An Unsupervised Approach towards Varying Human Skin Tone Using Generative Adversarial Networks

Debapriya Roy, Diganta Mukherjee and Bhabatosh Chanda

Indian Statistical Institute, Kolkata, India
Email: debapriyakundu1@gmail.com, diganta@isical.ac.in, chanda@isical.ac.in

Abstract

We propose a model to change skin tone of a person. Given any input image of a person or a group of persons with some value indicating the desired change of skin color towards fairness or darkness, this method can change the skin tone of the persons in the image. An illustrative example is shown in Fig. 1.

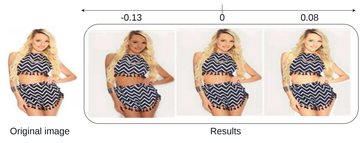

Fig. 1: Illustrating the objective of the present work.

- This is an unsupervised method.
- Unconstrained in terms of pose, illumination, number of persons in the image etc.
- The goal of this work is to reduce the time and effort which is generally required for changing the skin tone using existing applications (e.g., Photoshop) by professionals or novice.

Methodology

- We first segment the given image pixels in two classes, skin and non-skin. This is achieved by our skin segmentation network which is a Convolutional Neural Network (CNN).

  - This network consists of two sub-networks. The first half of this network is dedicated to skin segmentation objective. This part of the network is trained using the following loss function,

$$L_{seg} = L_c(\hat{x}_{seg}, x_{seg}) + L_p(\hat{x}_{seg}, x_{seg}) + L_s(\hat{x}_{seg}, x_{seg}), \tag{1}$$

  where L_c $(.\ ,\ .)$ represents count loss and L_p $(.\ ,\ .)$, L_s $(.\ ,\ .)$ represents the perceptual [5] and SSIM [10] loss respectively between the ground truth (say, x_{seg}) and the predicted segmentation result (say, $\hat{x}_{seg}$).

  - Skin color estimation sub-network - estimates a rgb value indicating the skin tone. We used MS-SSIM [9] loss, which is a variant of SSIM [10]. As this method is unsupervised therefore we do not use any ground truth skin color annotations for training this network. Instead we fill the detected skin areas with the predicted skin color and minimize the structural dissimilarity between the input and the manipulated image.

- The next part of this work which is the final image synthesizer employs a Conditional Generative Adversarial Network [7] (cGAN). It takes the image as input, along with the value of a conditional variable and synthesizes a new image with the skin tone of the persons in the image changed in accordance with the value of the variable.

  - We formulate this problem as a conditional image generation problem, where the source image, along with its skin segmentation (obtained from the skin segmentation network discussed in the previous section) and a control variable z is given as input to a cGAN. z controls the amount of change of skin tone. The value of $z = 0$ indicates no change of skin color while, values less than zero and above zero indicates the amount of change towards darkness and fairness of skin respectively. Therefore z here plays the role of a skin tone regulator.

  - We formulate the objective function in the following way,

$$L_{cGAN} = l^1 + l^2 + \lambda(m \times z + l^3 - \epsilon) + L_{ADV}. \tag{2}$$

Where considering $\hat{x}_{z=0} = f_g(x, z = 0, \hat{x}_{seg})$ and $\hat{x}_{z\neq 0} = f_g(x, z \neq 0, \hat{x}_{seg})$, we define,

$$\begin{aligned} l^1 &= L_p(\hat{x}_{z=0}, x) \\ l^2 &= L_p(\hat{x}_{z=0} \times \hat{x}'_{seg}, x \times \hat{x}'_{seg}) \\ l^3 &= \log(0.5 - L_{color}) \\ L_{color} &= L_p^{color}(\hat{x}_{z\neq 0}, x). \end{aligned} \tag{3}$$

Here λ, m and ϵ are parameters. L_{ADV} denotes the adversarial loss. The function $L_p(.,.)$ indicates VGG-perceptual loss and $L_p^{color}(.,.)$ indicated a loss similar in concept to perceptual loss but the underlying network is the skin color estimation network

Datasets

We evaluate our model based on the following datasets, Category and Attribute Prediction Benchmark, In Shop Cloth Retrieval dataset of DeepFashion [6] and MPV [3]. The In Shop Cloth Retrieval dataset contains in total 52,712 images of multiple views of each person (front, side, back and full) while the Category and Attribute Prediction Benchmark dataset contains 289,222 number of clothing images, where the images are mostly of models wearing the clothing. MPV dataset contains in total 35,687 images of multiple views of each person.

Quantitative Analysis

For quantitative analysis we have reported the scores on the following metrics, Inception Score [8] (IS) and Frechet Inception Distance [4] (FID), SSIM [10]. We report the value of Kolmogorov-Smirnov test [2] statistic which is a goodness of fit test. The values of SSIM are based on the results of the cGAN generator with $z = 0$. The scores of IS and FID, suggests that our method synthesizes quite good quality images which can also be verified visually from the results presented in the qualitative analysis section.

Dataset	IS↑	FID ↓	SSIM↑
In Shop	3.21 ± 0.17	38.33	0.93
Category-and-Attribute	3.58 ± 0.19	36.19	0.95
MPV	3.03 ± 0.23	42.56	0.92

Tab. 1: Values of Incepion Score (IS) and Frechet Inception Distance (FID) and SSIM on results of different data sets.

Dataset	KS statistic ↓	P-Value↑
DeepFashion (Category-and-Attribute)	0.0249	0.5545
MPV	0.0450	0.0837

Tab. 2: Values of Kolmogov-Smirnov test (KS test) statistic along with the corresponding P-values on results of different data sets.

Qualitative Analysis

Fig. 2: Results of our method. The source person images are shown in the 1st column, the next column shows the skin pixel segmentation results. The following columns show the results for different values of z. Walking along the axis from negative to positive direction increases the fairness of skin.

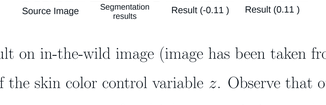

Fig. 3: Demonstration of result on in-the-wild image (image has been taken from [1]). The values within the brackets indicates the value of the skin color control variable z. Observe that our segmentation module is not constrained to background clutter or presence of multiple persons in image. Also it is worth noticing that the results of the cGAN is perceptually convincing in terms of skin tone variation.

References

[1] `https://p4t6u7k5.stackpathcdn.com/wp-content/uploads/shutterstock_605616227.jpg`.

[2] Laha Chakravarti. *Roy, Handbook of Methods of Applied Statistics, 1 (1967)*.

[3] Haoye Dong et al. "Towards multi-pose guided virtual try-on network". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 9026–9035.

[4] Martin Heusel et al. "Gans trained by a two time-scale update rule converge to a local nash equilibrium". In: *Advances in Neural Information Processing Systems*. 2017, pp. 6626–6637.

[5] Justin Johnson, Alexandre Alahi and Li Fei-Fei. "Perceptual losses for real-time style transfer and super-resolution". In: *European conference on computer vision*. Springer. 2016, pp. 694–711.

[6] Ziwei Liu et al. "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 1096–1104.

[7] Mehdi Mirza and Simon Osindero. "Conditional generative adversarial nets". In: *arXiv preprint arXiv:1411.1784* (2014).

[8] Tim Salimans et al. "Improved techniques for training gans". In: *Advances in neural information processing systems*. 2016, pp. 2234–2242.

[9] Zhou Wang, Eero P Simoncelli and Alan C Bovik. "Multiscale structural similarity for image quality assessment". In: *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. Vol. 2. Ieee. 2003, pp. 1398–1402.

[10] Zhou Wang et al. "Image quality assessment: from error visibility to structural similarity". In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.