Faces as Lighting Probes via Unsupervised Deep Highlight Extraction

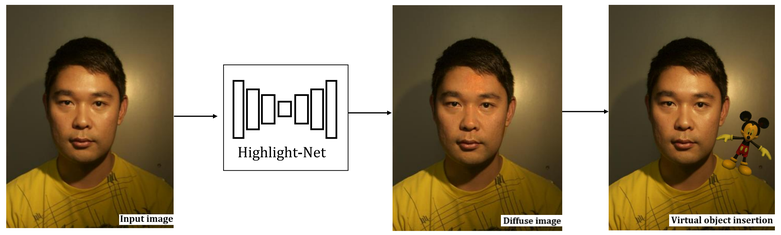
Renjiao Yi[1,2], Chenyang Zhu[1,2], Ping Tan[1], Stephen Lin[3]

{renjiaoy, cza68, pingtan}@sfu.ca, stevelin@Microsoft.com

[1]Simon Fraser University, [2]NUDT, [3]Microsoft Research

Motivations

- Remove specular highlights from a single face photo
- Estimate illumination environment from the specular highlights for rendering virtual objects realistically



Challenges

Lack of training data:

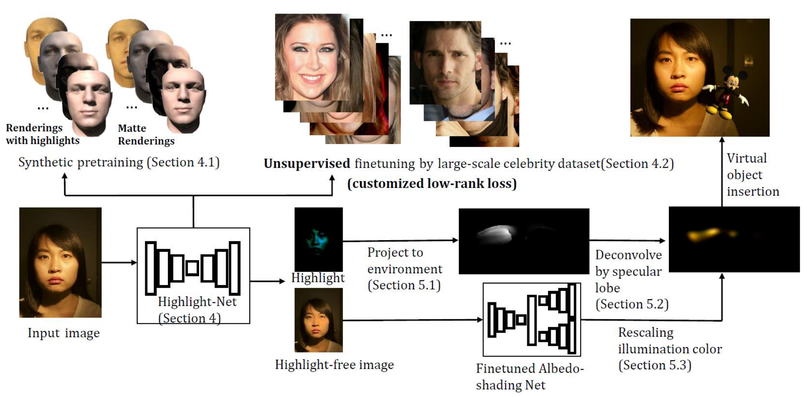
- Synthetic data does not work well
- Impractical to capture ground truth real data under natural illumination

Our observation

- Diffuse chromaticity over a given person's face is invariant across images

Overview

The highlight reflections are predicted from a single image by Highlight-Net, then they are traced back to the scene to recover a non-parametric environment map, with which virtual objects can be inserted into the input image with consistent lighting.



Main contributions

- An unsupervised method for training a highlight extraction network using unlabeled real data
- Recovery of a non-parametric illumination representation that includes both low- and high-frequency components

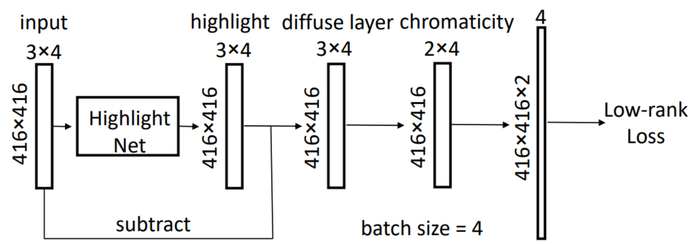
Method

Pretraining by synthetic data

- **Training data**: a small set of synthetic data, with ground truth
- **Problem**: does not work well for real data, as shown in column 4 of Fig. 1, due to the gap in appearance between synthetic and real images
- **Solution**: an indirect method for finetuning with unlabeled real data

Unsupervised finetuning by real data

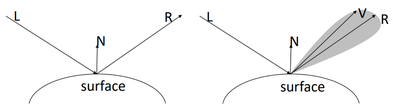
- **Dataset**: the Microsoft-Celeb-1M dataset (each celebrity has more than 100 unlabelled images)
- **Preprocessing**: a set of calibrations



- **Finetuning**: 4 input images of the same person form a batch. Diffuse layers are obtained by Highlight-Net and transformed to chromaticity, then each of 4 chromaticity maps are reshaped to a vector and stacked together as a matrix D.
- **Unsupervised loss**: since the diffuse chromaticity of a person's face should be the same in all images, matrix D should be low rank if the highlights are correctly removed. We thus define a low rank loss as the second singular value σ_2:

$$D = U\Sigma V^T \text{ (SVD decomposition)}$$
$$\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \sigma_3, \sigma_4)$$
$$loss_{lowrank} = \sigma_2$$

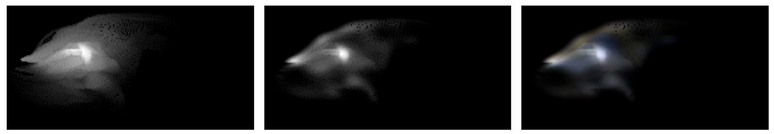
Illumination estimation



Mirror reflection vs Phong specular reflection

Step 1: Highlights on faces are treated as mirror reflection, and projected back to environment map (with estimated face normal)

Step 2: To account for the non-mirror reflections, the environment map is deconvolved by Phong lobe defined by statistics in the MERL/ETH Skin Reflectance Database.



Initial traced environment map | Map after deconvolution | Final environment maps after illumination color rescaling (Sec 5.3)

Step 3: Repeating for 3 channels, we get the color environment map.

Results

Fig.1 Comparisons of highlight removal on real data (ground truth diffuse images captured by cross-polarization)

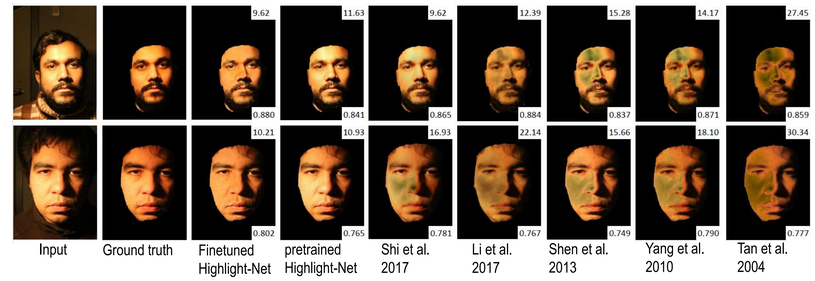
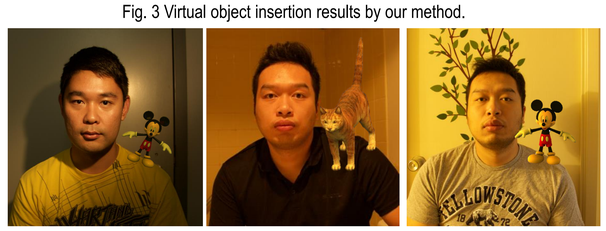


Fig. 2 Highlight removal results on challenging data with strong expressions, occluders, various skin tones, and ages

Fig. 3 Virtual object insertion results by our method.



Scan to check the paper and codes: