 0.4138

Threshold: 0.3013
Average Precision: 0.2780
Average Recall: 0.7200
Average F1 Score: 0.4011

Threshold: 0.3108
Average Precision: 0.2657
Average Recall: 0.7200
Average F1 Score: 0.3881

Threshold: 0.3247
Average Precision: 0.2598
Average Recall: 0.7300
Average F1 Score: 0.3832

Threshold: 0.3420
Average Precision: 0.2607
Average Recall: 0.7900
Average F1 Score: 0.3921

Threshold: 0.3584
Average Precision: 0.2455
Average Recall: 0.8100
Average F1 Score: 0.3767

Threshold: 0.3461
Average Precision: 0.2581
Average Recall: 0.8000
Average F1 Score: 0.3902

Threshold: 0.3626
Average Precision: 0.2426
Average Recall: 0.8200
Average F1 Score: 0.3744
```

```
At Threshold: 0.3626, Precision: 0.36, Recall: 0.59, F1 Sc
Average values for all thresholds for Query Index: 373
avg-Precision: 0.28
avg-Recall: 0.35
avg-F1 Score: 0.27
```

Discussion of 3.3:

Previous Results with Mean, Standard Deviation, and Skewness:

- The baseline with equal weights (mean, standard deviation, skewness) achieved an AUC of 0.67 and an F1 score of 0.1604.
- Adjusting weights to favor standard deviation showed some improvements in the F1 score but mixed results for AUC.

Results with Additional Moments:

- Balanced Approach with All Moments:

Weights: [1, 1, 1, 1, 1, 1] * 3

AUC: 0.68, F1 score: 0.2327

Including all moments with equal weighting resulted in a slight improvement in AUC and a significant increase in the F1 score compared to the baseline, indicating that these additional moments contribute positively to the system's performance.

Emphasizing Median Color Value:

Slight emphasis ([1, 1, 1, 3, 1, 1] * 3): AUC decreased to 0.66, F1 score to 0.1850.

Strong emphasis ([1, 1, 1, 10, 1, 1] * 3): AUC further decreased to 0.65, F1 score to 0.1643.

It appears that emphasizing the median color value diminishes the system's effectiveness, as both the AUC and F1 score decreased compared to the balanced approach.

Emphasizing Most Common Color (Mode):

Moderate emphasis ([1, 1, 1, 1, 3, 1] * 3): AUC was 0.67, F1 score 0.2597.

Strong emphasis ([1, 1, 1, 1, 15, 1] * 3): AUC decreased to 0.66, F1 score to 0.2319.

Different strong emphasis ([1, 1, 1, 1, 10, 1] * 3): AUC remained at 0.66, F1 score slightly higher at 0.2331.

These results suggest that while the mode is an important feature, overemphasis beyond a certain threshold does not necessarily yield better retrieval performance.

Emphasizing Color Distribution Shape (Kurtosis):

Moderate emphasis ([1, 1, 1, 1, 1, 3] * 3): AUC was 0.68, F1 score 0.2332.

Strong emphasis ([1, 1, 1, 1, 1, 10] * 3): AUC was consistent at 0.68, F1 score marginally increased to 0.2336.

Emphasizing kurtosis slightly improves the F1 score without affecting the AUC, indicating that kurtosis contributes to a balanced precision-recall trade-off.

Additional Custom Scenario:

Weights: [2, 10, 2, 3, 3, 6] * 3

AUC was 0.69, F1 score 0.2897

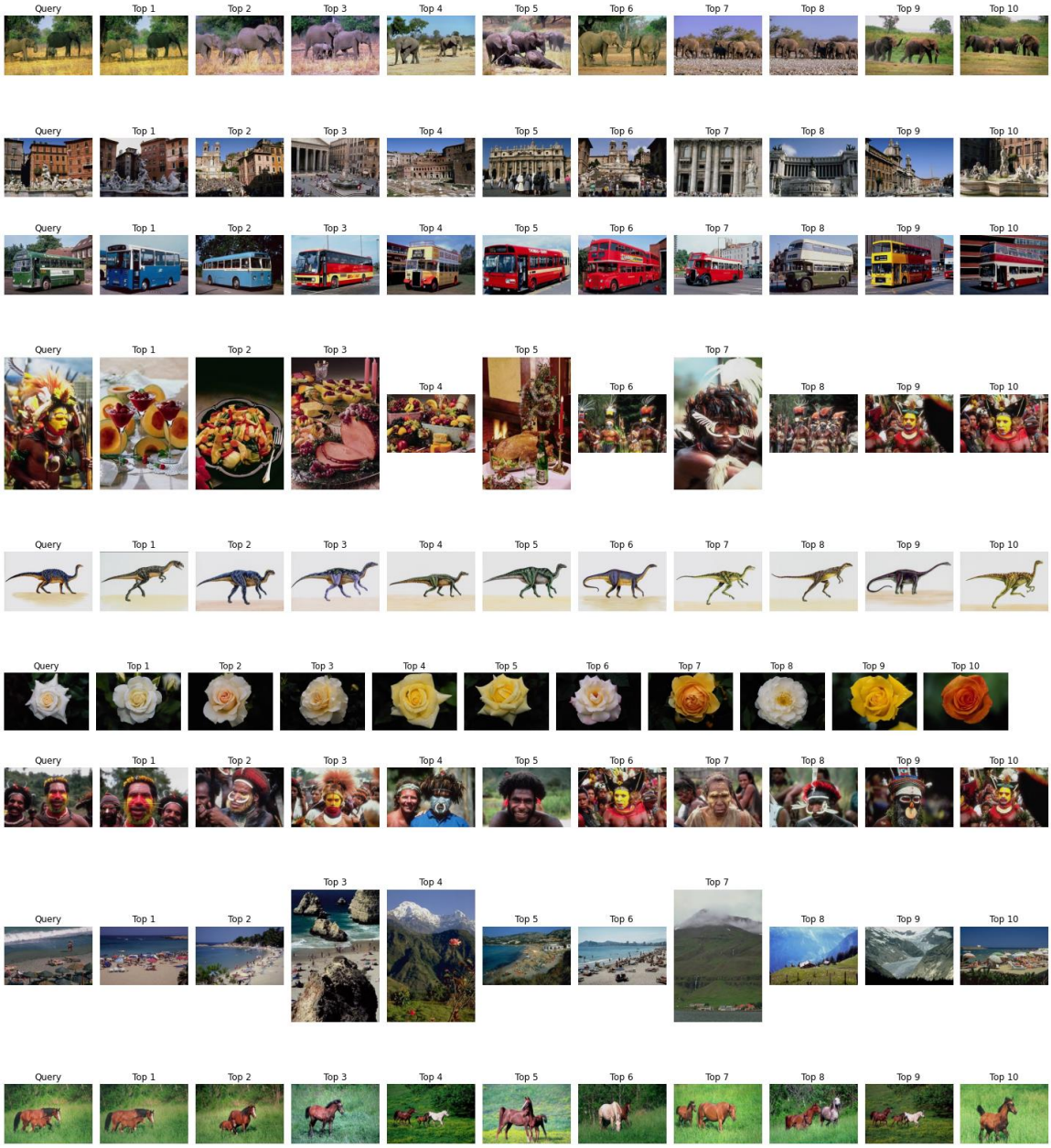
This custom scenario, which gives varying emphasis across different moments, resulted in the best AUC and F1 score among all the configurations tested, suggesting that a more nuanced combination of all moments leads to the most effective performance.

Comment on task 2 & 3 :

A CBIR system using color moments seems to be more sensitive to threshold adjustments, and fine-tuning this parameter can lead to significant changes in performance. The histogram-based approach seems to be more sensitive to the number of bins, which affects its ability to discriminate between different colors in the images.

3.3 Task 4: Improvement Using CNN

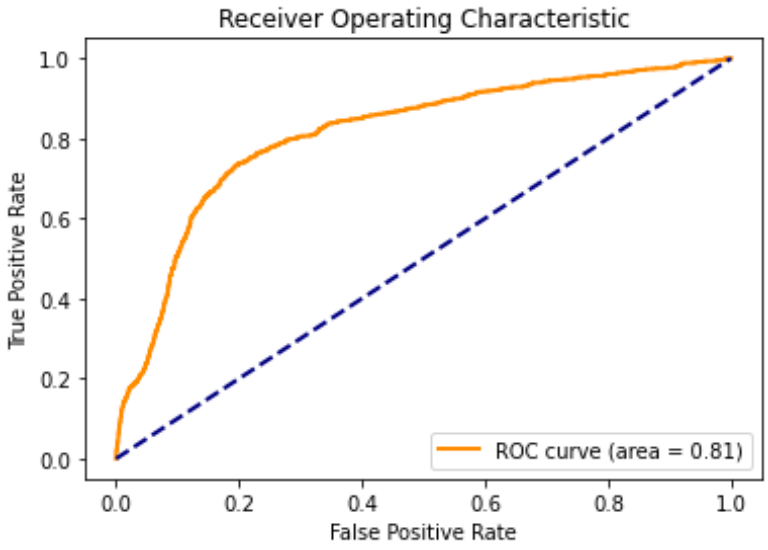
Each query with its top 10 similar images:



Appling measures for the top 100 similar images to the query.

