



Evaluating Photo Aesthetics Using MobileNetV3 and Feature Engineering

Frank Zhao¹, Scott Xu¹

Department of Computer Science, Stanford University

Stanford
Computer Science

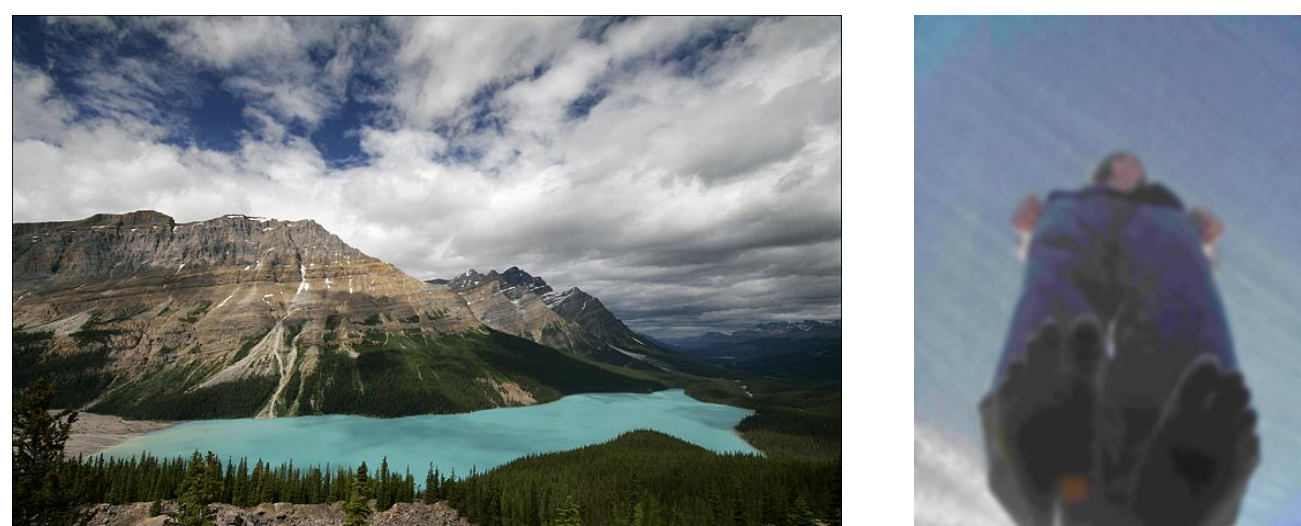
Project Overview

- For a professional photographer, sorting through thousands of images can be time-consuming and inefficient. Thus, **automatic image aesthetics assessment** can be a game changer.
- It can also be applied in a **wide variety of applications**, such as image thumbnailing and automatic editing.
- Pass original image through **MobileNetV3 - Large**.
- Different from previous work, we also pass the image's **saliency map** generated from a pretrained model through **MobileNetV3 - Small**.
- Color histogram and HOG as augmentation** before 2 fully connected layers.
- Achieves **77.2%** accuracy in binary classification.
- Problem Statement:** takes an image and its auxiliary features to predict a discrete probability distribution over the rating 1-10.
- Main contribution:** Feature Engineering, an Aggregation of MobileNetV3 networks, and demonstration of some important applications.

Datasets & Metrics

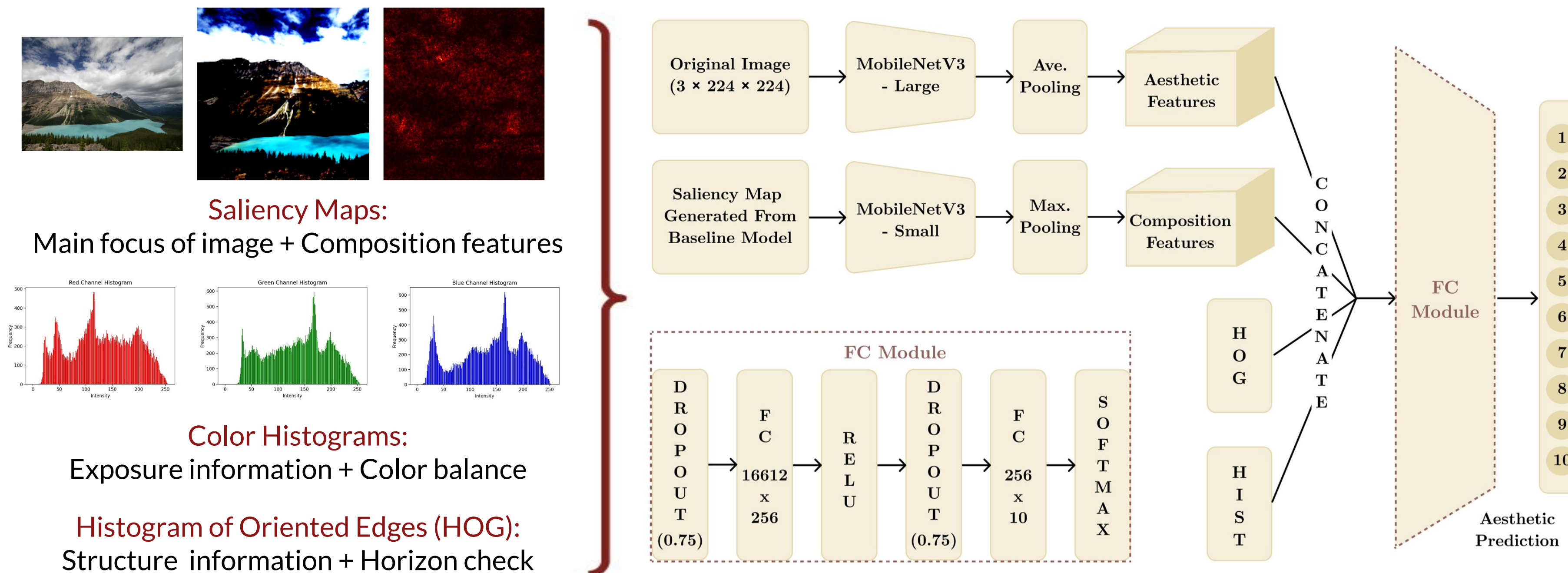
- Uses **AVA (Aesthetics Visual Analysis)** dataset.
- 255,000 images** rated manually based on aesthetic qualities.
- The image rating ranges from **1 to 10** with 10 being the highest. The mean rating is around 5.5.
- Use **5 as cutoff score**. Images with mean rating less than 5 are "bad" and higher than 5 are "good".
- Objective is to minimize the **EMD (Earth Mover's Distance)** loss to compare ordered classes:

$$EMD(p, \hat{p}) = \left(\frac{1}{N} \sum_{k=1}^N |CDF_p(k) - CDF_{\hat{p}}(k)|^r \right)^{\frac{1}{r}}$$



Examples: Image with highest predicted rating (Left), Image with lowest predicted rating (Right).

Methods & Model Architecture



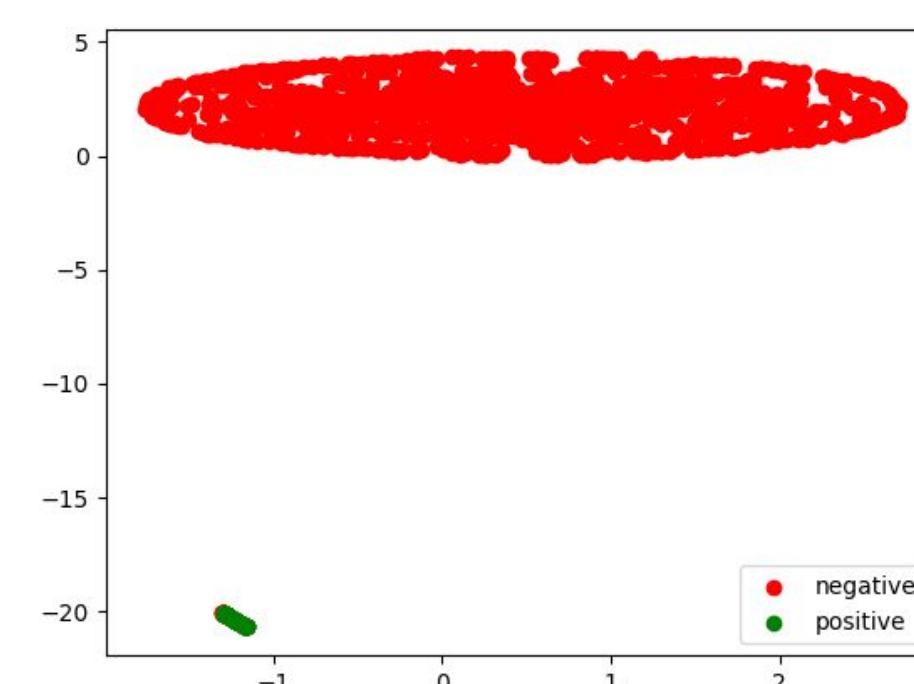
Evaluations & Applications

Feature Importance

Feature Type	Average Attribution
Image Features	0.01465
Saliency Map	0.00136
Color Histogram	1.22×10^{-6}
HOG	2.22×10^{-6}

