Understanding Biases in Human Portrait Colorization by a Conditional Wasserstein GAN

Edward Deleu and Vishva Ilavelan

Princeton University, Princeton NJ, USA

Supported By:
Princeton Research Computing

COS 429, Fall 2024
Professors Olga Russakovsky and Vikram Ramaswamy

Background

- Realistically adding color to grayscale portraits is challenging!
- Models must naturally colorize hair, skin-tones, and clothing for a diverse range of people
  - It is difficult to infer skin-tones and color from grayscale images, especially considering variable lighting conditions



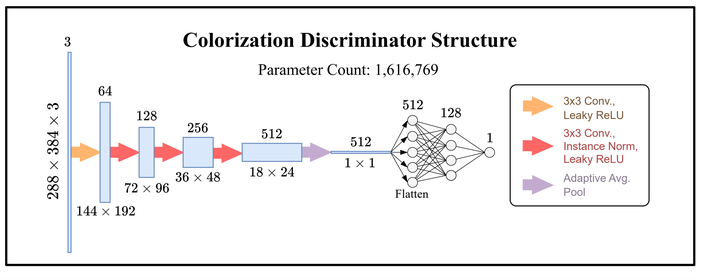
- AISegment Dataset: 34,427x clear front-facing portrait images at 600x800 with corresponding background removed images
  - Heavily biased towards white individuals, but it has reasonable age and gender diversity

Approach

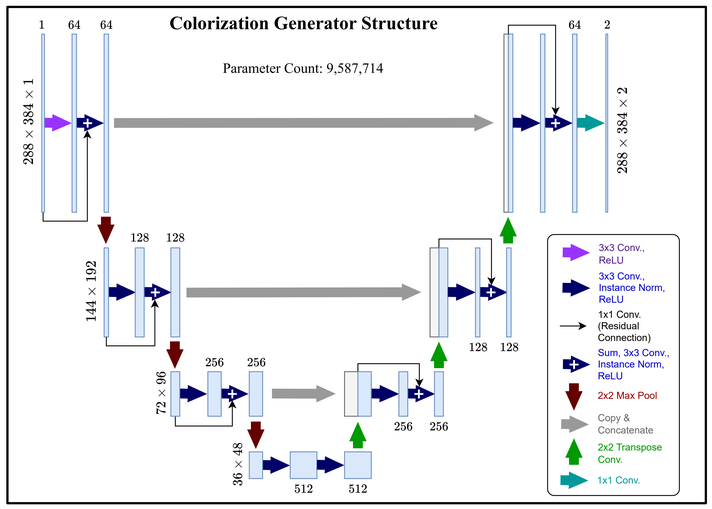
- Conditional Wasserstein Generative Adversarial Network

$$\min_G \max_{D \in \mathcal{D}} \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathbb{P}_R} [D(\boldsymbol{x})] - \mathop{\mathbb{E}}_{\tilde{\boldsymbol{x}} \sim \mathbb{P}_G} [D(\tilde{\boldsymbol{x}})]$$

- Where $\mathbb{P}_R$ is the real data and $\mathbb{P}_G$ is the generated data
- WGAN loss constraint stabilizes & improves GAN training
- Utilize *Lab* color space (cf. RGB color space)
  - *L* is grayscale image, *a* & *b* are color channels
- Generator colorizes images conditioned by L color map, outputting fake *a* & *b* channels
  - Randomness induced by dropout between all layers
- Discriminator scores fakeness of colorized and real images



Training

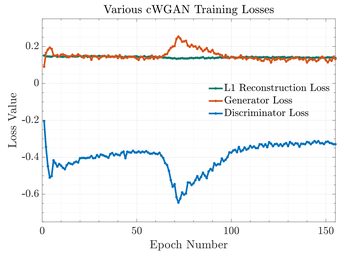


- 25% of the AISegment dataset used for training: 8,607 images
  - Images resized to 288x384 for faster training and random horizontal flips are applied
- The WGAN Generator Loss is and L1 loss is added between real and reconstructed images

$$\mathcal{L}_{WGAN,Gen} = \frac{1}{N}\sum_{i=1}^{N} -D(L_i, G(L_i))$$

- The Discriminator Loss is:

$$\mathcal{L}_{Discrim} = \frac{1}{N}\sum_{i=1}^{N} -D(L_i, a_i, b_i) + D(L_i, G(L_i)))$$



- 160 hours of hyperparameter tuning on A100 GPUs
- Final model trains in 20 hours for 155 epochs

Grayscale images and generator's reconstructions (below):



Results

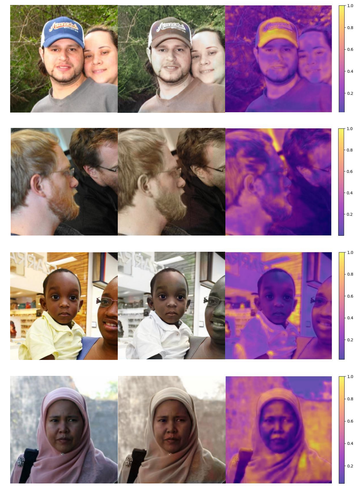
- FairFace dataset contains ~100k facial images categorized by age, race, and gender

(Below): White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino – Real images in top row, generated below



- Despite the lower quality images in this dataset, results are consistent across racial demographics
  - Notably, colorized images are less vibrant and are more dull
  - Some regions of images lack color or are overly saturated red
- For 300 random images of each race from the FairFace dataset, per-channel MAE and overall MSE are computed between real and generated images
  - Similar results across demographics, with slightly worse accuracy on Southeast Asian, Black, and Indian classes
  - Blue MAE values are much lower than red and green, potentially stemming from the Generator's *Lab* color space

Demographic	Red MAE	Blue MAE	Green MAE	RGB MSE
White	13.4165	5.0175	15.9579	23.4175
Black	14.4832	5.1362	16.9982	25.1666
Indian	14.4709	5.4780	16.1638	24.4988
East Asian	14.1309	5.4172	15.7837	24.0005
Southeast Asian	14.6744	5.5838	17.0340	25.2815
Middle Eastern	12.9918	4.5700	15.5266	22.4669
Latino	13.4782	5.3508	15.5070	23.2841



(Left): Real, generated images, and heatmaps of MSE between real and generated images

- Colorization of clothing and backgrounds is less accurate and more irregular
- Vibrant colors are colorized as neutral tones.
- Age and gender appear to have limited influence on colorization quality
- The model appears consistent among races, even with the biased training data!