Synthesizing Normalized Faces from Facial Identity Features

Forrester Cole, David Belanger, Dilip Krishnan, Aaron Sarna, Inbar Mosseri, William T. Freeman,
Google, Inc. University of Massachusetts Amherst, MIT CSAIL
CVPR 2017

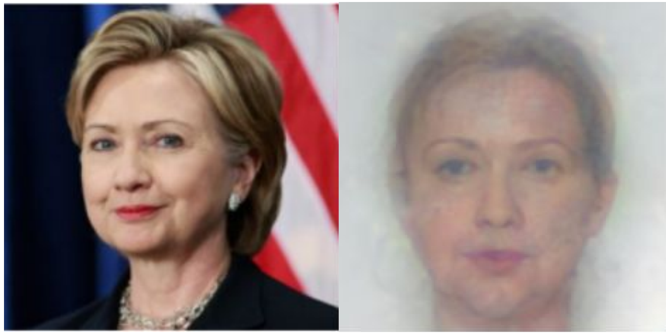
Presented by: Kapil Krishnakumar

Problem

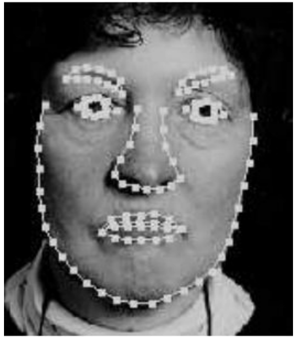
- Want method for synthesizing a frontal, neutral expression image of a person's face given an input face photograph
- One-to-one mapping from identity to image
- Method of pre-processing images to remove irregularities

Image Credit: Cole et al.

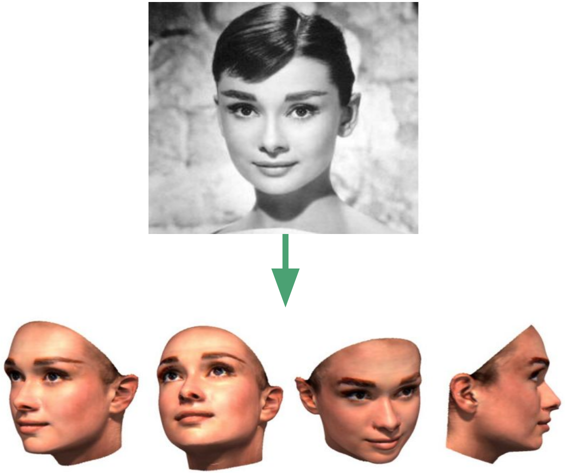
Related Work



Zhmoginov and Sandler et al.



Cootes et al.



Blanz and Vetter et al.

Hassner et al.

Image Credit: Zhmoginov and Sandler. Inverting face embeddings with convolutional neural works.
Blanz and Vetter et al. A Morphable Model For The Synthesis Of 3D Faces
Cootes et al. Active Appearance Models
Hassner et al. Effective Face Frontalization in Unconstrained Images

Approach

- Morphing of Images (Data Augmentation)
- Encoder (Image to Feature Vector)
- Decoder (Feature Vector to Normalized Image)
    - Landmarks
    - Texture

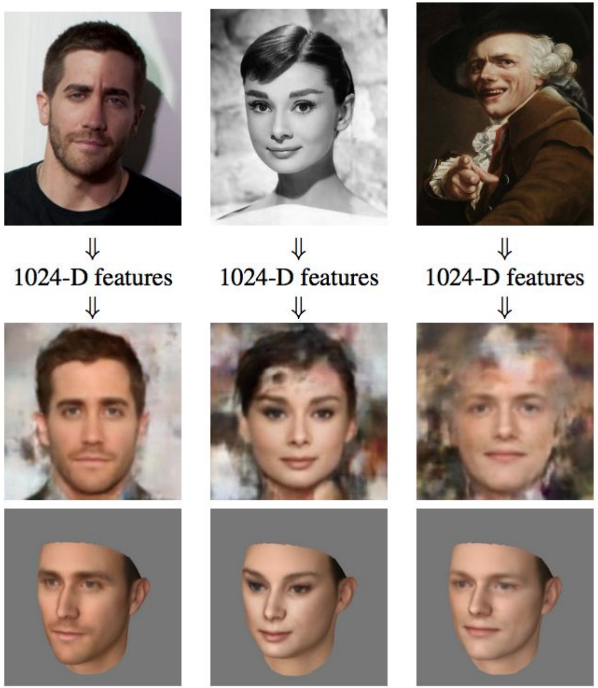


Image Credit: Cole et al.

Architecture

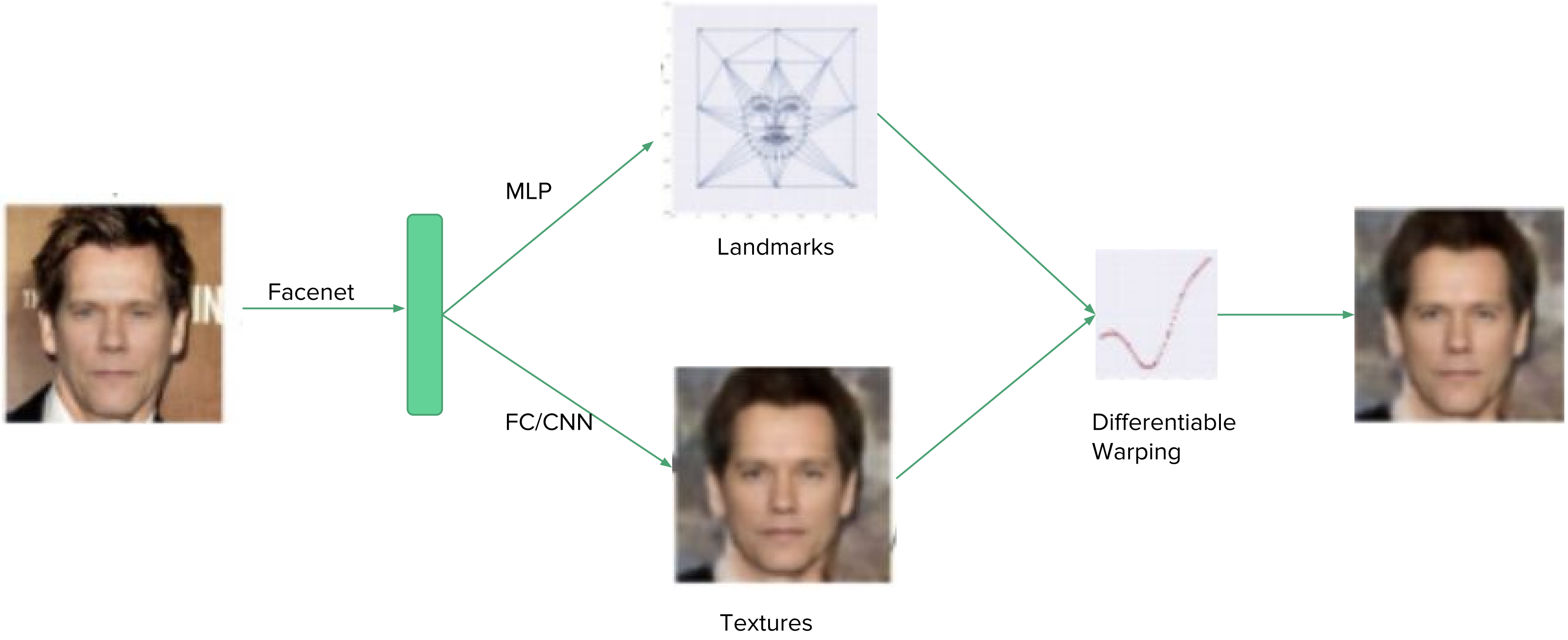
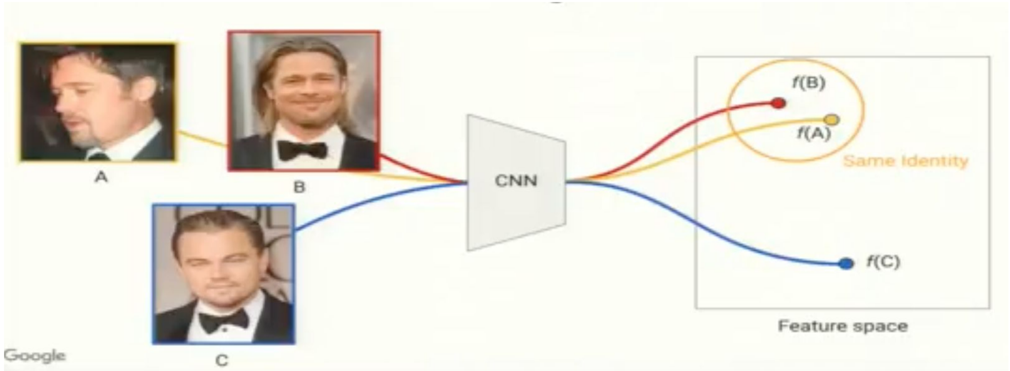


Image Credit: Cole et al.

FaceNet (Background) (Schroff et al. 2015)

- Face Images -> 128-D vectors
- Trained using triplet loss. Embeddings of two pictures of A should be more similar than picture of person A and person B.
- Uses GoogLeNet's NN2 Architecture



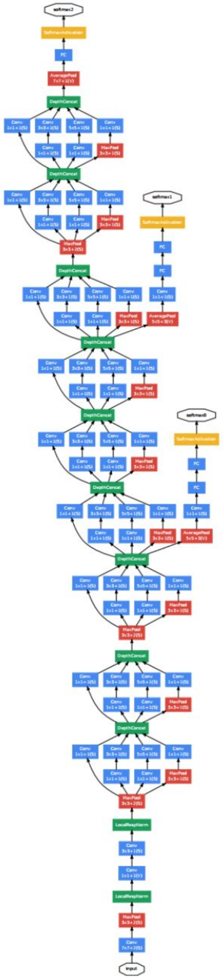


Image Credit: Cole CVPR 2017 Talk (https://www.youtube.com/watch?v=jVACIXpHgAI) | Szegedy et al. Going deeper with convolutions

Encoder

- Use pretrained FaceNet
- Extract 1024-D “avgpool” layer of “NN2” architecture
- Append and train Fully Connected Layer from 1024 to F dimensions on this layer.

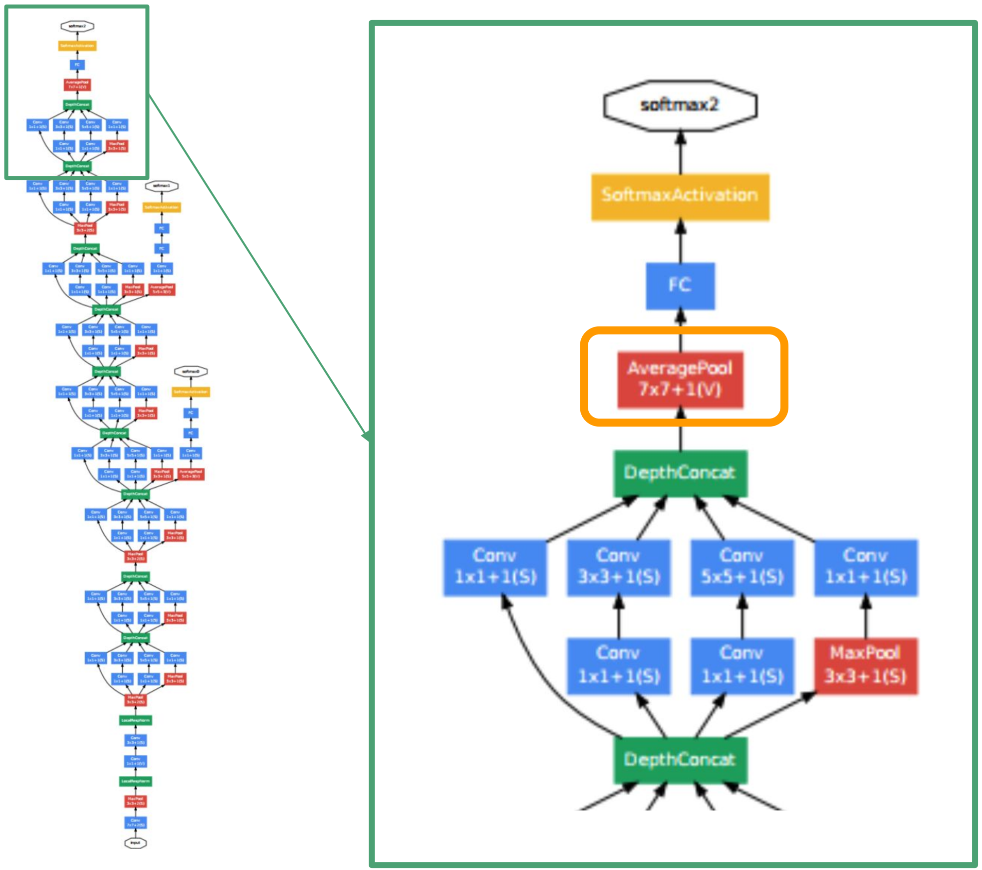


Image Credit: Szegedy et al. Going deeper with convolutions

Encoder

- Use pretrained FaceNet
- Extract 1024-D “avgpool” layer of “NN2” architecture
- Append and train Fully Connected Layer from 1024 to F dimensions on this layer.

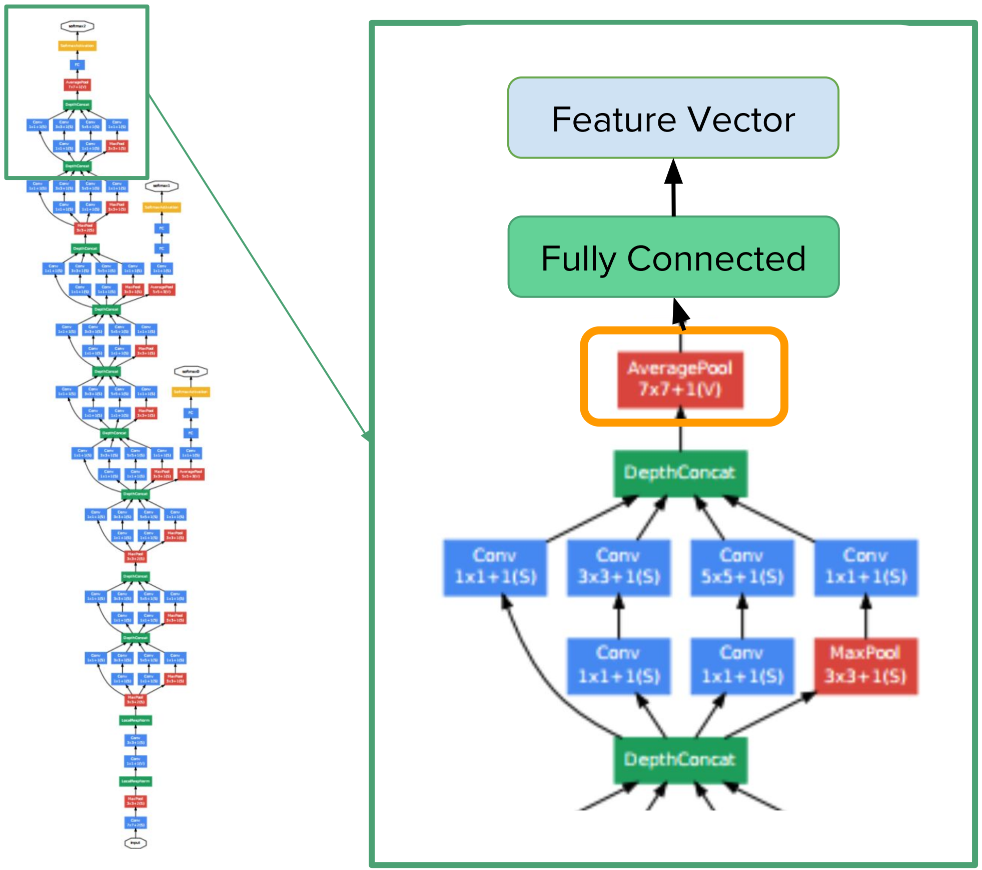


Image Credit: Szegedy et al. Going deeper with convolutions

Decoder

- Separating landmarks and textures more effective than just predicting image
- Landmarks estimated using shallow MLP with ReLUs applied on feature vector
  - FV -> [(x,y),.....]
- Textures estimated using fully connected or CNN
  - FV -> Image

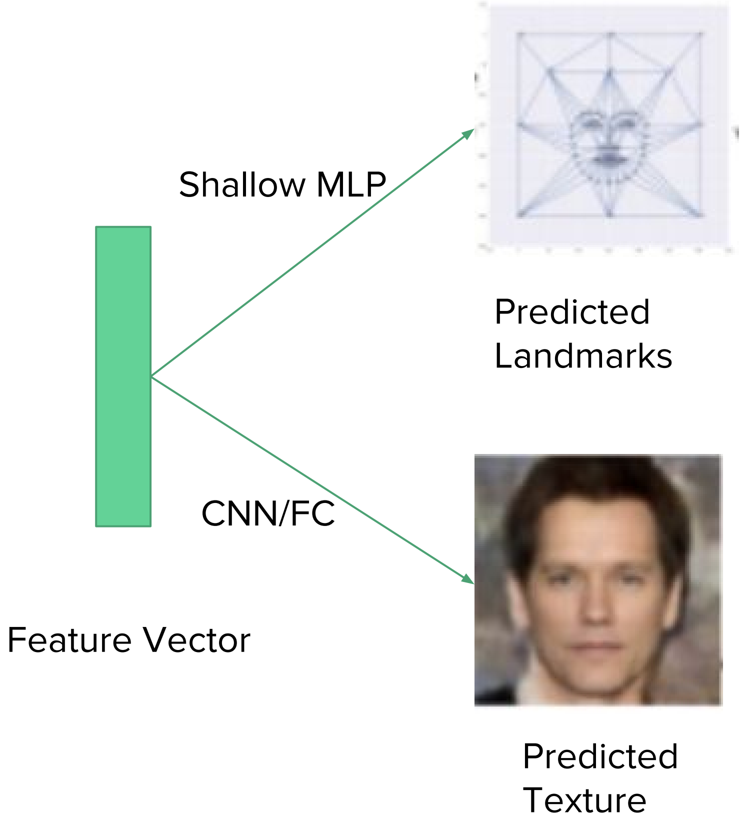


Image Credit: Cole et al

Decoder

- Use differentiable image warping to combine landmarks and textures

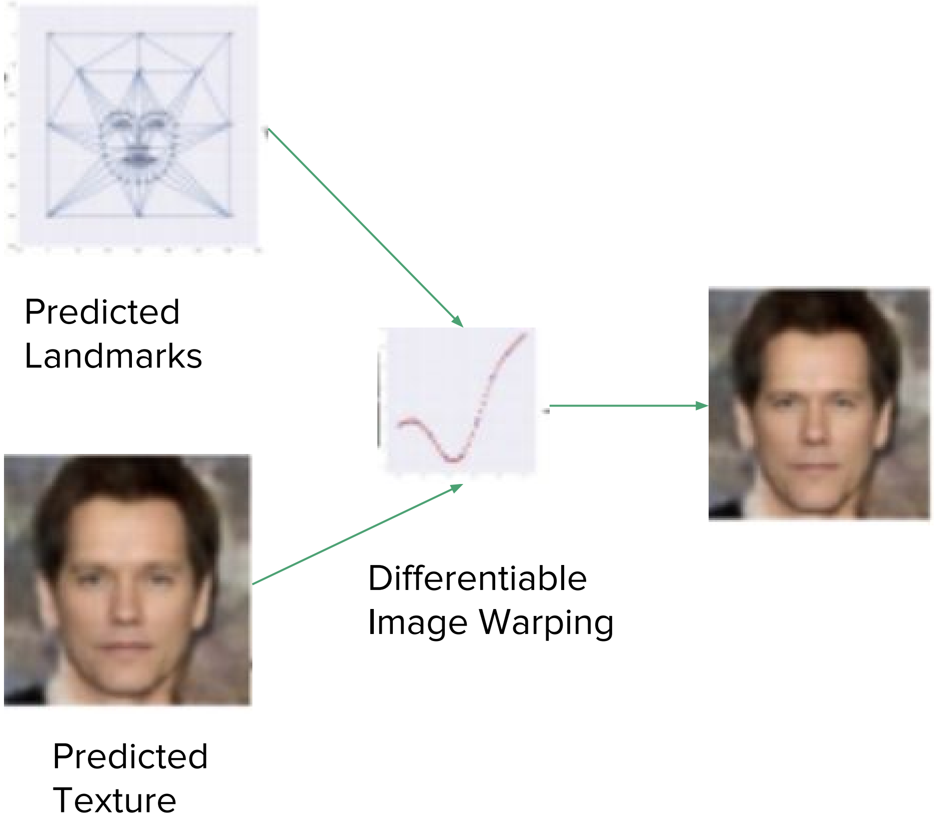


Image Credit: Cole et al

Decoder

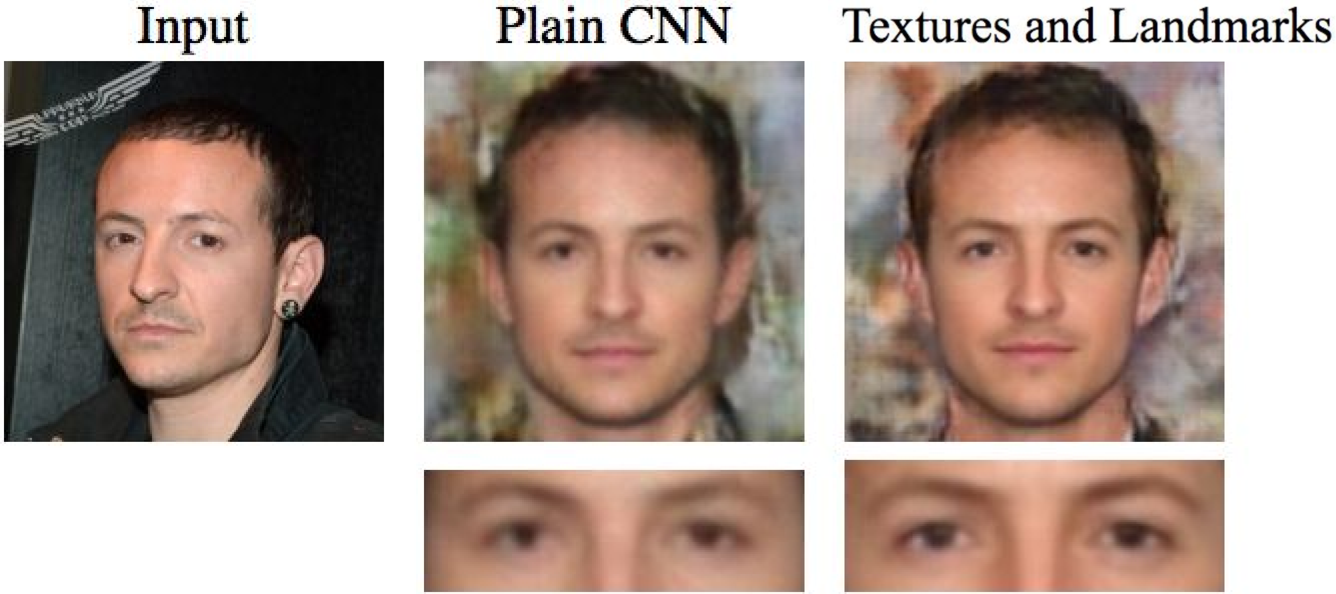


Image Credit: Cole et al

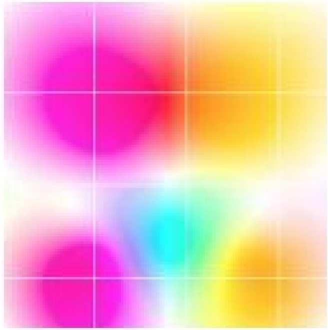
Differentiable Image Warping

Input Image
with Landmarks

Textures with
Landmarks

Final Landmarks

Mean
Landmarks of
training data



Dense Flow
Field with Spline
Interpolation

Final Output

Image Credit: Cole et al

Differentiable Spline Interpolation

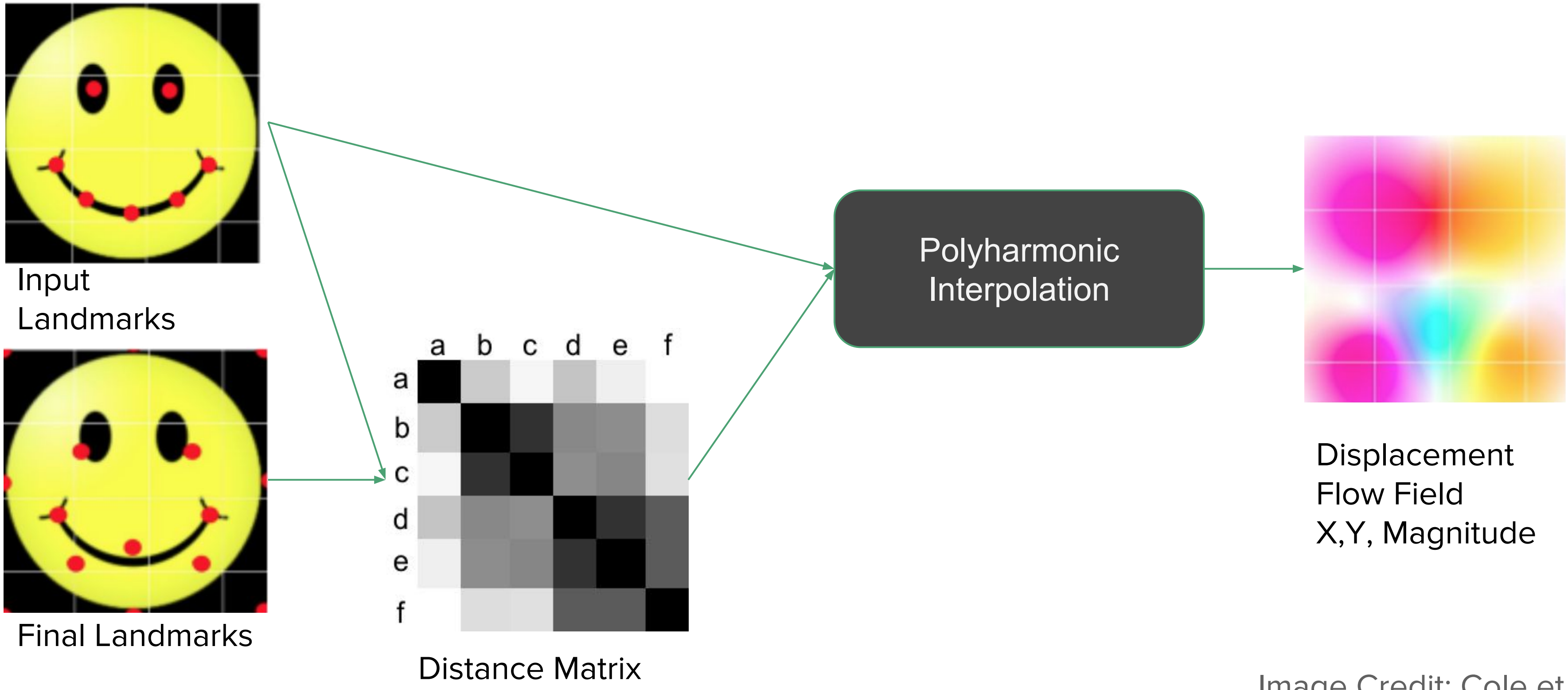


Image Credit: Cole et al

Training

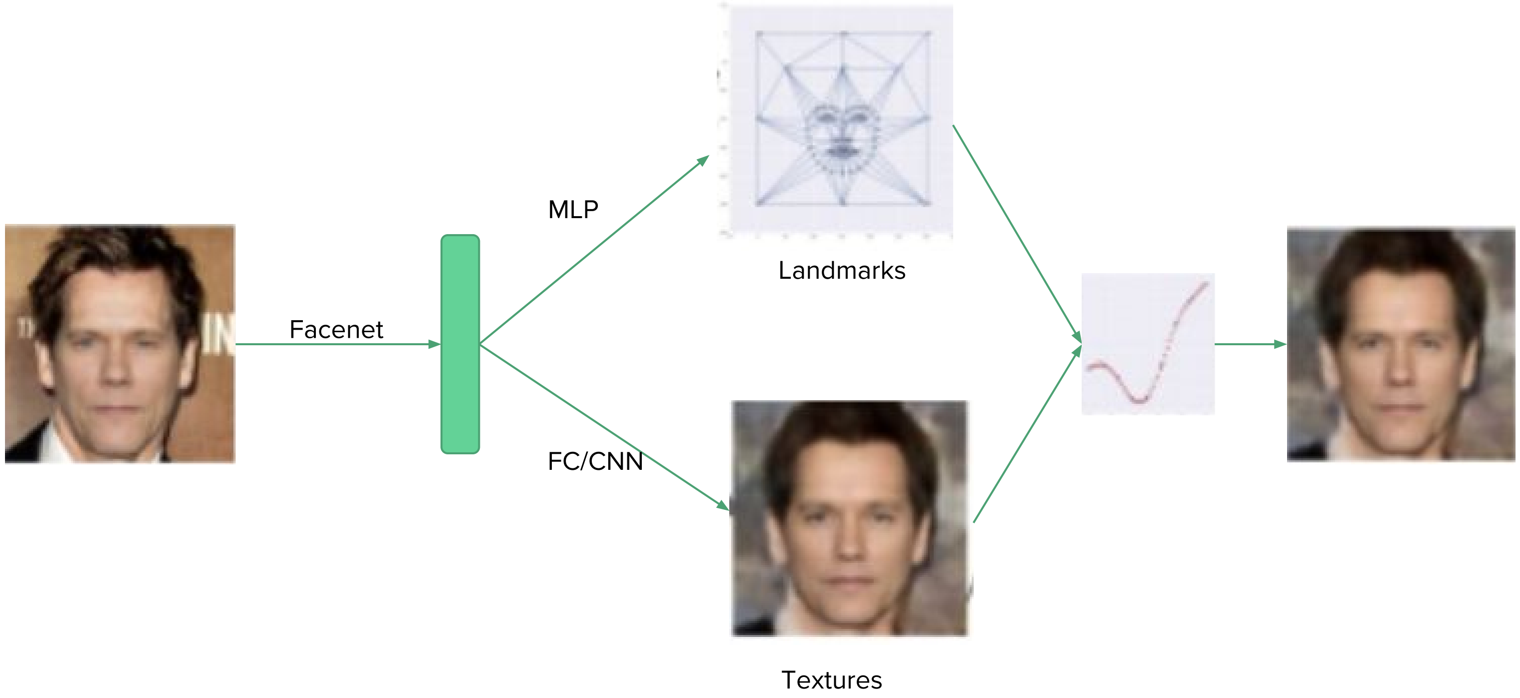


Image Credit: Cole et al.

Training

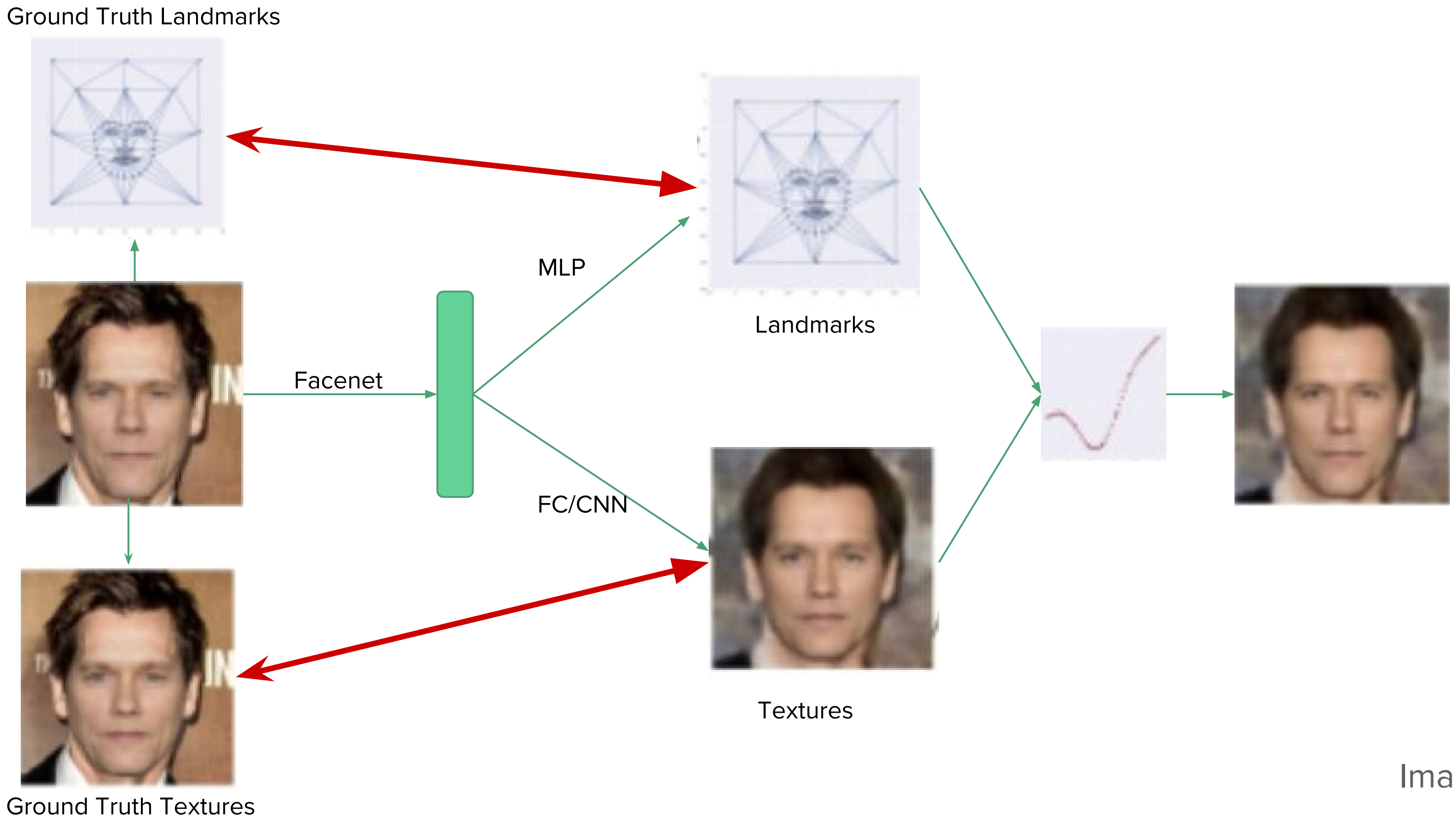


Image Credit: Cole et al.

Training with FaceNet Loss

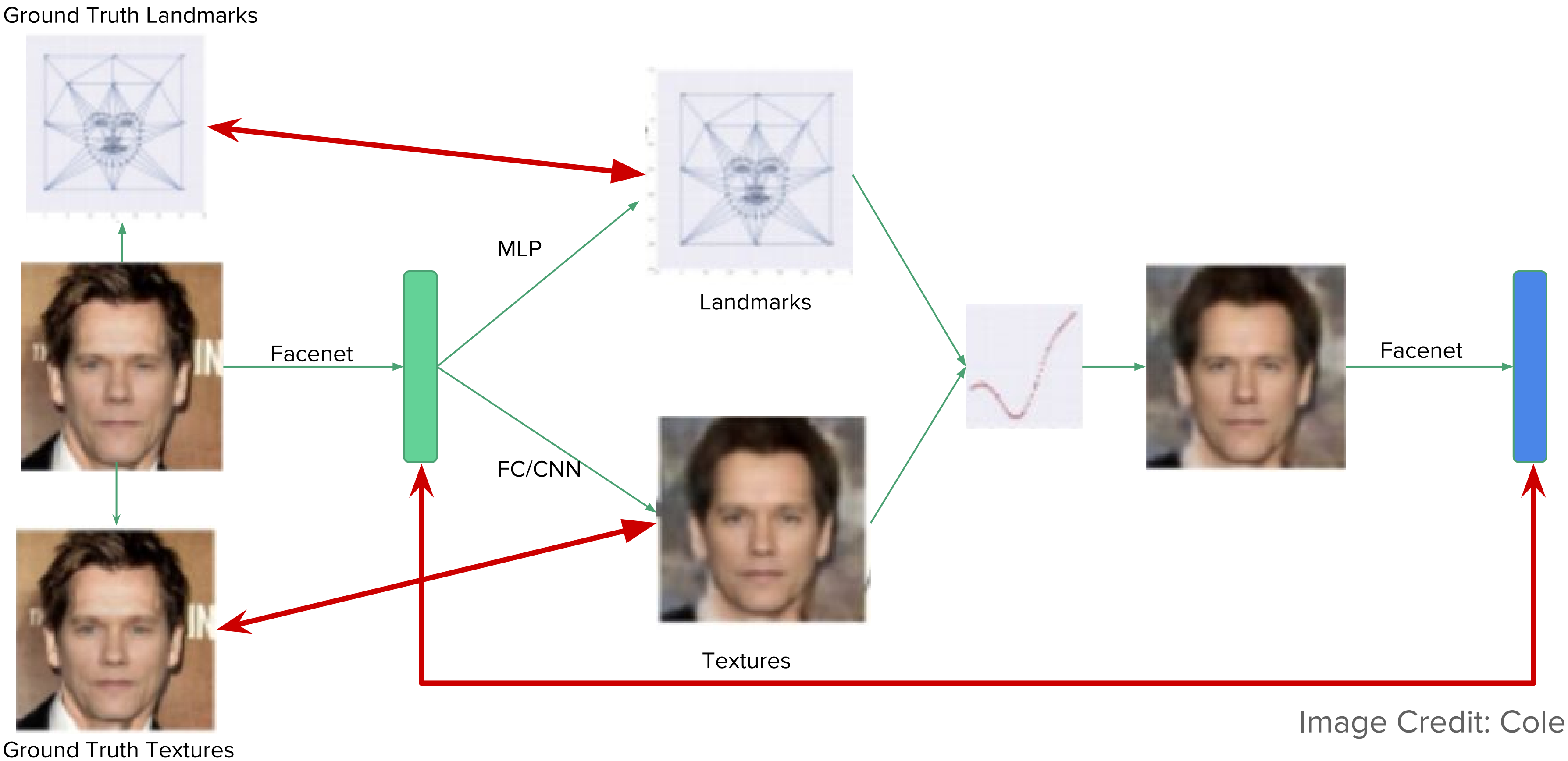


Image Credit: Cole et al.

Training Loss

- Separately penalize predicted landmarks and textures using mean squared error
- Penalize differences in resulting encodings from input image and rendered image when passed through FaceNet
    - Highly expensive to train

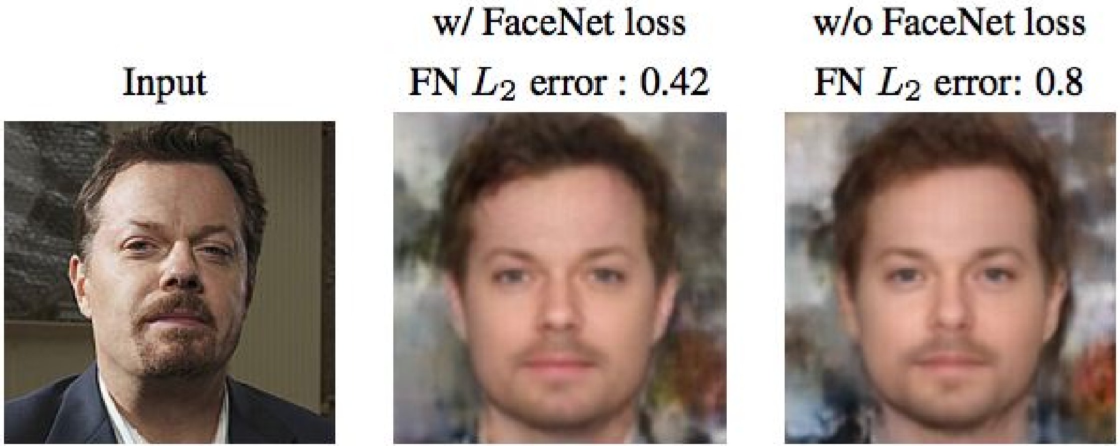
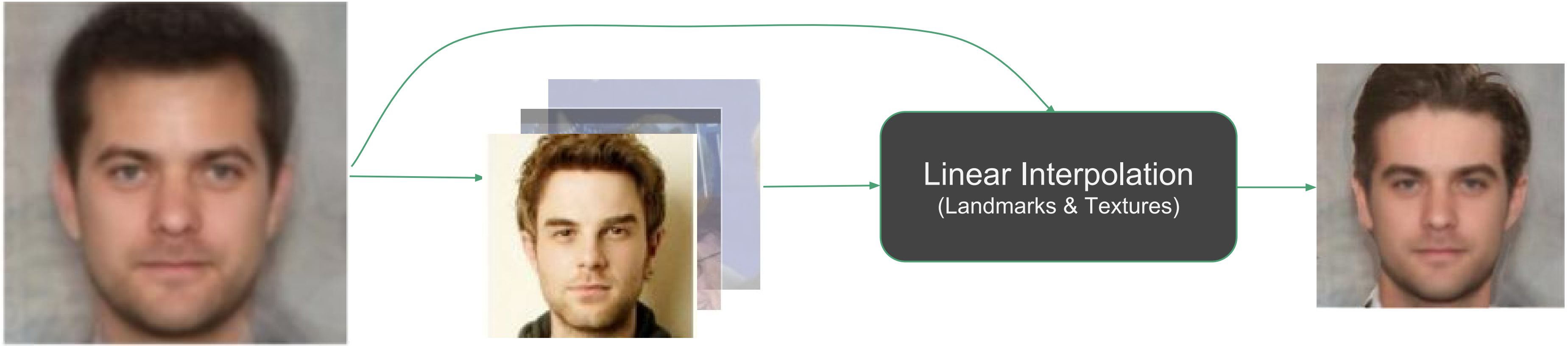


Image Credit: Cole et al

Data Augmentation: Random Morphs

- Problem: Don’t have database of normalized face photos to train decoder network on
- Solution: Morphing Data Augmentation



Select one of k=200 Nearest Neighbors using distance defined by Landmarks and Textures

Image Credit: Cole et al

Data Augmentation: Gradient Domain Compositing

- Morphing cannot capture hair and background detail
- Combine morphed image onto an original background using gradient domain compositing

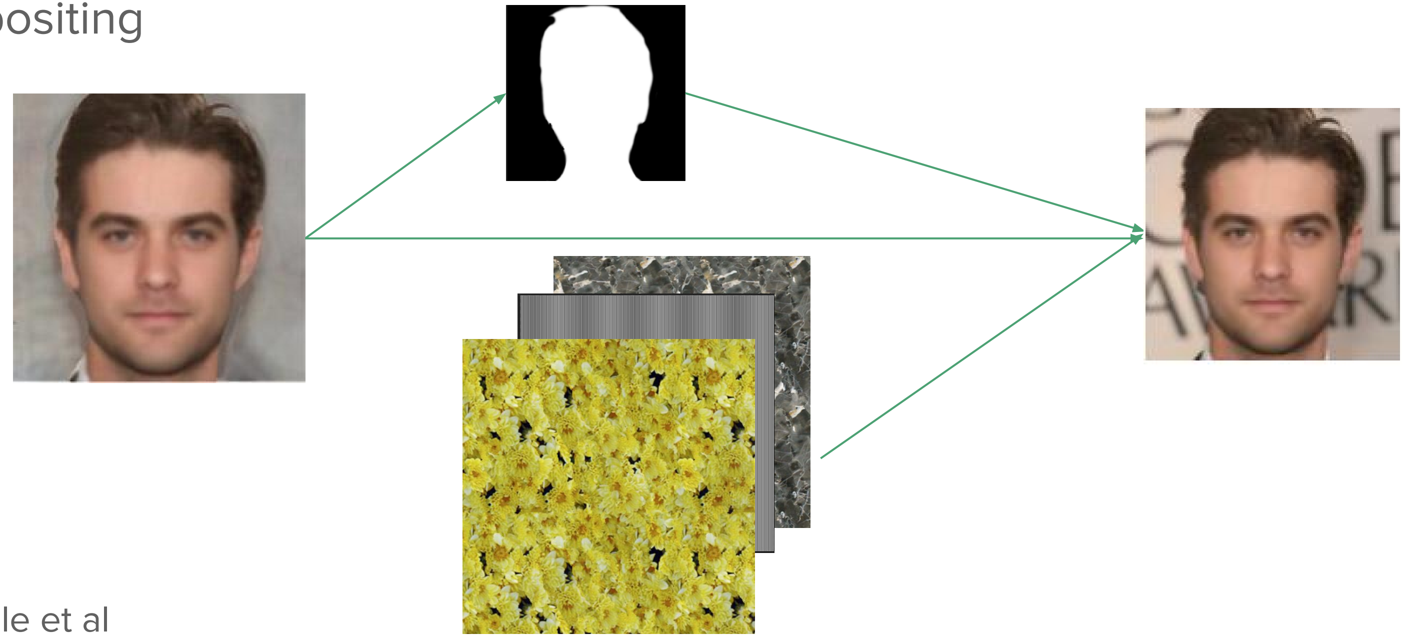
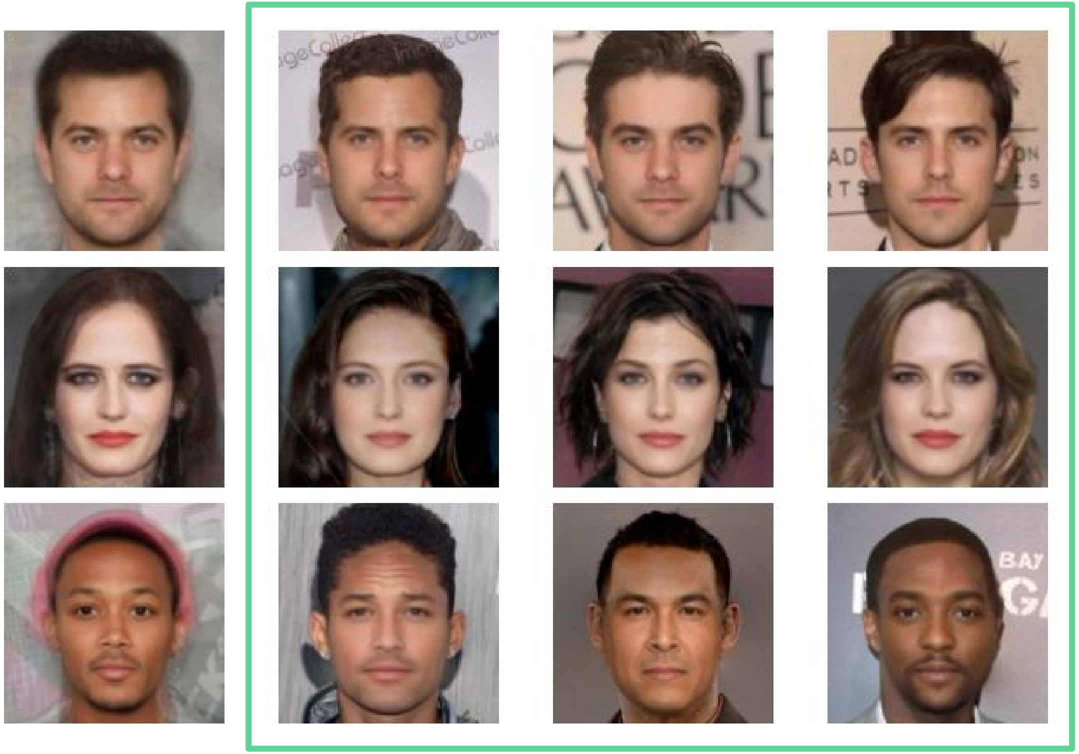


Image Credit: Cole et al

Data Augmentation



Input

Augmented

Image Credit: Cole et al

Data Augmentation

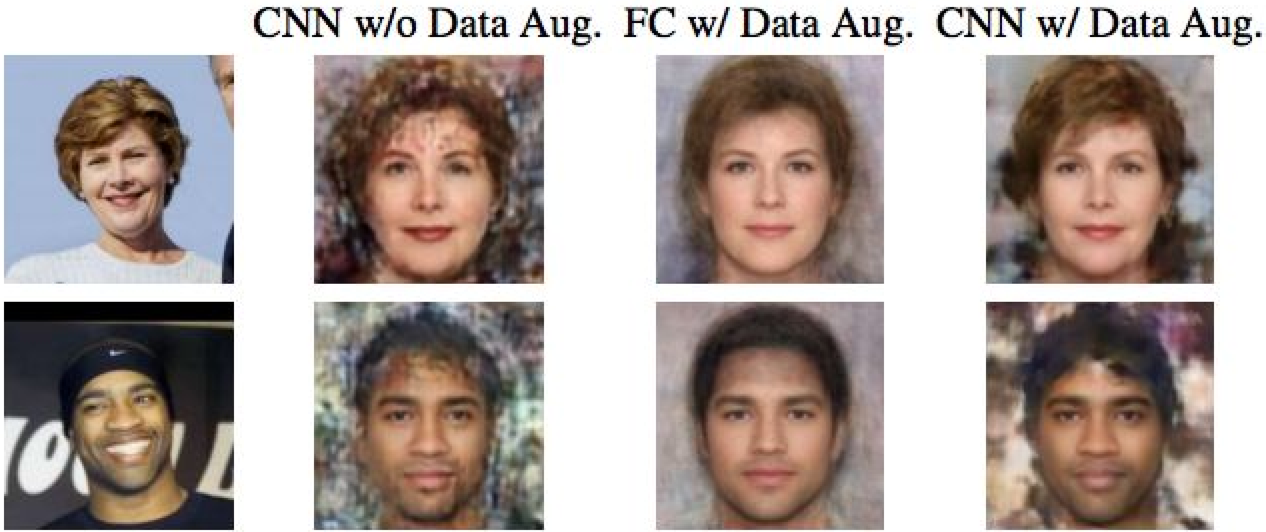
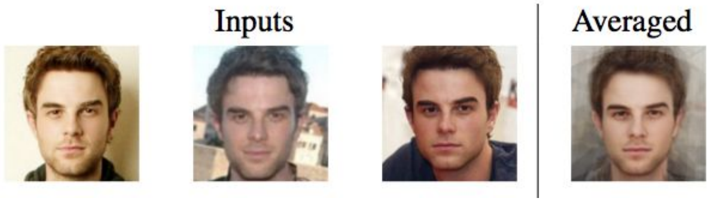


Image Credit: Cole et al

Training Data



- Dataset used to train VGG-Face network. 2.6M photos
- Processing:
  - Average all images for each individual by morphing
  - Each image is then warped to average landmarks of individual
  - Pixel values are averaged to form average image of individual.
- Gives 1K unique identities images
- Use Kazemi and Sullivan for extracting groundtruth Landmarks
- Augmentation produces 1M images

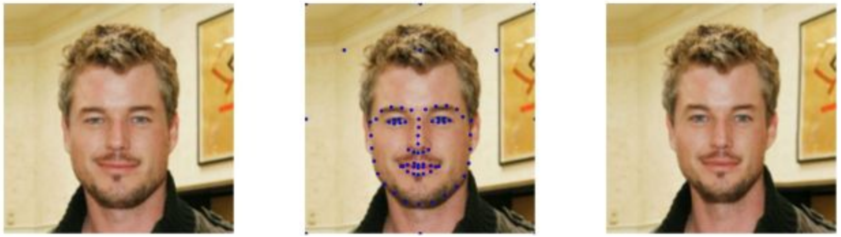


Image Credit: Cole et al

Experiments: Labeled Faces in the Wild

- Identities mutually exclusive to VGG face dataset

Image Credit: Cole et al

Experiments: Labeled Faces in the Wild

- Histograms of FaceNet L2 error between input and synthesized images.
- 1.242 is threshold for clustering identities in FaceNet feature space
- **Blue**: With Facenet Training Loss
- **Green**: Without Facenet Training Loss

Image Credit: Cole et al

Robustness to Occlusions

Image Credit: Cole CVPR 2017 Talk (https://www.youtube.com/watch?v=jVACIXpHgAI)

Extensions: 3-D Model Fitting

- Easier to fit normalized face image on 3D morphable model.

Image Credit: Cole et al

Extensions: Automatic-Photo Adjustment

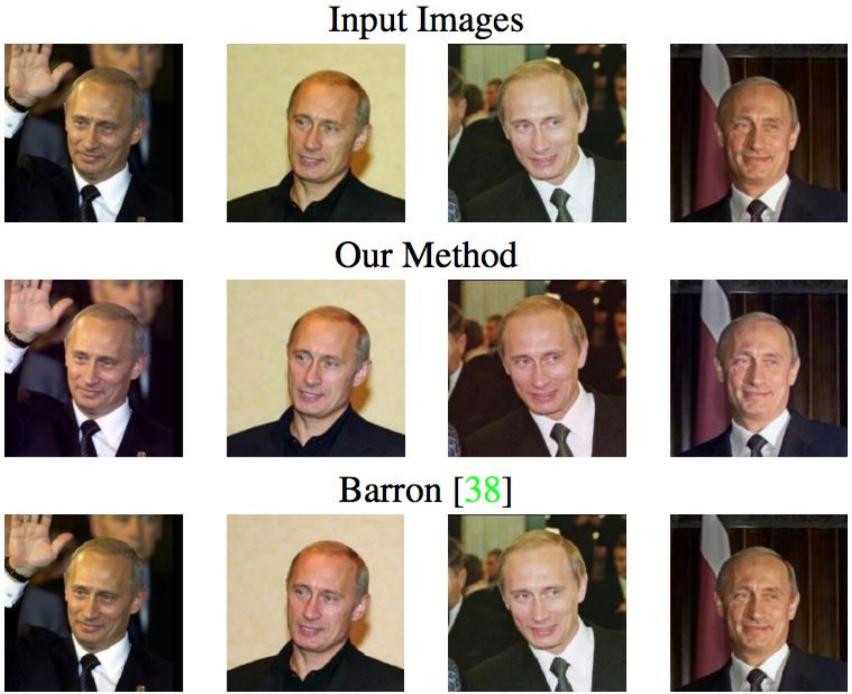


Image Credit: Cole et al

Extensions: Automatic-Photo Adjustment

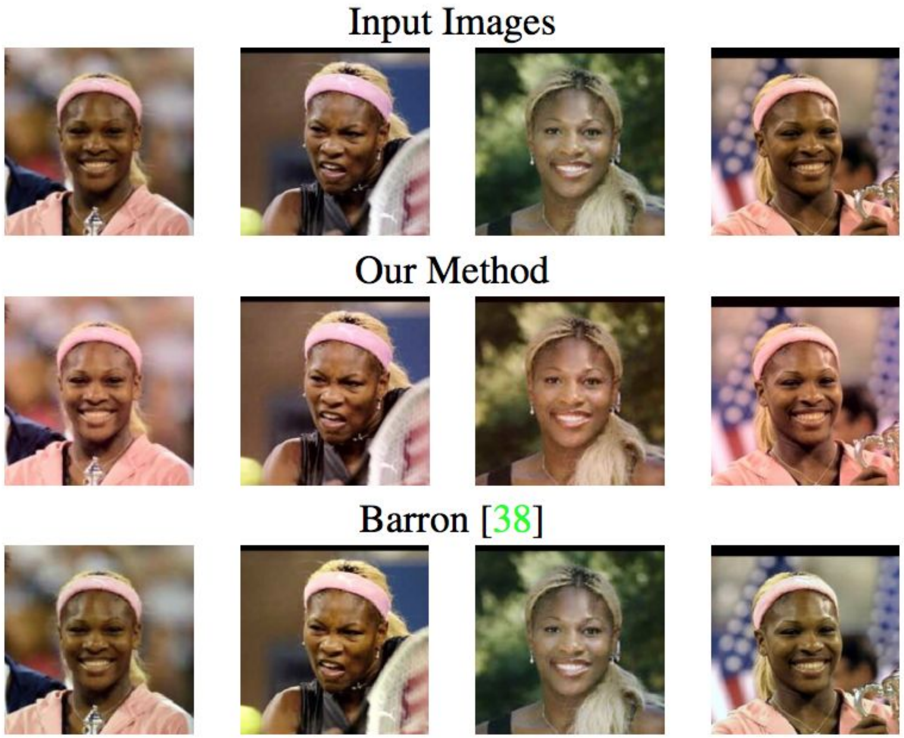


Image Credit: Cole et al

Advantages

- Splitting of generative tasks (Landmarks and Textures) can be better than directly outputting result
- Fresh use of spline interpolation as differentiable module in NN
- Augmentation technique allows training of decoder with only 1K images to perform extremely well.
- Tough features like hair and eyes are well defined in normalized images
- Robustness to occlusions

Disadvantages

- No “ground truth” to compare Normalized Images
  - Though measure of performance can be defined as FaceNet closeness between image and normalized image
  - Cannot get human annotated ground truth
- Dependent on out of box methods for getting Landmarks and Textures labels
  - Paper doesn’t show experiments on other techniques other than Kazemi
  - Unclear on how Texture labels are generated.
- Backgrounds are unrealistic and blurry