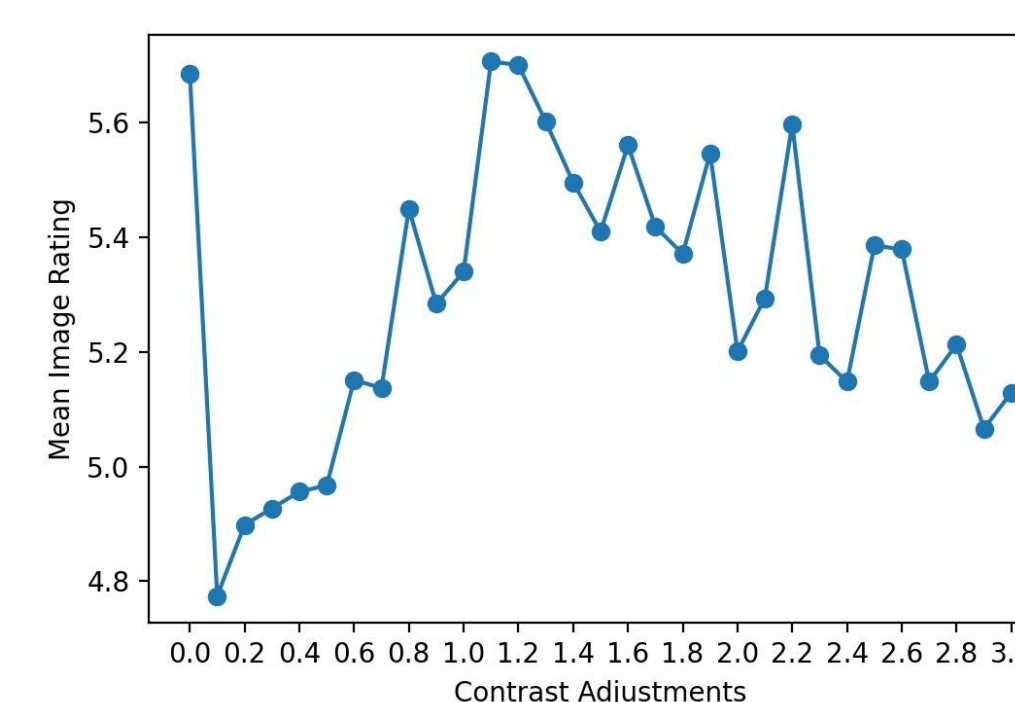
- Integrated gradients method in Captum.
- Color histogram and HOG not so useful.
- Saliency map more useful.
- In future work, get rid of histogram and HOG, and replace the original image's network to Inception to compare.

t-SNE Graph Before Last FC Layer



- High separability.
- Model is expressive.
- Maybe 2 FC layers is not necessary.
- Future work: compare with 1 FC layer architecture.

Effect of Image Adjustments on Aesthetic Scoring



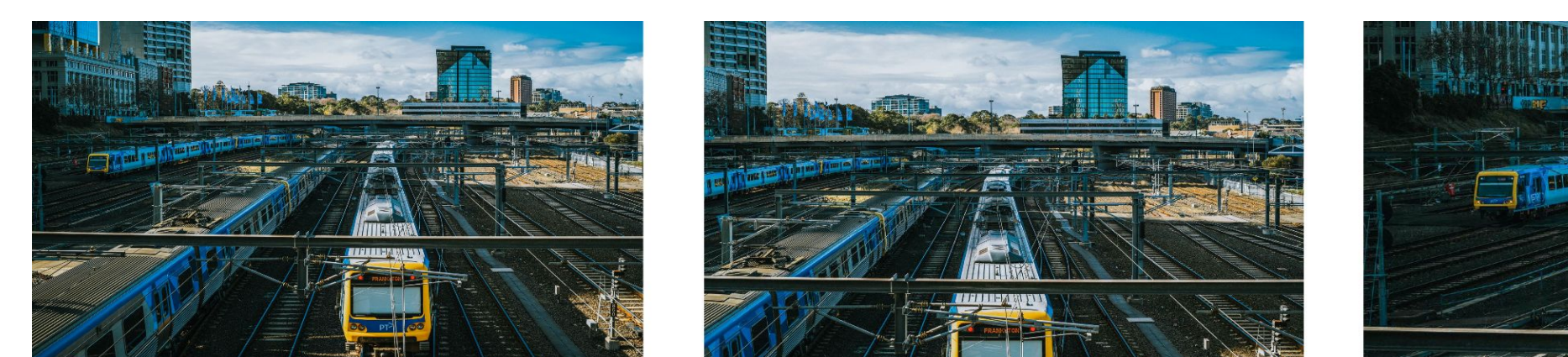
- Inverted U-shaped: if contrast is too low or too high, predicted aesthetics scores decline.
- Similar shape for brightness.
- Monotonic increase for sharpness.



