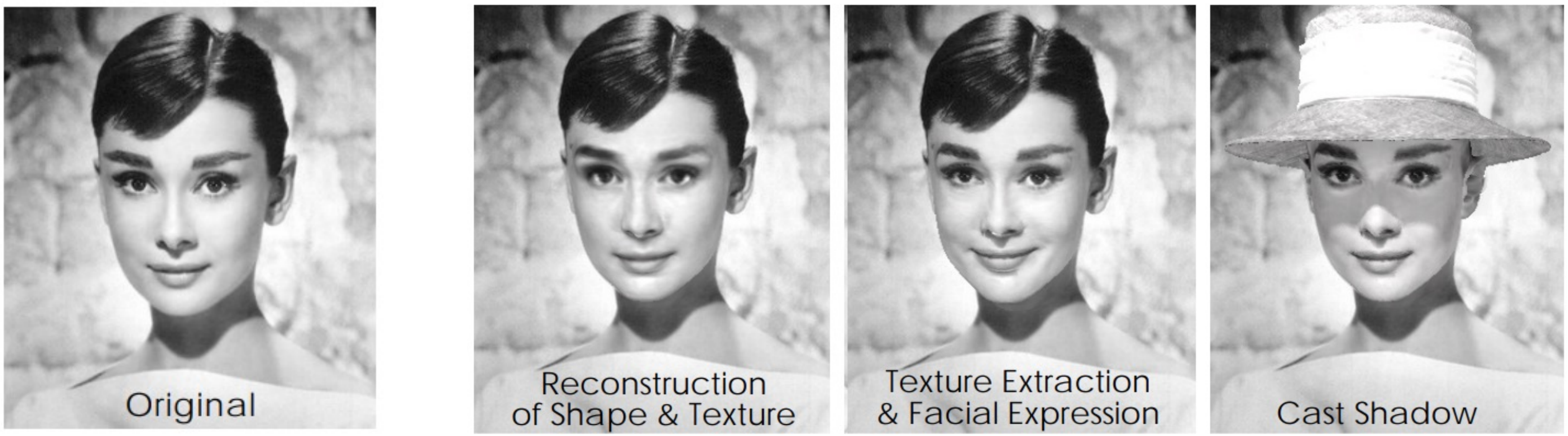


Face modeling

Jun-Yan Zhu

16-726 Learning-based Image Synthesis, Spring 2023

Slides adopted from Alexei A. Efros, Hyeongwoo Kim et al.

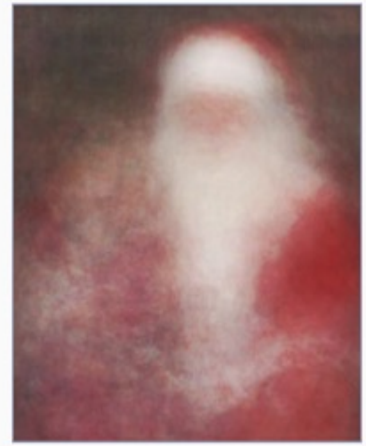
Why Human Faces?

- Face is an important subject.
  - We are humans.
  - Many commercial applications.
- Lots of useful tools
  - 3D data: geometry-based synthesis.
  - 2D/3D Computer vision works for faces.

© Weise et al., SIGGRAPH 2011

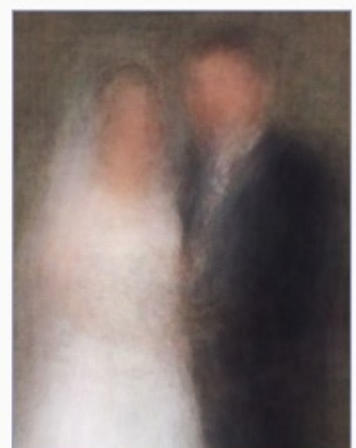
“100 Special Moments” by Jason Salavon

Little Leaguer



Kids with Santa

The Graduate

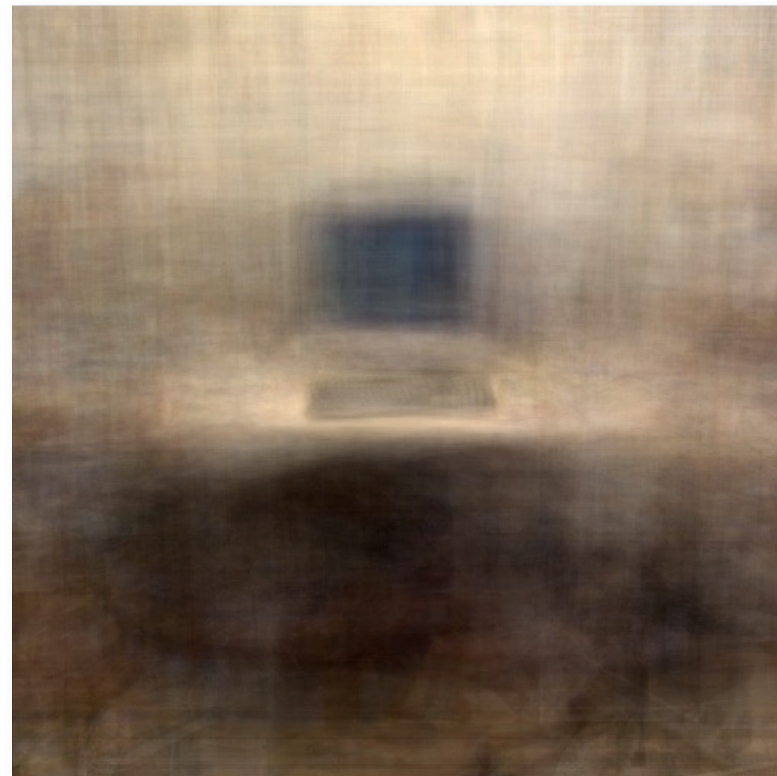


Newlyweds

Why blurry?