```
N_QUERIES = 10
TOP_K = 100
```

AUC was 0.81 as shown below:



The AUC of 0.81 indicates a good level of separability; the model is able to distinguish between the positive class (similar images) and the negative class (dissimilar images) with a reasonable degree of accuracy.

Measures for each query and average measures for all queries :

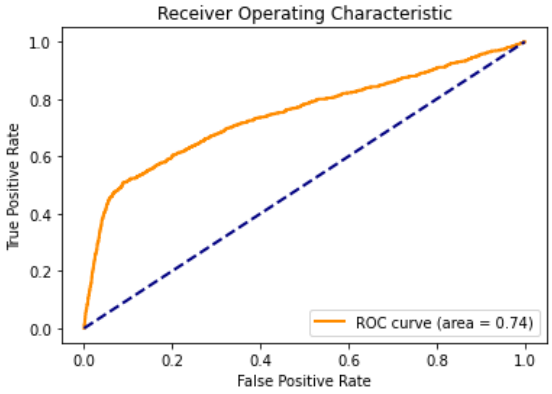
```
Query 1: Precision = 0.5500, Recall = 0.5500, F1-score = 0.5500
1/1 [==============================] - 0s 117ms/step
Query 2: Precision = 0.4000, Recall = 0.4000, F1-score = 0.4000
1/1 [==============================] - 0s 113ms/step
Query 3: Precision = 0.5000, Recall = 0.5000, F1-score = 0.5000
1/1 [==============================] - 0s 112ms/step
Query 4: Precision = 0.3800, Recall = 0.3800, F1-score = 0.3800
1/1 [==============================] - 0s 111ms/step
Query 5: Precision = 0.6800, Recall = 0.6800, F1-score = 0.6800
1/1 [==============================] - 0s 119ms/step
Query 6: Precision = 0.3800, Recall = 0.3800, F1-score = 0.3800
1/1 [==============================] - 0s 113ms/step
Query 7: Precision = 0.3800, Recall = 0.3800, F1-score = 0.3800
1/1 [==============================] - 0s 114ms/step
Query 8: Precision = 0.5800, Recall = 0.5800, F1-score = 0.5800
1/1 [==============================] - 0s 131ms/step
Query 9: Precision = 0.3000, Recall = 0.3000, F1-score = 0.3000
1/1 [==============================] - 0s 110ms/step
Query 10: Precision = 0.4500, Recall = 0.4500, F1-score = 0.4500
Average values for all thresholds for Query Index: 9

Average across all queries: Precision = 0.4600, Recall = 0.4600, F1-score = 0.4600
```

```
Execution Time: 7.88 seconds
```

The average precision across all queries is 0.4600, which suggests that when the model retrieves a set of images, about 46% of the retrieved images are relevant. The average recall (also 0.4600) indicates that the model retrieves about 46% of all relevant images in the dataset. The average F1-score, which balances precision and recall, is also 0.4600, showing that there is a balance between precision and recall in this model's performance.

Increasing top-K affects AUC and the whole measures decrease as shown:

```
TOP_K = 150
```




```
1/1 [==============================] - 0s 90ms/step
Query 1: Precision = 0.2133, Recall = 0.3200, F1-score = 0.2560
1/1 [==============================] - 0s 126ms/step
Query 2: Precision = 0.5000, Recall = 0.7500, F1-score = 0.6000
1/1 [==============================] - 0s 109ms/step
Query 3: Precision = 0.0600, Recall = 0.0900, F1-score = 0.0720
1/1 [==============================] - 0s 122ms/step
Query 4: Precision = 0.0600, Recall = 0.0900, F1-score = 0.0720
1/1 [==============================] - 0s 122ms/step
Query 5: Precision = 0.4733, Recall = 0.7100, F1-score = 0.5680
1/1 [==============================] - 0s 117ms/step
Query 6: Precision = 0.3200, Recall = 0.4800, F1-score = 0.3840
1/1 [==============================] - 0s 134ms/step
Query 7: Precision = 0.4067, Recall = 0.6100, F1-score = 0.4880
1/1 [==============================] - 0s 114ms/step
Query 8: Precision = 0.4933, Recall = 0.7400, F1-score = 0.5920
1/1 [==============================] - 0s 108ms/step
Query 9: Precision = 0.4067, Recall = 0.6100, F1-score = 0.4880
1/1 [==============================] - 0s 125ms/step
Query 10: Precision = 0.3333, Recall = 0.5000, F1-score = 0.4000
Average values for all thresholds for Query Index: 9

Average across all queries: Precision = 0.3267, Recall = 0.4900, F1-score = 0.3920

Execution Time: 8.04 seconds
```

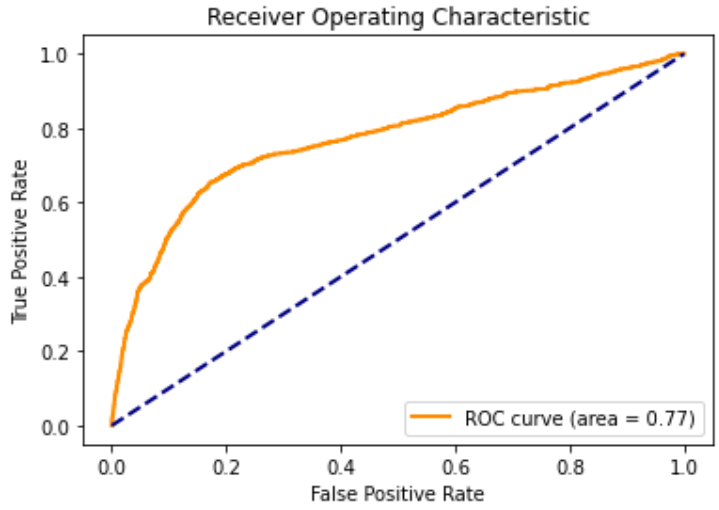
TOP_K = 150, AUC = 0.74:

- The AUC is slightly lower compared to TOP_K = 100, suggesting that increasing the number of top results considered introduces more false positives, thus reducing the model's ability to distinguish between similar and dissimilar images effectively.
- The precision drops to 0.3267, indicating that the inclusion of more images in the top results leads to more false positives.
- The recall increases to 0.4900, suggesting that while more relevant images are retrieved, the precision suffers.

- The F1-score is 0.3920, which is lower than the TOP_K = 100 scenario, reflecting the trade-off between precision and recall.

Decreasing top-k less than 100:

```
TOP_K = 77
```



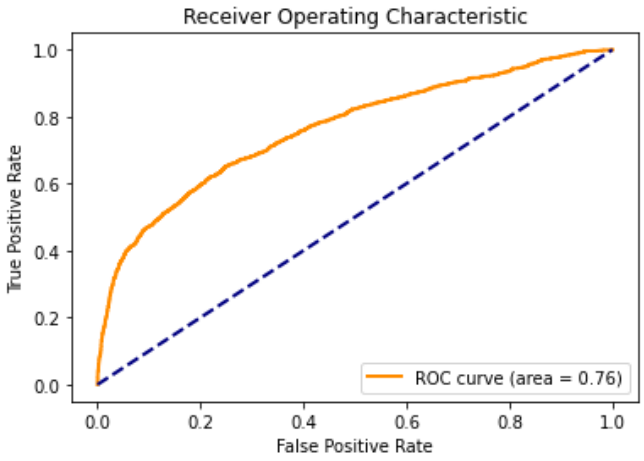

```
Query 1: Precision = 0.8701, Recall = 0.6700, F1-score = 0.7571
1/1 [==============================] - 0s 104ms/step
Query 2: Precision = 0.0649, Recall = 0.0500, F1-score = 0.0565
1/1 [==============================] - 0s 110ms/step
Query 3: Precision = 0.4156, Recall = 0.3200, F1-score = 0.3616
1/1 [==============================] - 0s 110ms/step
Query 4: Precision = 0.5455, Recall = 0.4200, F1-score = 0.4746
1/1 [==============================] - 0s 120ms/step
Query 5: Precision = 0.2727, Recall = 0.2100, F1-score = 0.2373
1/1 [==============================] - 0s 110ms/step
Query 6: Precision = 0.6234, Recall = 0.4800, F1-score = 0.5424
1/1 [==============================] - 0s 109ms/step
Query 7: Precision = 0.5974, Recall = 0.4600, F1-score = 0.5198
1/1 [==============================] - 0s 113ms/step
Query 8: Precision = 0.7403, Recall = 0.5700, F1-score = 0.6441
1/1 [==============================] - 0s 113ms/step
Query 9: Precision = 0.5584, Recall = 0.4300, F1-score = 0.4859
1/1 [==============================] - 0s 110ms/step
Query 10: Precision = 0.4935, Recall = 0.3800, F1-score = 0.4294
Average values for all thresholds for Query Index: 9

Average across all queries: Precision = 0.5182, Recall = 0.3990, F1-score = 0.4508

Execution Time: 8.64 seconds
```

TOP_K = 77, AUC = 0.77:

- This AUC is lower than the best performing TOP_K = 100 but higher than the TOP_K = 150, suggesting that there might be an optimal range around TOP_K = 100 where the model performs better.

- The average precision of 0.5182 is an improvement over both the TOP_K = 100 and 150 scenarios, suggesting that a smaller set of results might be more precise.
- The average recall drops to 0.3990, indicating fewer relevant images are retrieved compared to the TOP_K = 150 scenario.
- The F1-score of 0.4508 is slightly lower than the TOP_K = 100 scenario, which suggests a better balance between precision and recall was achieved at TOP_K = 100.

And when top-k is 50 , measures decreased:




```
Query 1: Precision = 0.3000, Recall = 0.1500, F1-score = 0.2000
1/1 [==============================] - 0s 121ms/step
Query 2: Precision = 0.5000, Recall = 0.2500, F1-score = 0.3333
1/1 [==============================] - 0s 120ms/step
Query 3: Precision = 0.3400, Recall = 0.1700, F1-score = 0.2267
1/1 [==============================] - 0s 113ms/step
Query 4: Precision = 0.5000, Recall = 0.2500, F1-score = 0.3333
1/1 [==============================] - 0s 124ms/step
Query 5: Precision = 0.5800, Recall = 0.2900, F1-score = 0.3867
1/1 [==============================] - 0s 116ms/step
Query 6: Precision = 0.6200, Recall = 0.3100, F1-score = 0.4133
1/1 [==============================] - 0s 113ms/step
Query 7: Precision = 0.7400, Recall = 0.3700, F1-score = 0.4933
1/1 [==============================] - 0s 113ms/step
Query 8: Precision = 0.7400, Recall = 0.3700, F1-score = 0.4933
1/1 [==============================] - 0s 114ms/step
Query 9: Precision = 0.3600, Recall = 0.1800, F1-score = 0.2400
1/1 [==============================] - 0s 113ms/step
Query 10: Precision = 0.5600, Recall = 0.2800, F1-score = 0.3733
Average values for all thresholds for Query Index: 9

Average across all queries: Precision = 0.5240, Recall = 0.2620, F1-score = 0.3493
```

TOP_K = 50, AUC = 0.76:

- The AUC is almost similar to TOP_K = 77, indicating consistent model performance in distinguishing between similar and dissimilar images.
- The precision increases to 0.5240, which is the highest among all scenarios, suggesting that a smaller set of top results yields more accurate predictions.
- However, the recall drops to 0.2620, which is the lowest among the scenarios, indicating that many relevant images are being missed.
- The F1-score of 0.3493 reflects the trade-off, showing that while the precision is high, the recall suffers significantly, resulting in a lower overall F1-score.