Original Image 5.651 (± 1.501) Auto-enhanced 5.735 (± 1.501)

- Use best adjustment factor in each to auto-enhance image.

Automatic Cropping

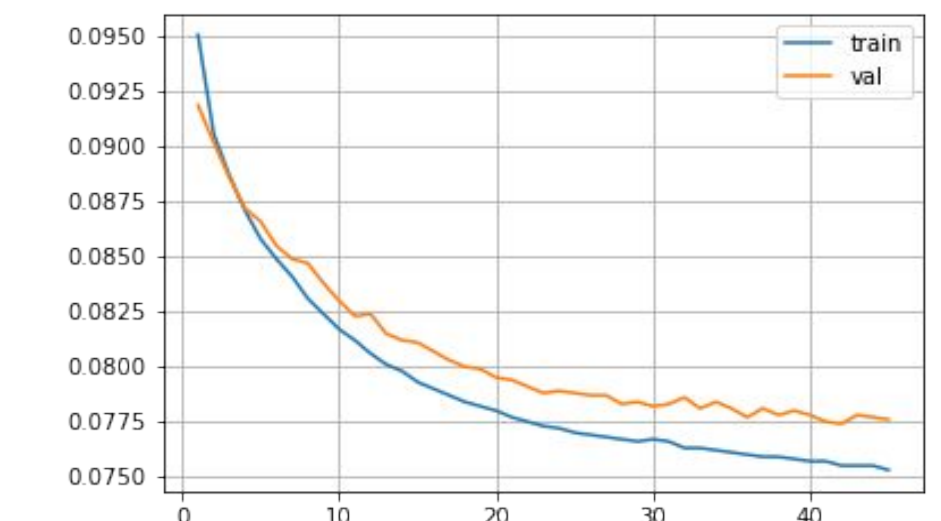


Original Image 5.708 (± 1.494)

Best Crop 5.757 (± 1.484)

Worst Crop 4.735 (± 1.445)

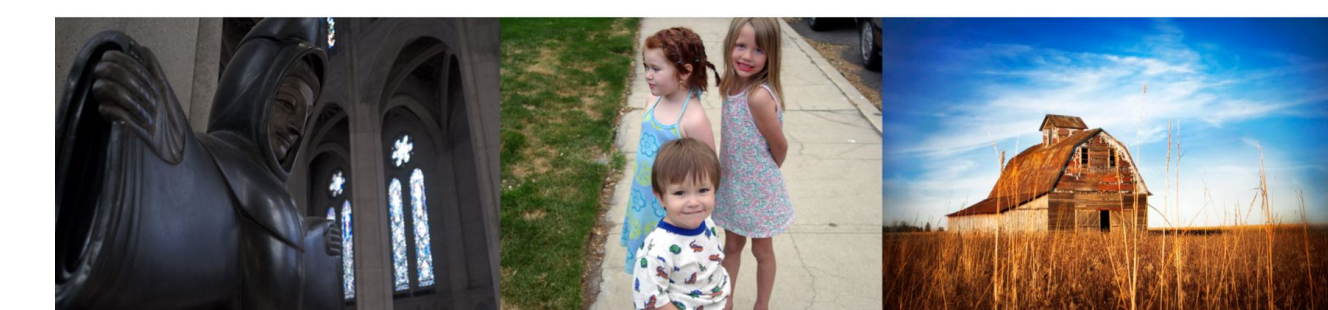
Results



Our model converges in 45 epoches, compared to 85 epoches for baseline.

	Baseline	Our model
Accuracy	76.5%	77.2%
Precision	78.4%	79.1%
Recall	92.0%	91.9%
F1 Score	84.6%	85.0%
LCC of Mean	0.595	0.608
LCC of std. dev.	0.206	0.078
EMD Loss	0.0712	0.0718

- Higher accuracy with half training epoches.
- Precision & recall reasonably high.
- Predicts mean very well, but predicts standard deviations rather poorly.



4.97 (± 1.42) 4.99 (± 1.40) 5.78 (± 1.49)
4.97 (± 1.45) 4.99 (± 1.57) 5.78 (± 1.29)

Top 3 Success Cases



4.92 (± 1.44) 4.95 (± 1.44) 4.78 (± 1.44)
1.81 (± 1.49) 8.00 (± 1.67) 2.08 (± 1.83)

Top 3 Failure Cases

- Predicts well for images with simple subject, mediocre content, and conventional style.
- Predicts poorly for monotone and unconventional images. Fails to predict extreme values.