Object-Centric Averages by Torralba (2001)

Manual Annotation and Alignment



Average Image

Computing Means

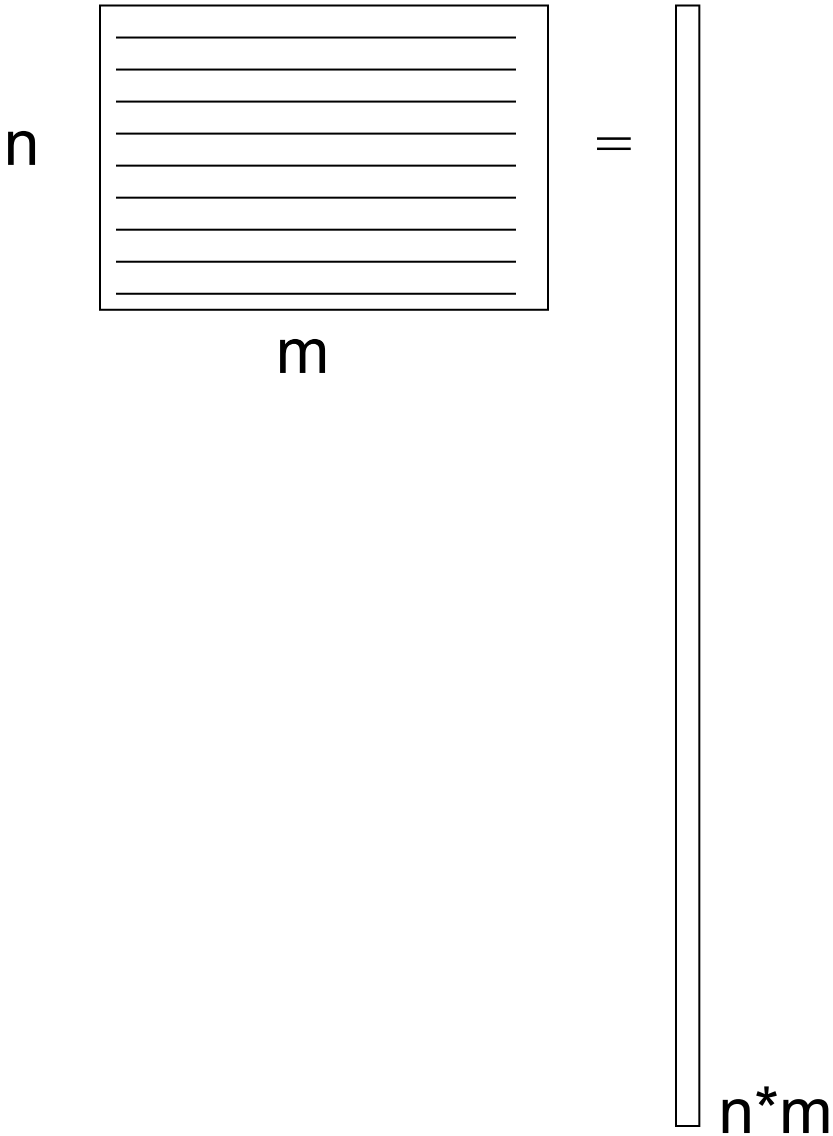
Two Requirements:

- Alignment of objects
- Objects must span a subspace

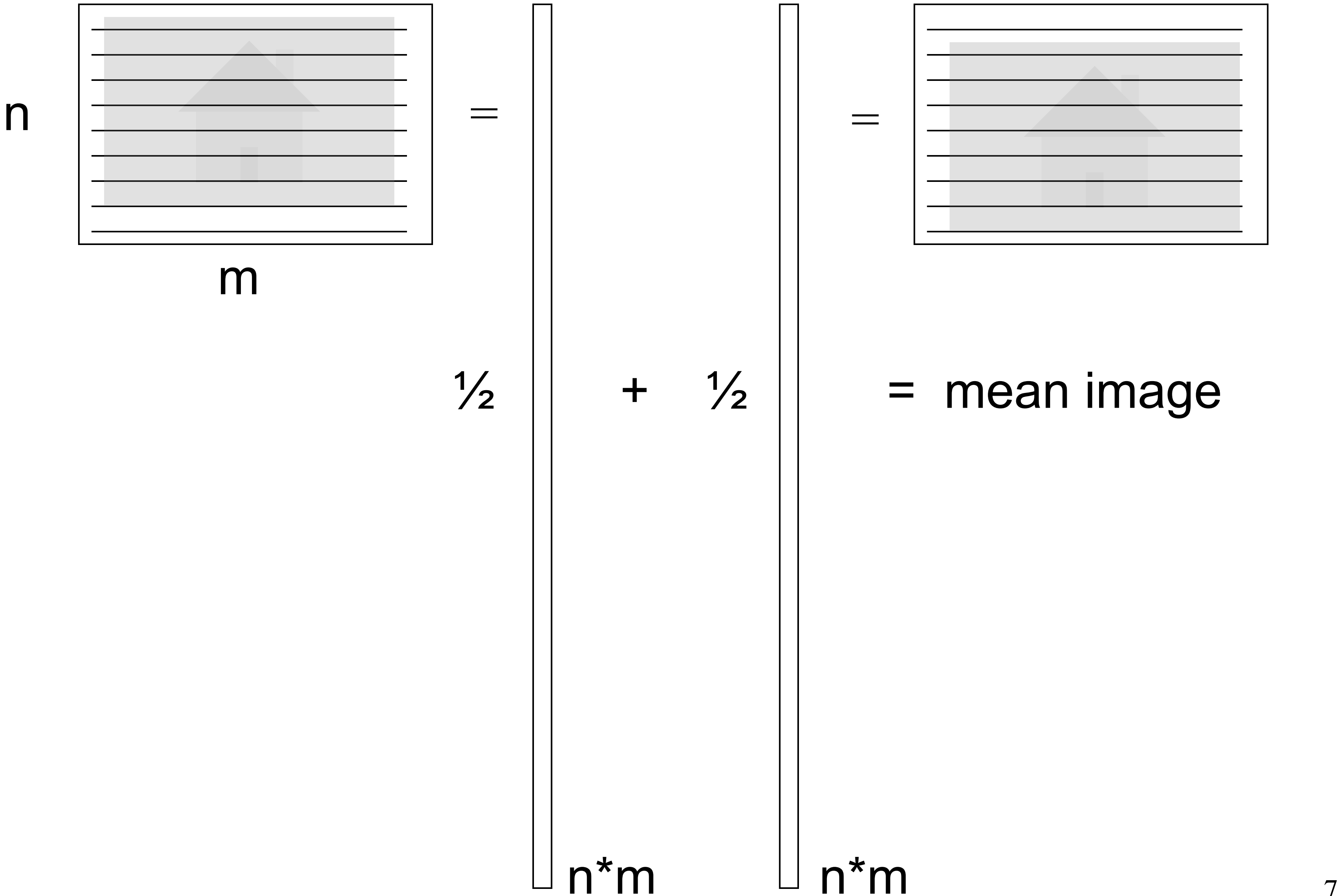
Useful concepts:

- Subpopulation means
- Deviations from the mean

Images as Vectors



Vector Mean: Importance of Alignment

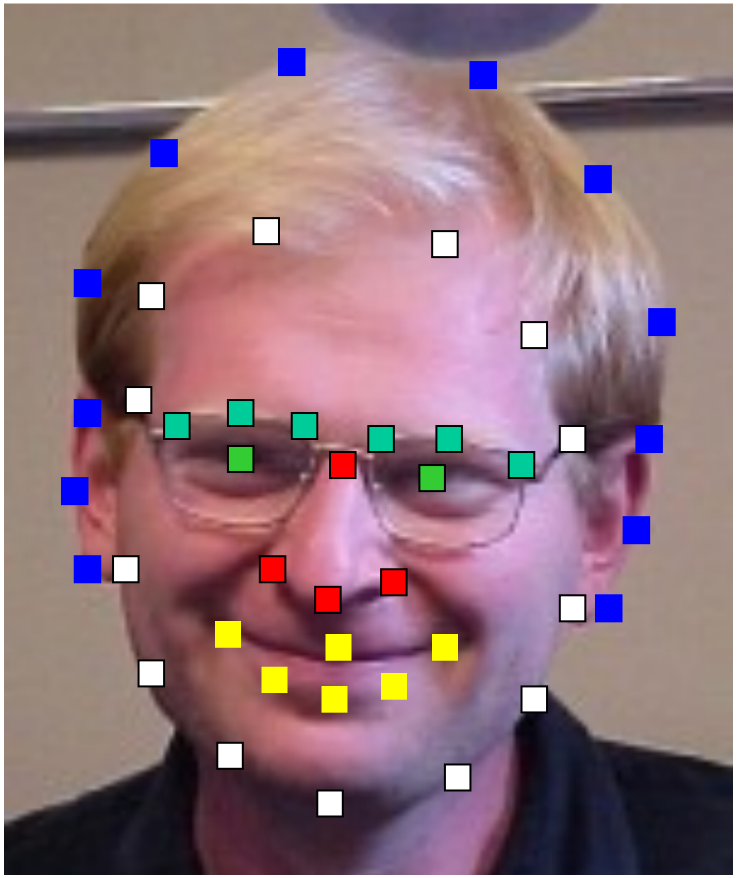


How to align faces?

Students and staff from Technical University of Denmark

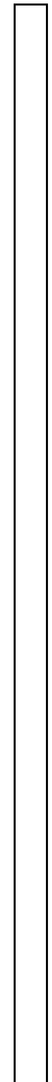
http://www2.imm.dtu.dk/~aam/datasets/datasets.html

Shape Vector



Landmark annotation

=



43

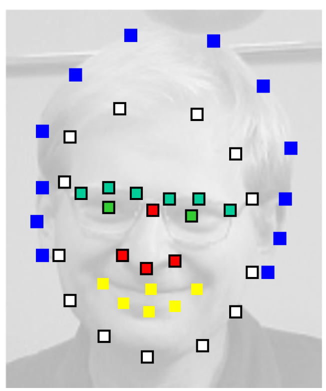
Appearance Vectors vs. Shape Vectors

Appearance Vector

200*150 pixels (RGB)

Vector of 200*150*3 Dimensions

Shape Vector



43 coordinates (x,y)

Vector of 43*2 Dimensions

- Manual annotation.

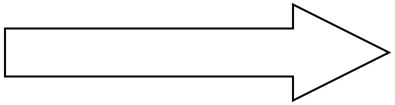
OR

- Face landmark detection.

Slide by Kevin Karsch

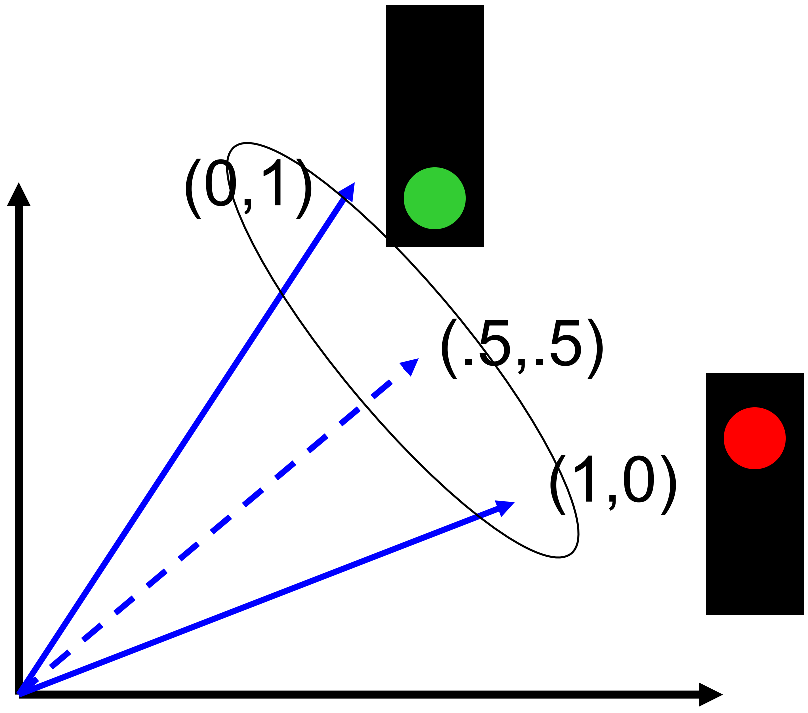
Average Face

1. Warp to mean shape
2. Average pixels



Students and staff from Technical University of Denmark
http://graphics.cs.cmu.edu/courses/15-463/2004_fall/www/handins/brh/final/

Objects must span a subspace



Subpopulation means

Examples:

- Male vs. female
- Happy vs. said
- Average Kids
- Happy Males
- Etc.
- http://www.faceresearch.org

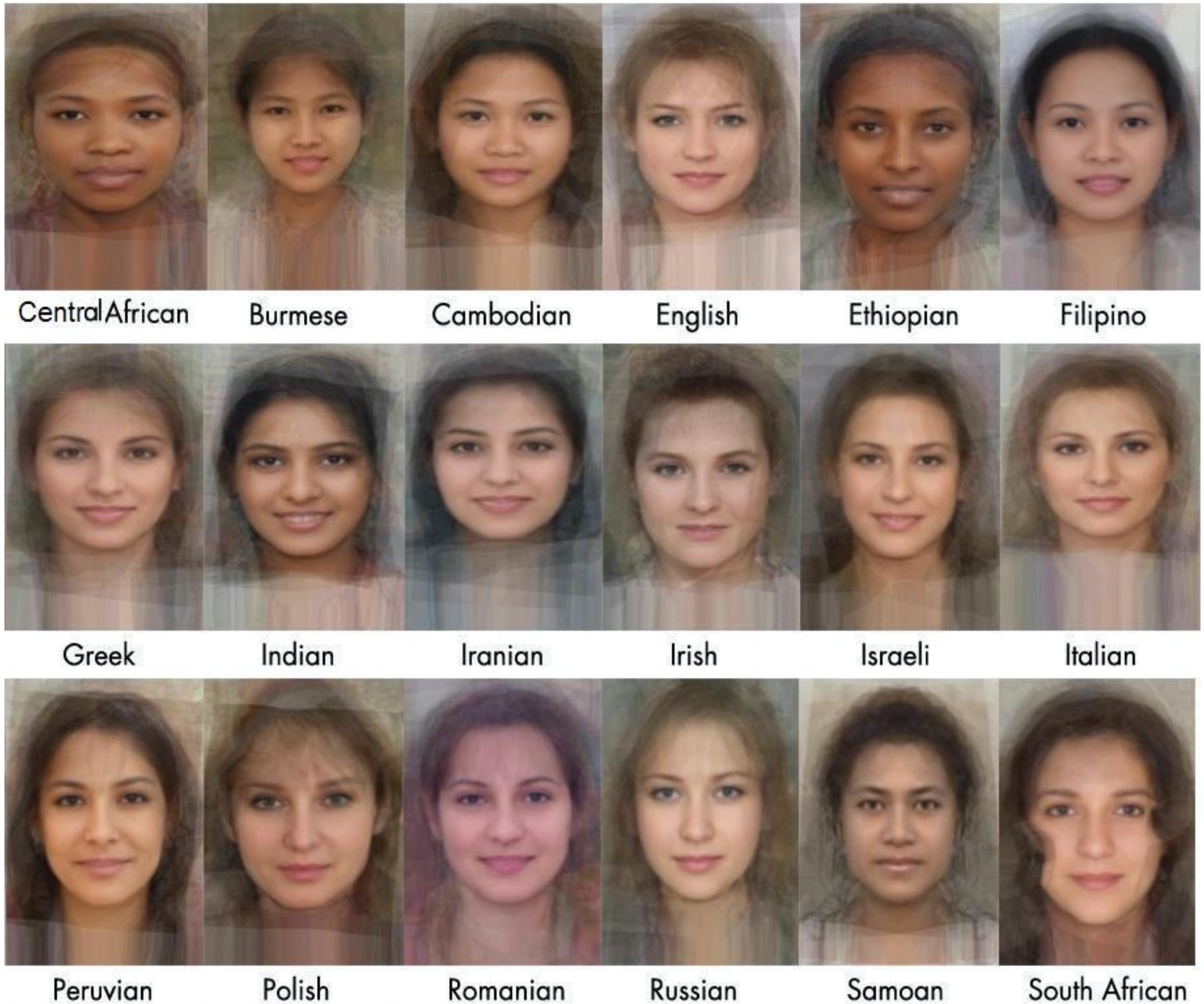
Average female

Average kid

Average happy male

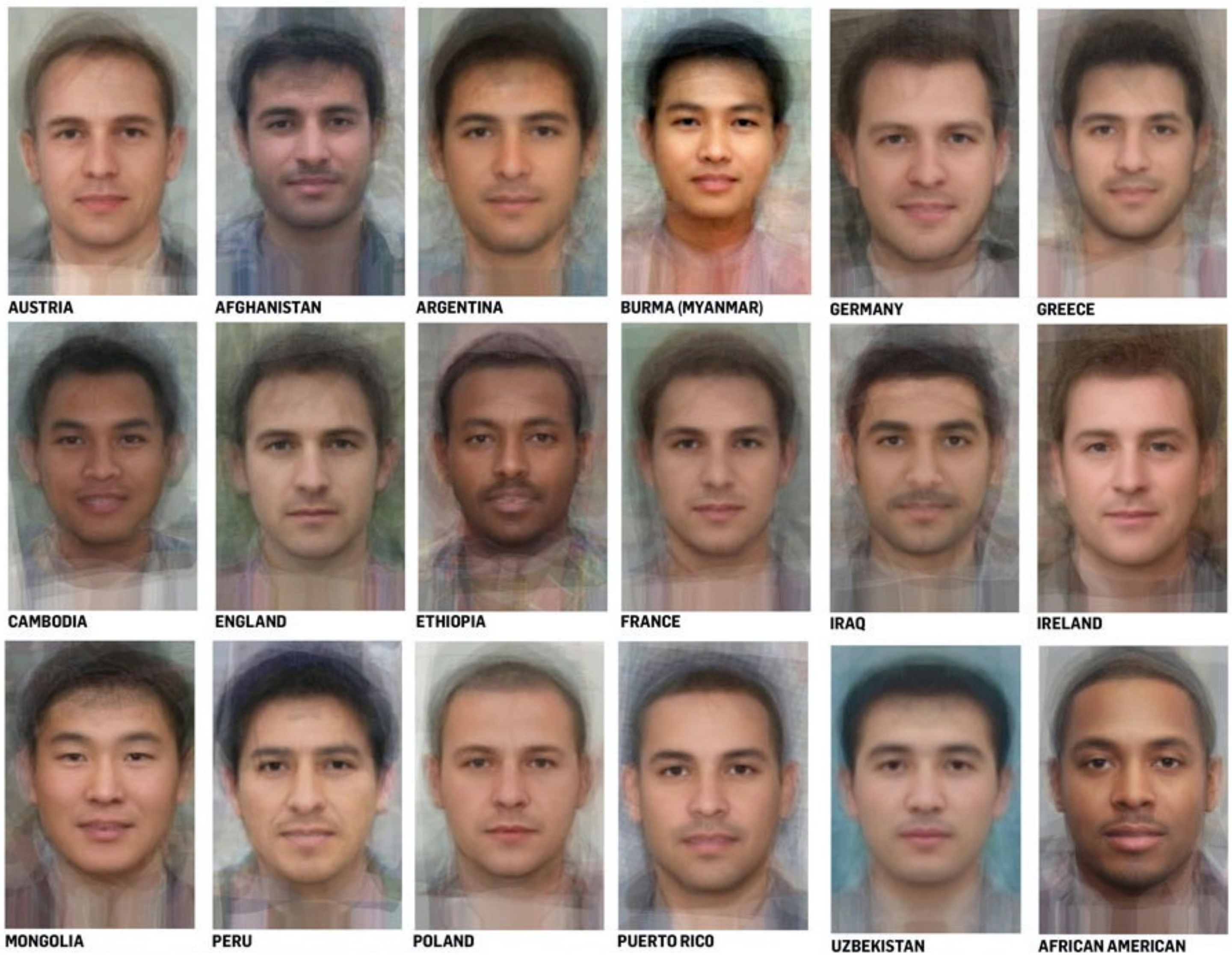
Average male

Average Women of the world



Several issues: 1. country ≠ race. 2. demographic diversity is lost. 3. bias in data source

Average Men of the world



Several issues: 1. country $\neq$ race. 2. demographic diversity is lost. 3. bias in data source

Deviations from the mean

Image X

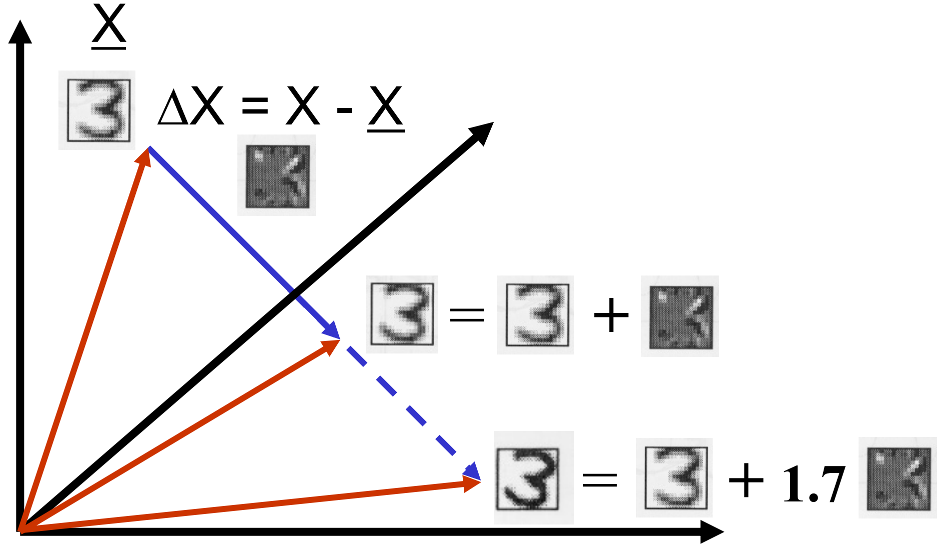
\-

Mean $\underline{X}$

=

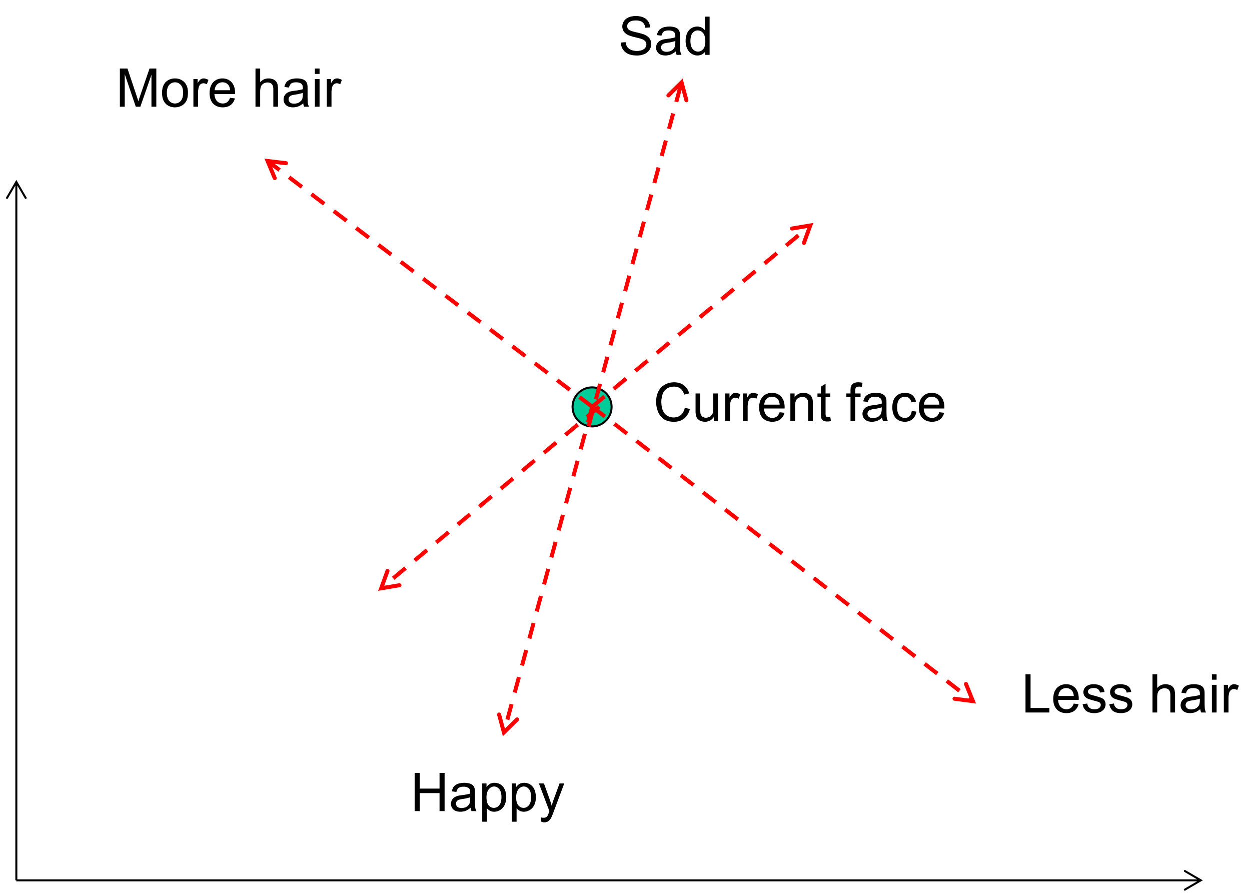
$\Delta X = X - \underline{X}$

Deviations from the mean



Extrapolating faces

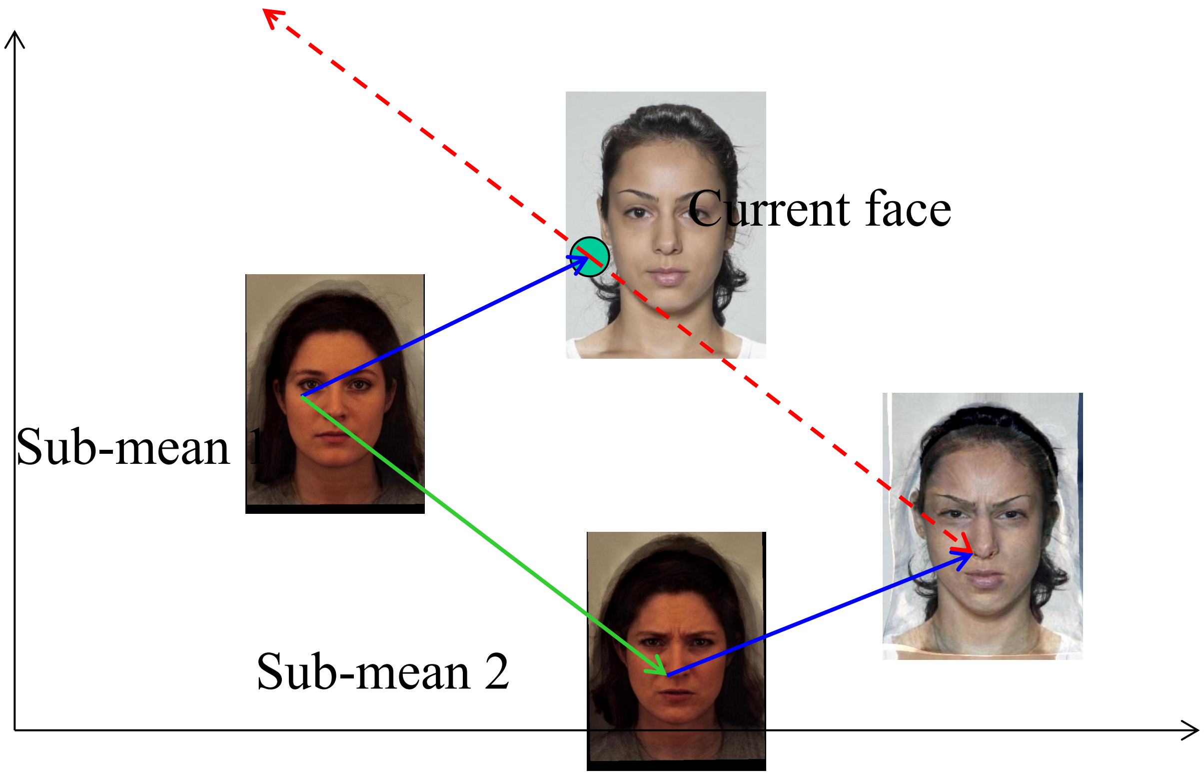
- We can imagine various meaningful <u>directions</u>.



Slide by Kevin Karsch

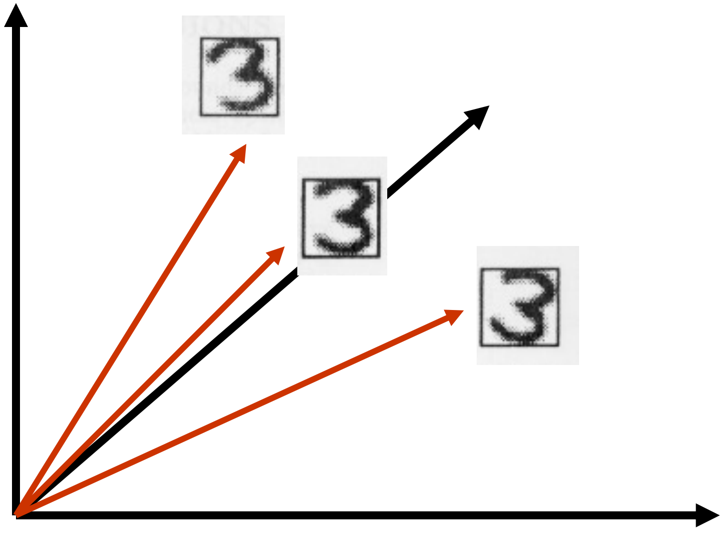
Manipulating faces

- How can we make a face look younger/older, or happy/sad, etc.?
- http://www.faceresearch.org/demos/transform

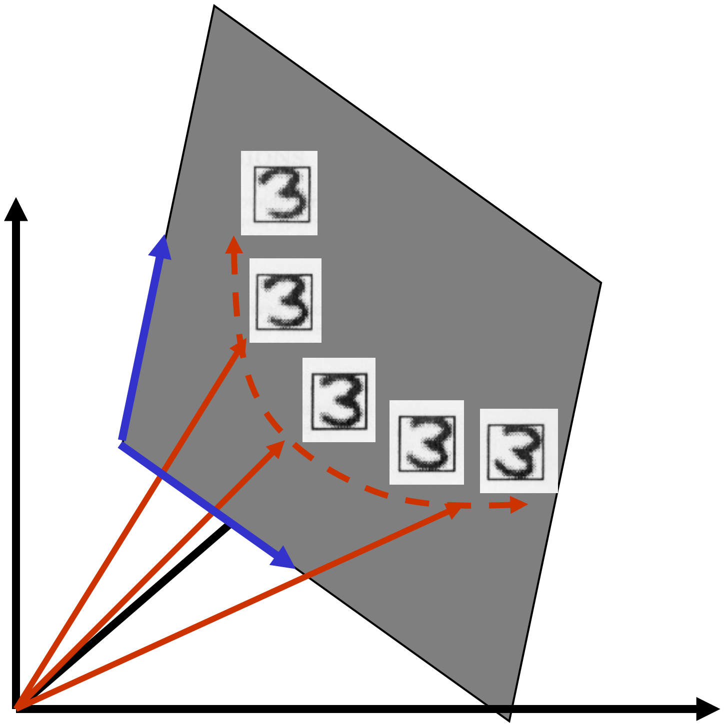


Slide by Kevin Karsch

Back to the Subspace



Linear Subspace: convex combinations



Any new image X can be obtained as weighted sum of stored “basis” images.

$$X = \sum_{i=1}^{m} a_i X_i$$

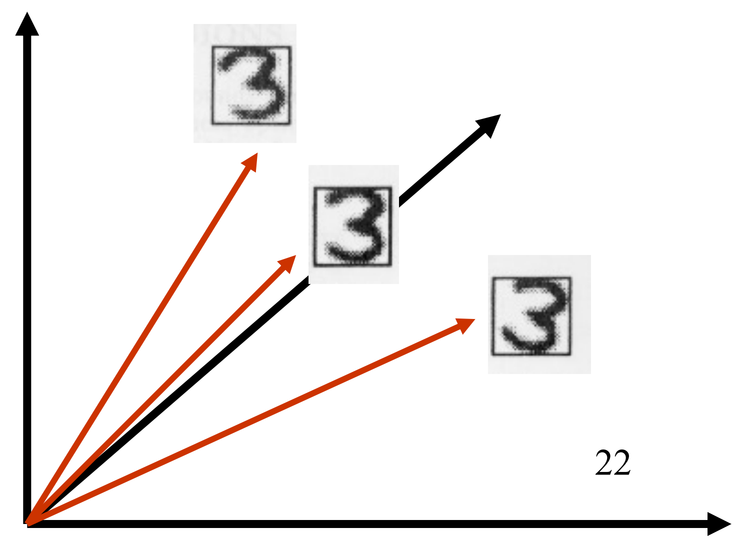
Our old friend, change of basis! What are the new coordinates of X?

Issues:

1. How many basis images is enough?
2. Which ones should they be?
3. What if some variations are more important than others?
   - E.g. corners of mouth carry much more information than haircut

Need a way to obtain basis images automatically, in order of importance!

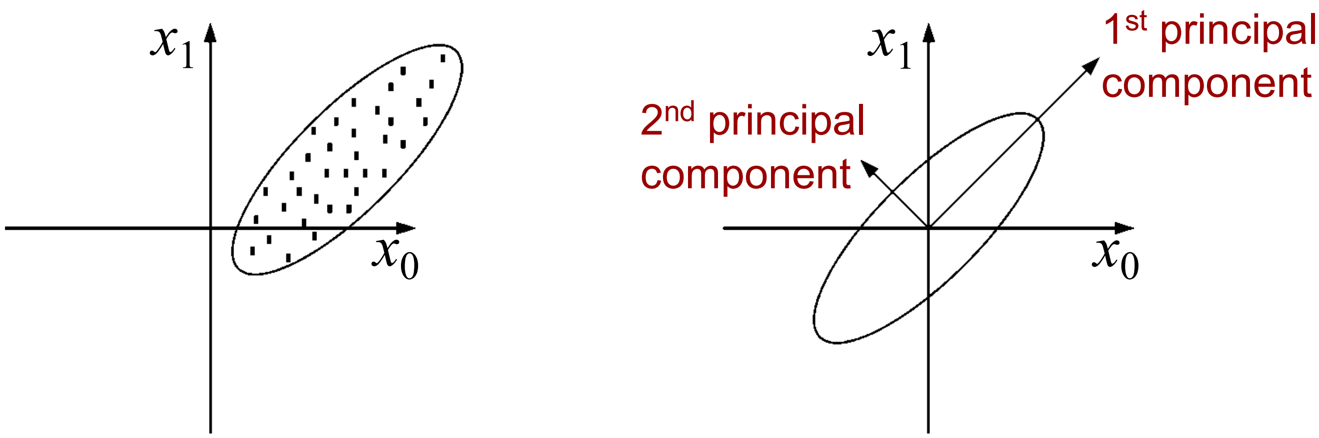
But what's important?



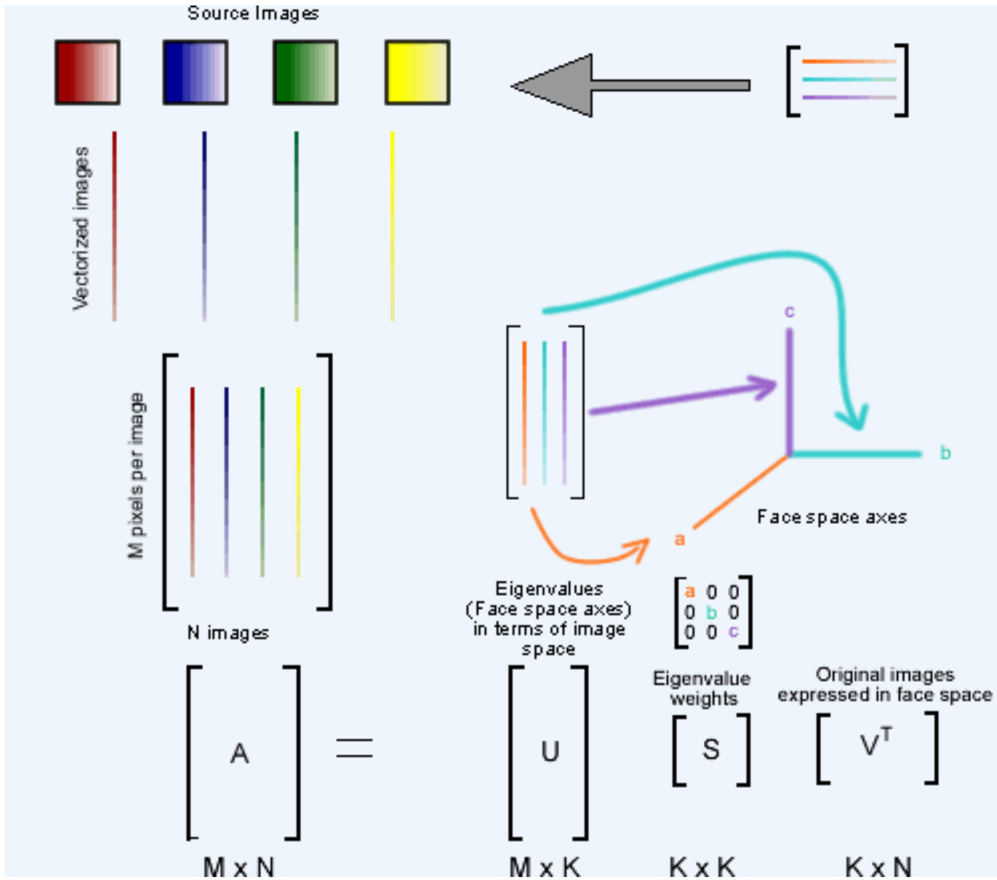
Principal Component Analysis

Given a point set $\{\vec{\mathrm{p}}_j\}_{j=1\ldots P}$, in an M-dim space, PCA finds a basis such that

- coefficients of the point set in that basis are uncorrelated
- first $r < M$ basis vectors provide an approximate basis that minimizes the mean-squared-error (MSE) in the approximation (over all bases with dimension r)



PCA via Singular Value Decomposition



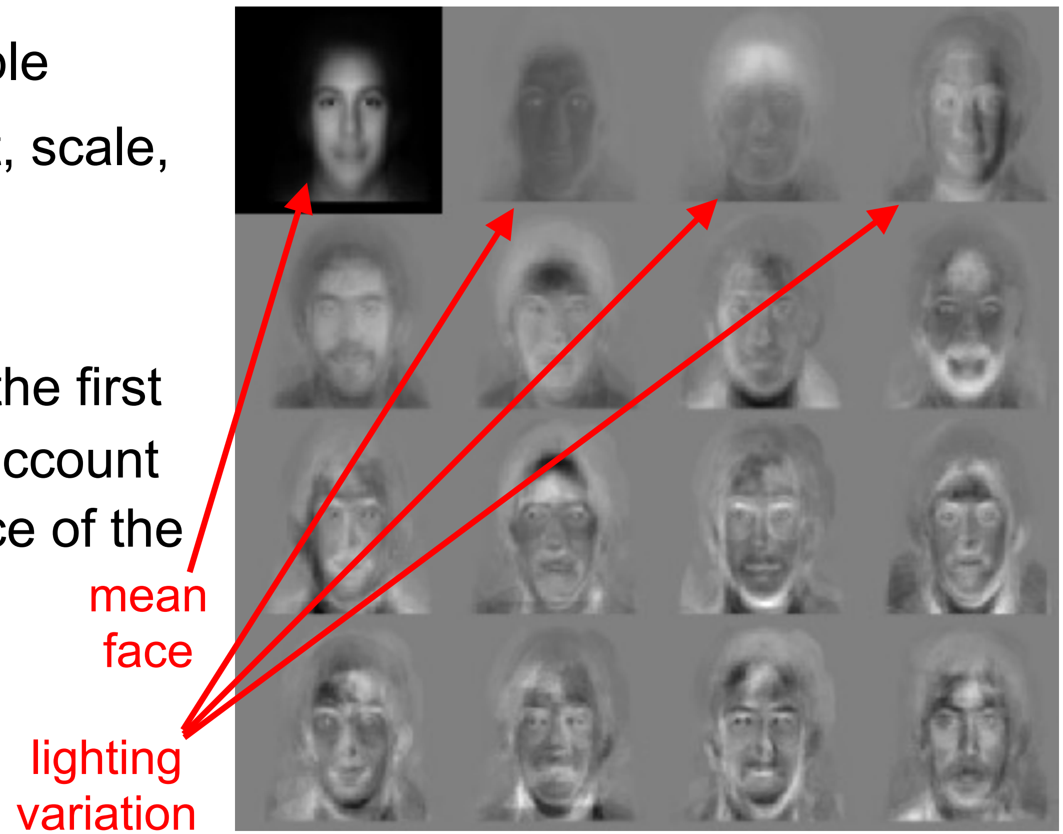
[u,s,v] = svd(A);

http://graphics.cs.cmu.edu/courses/15-463/2004_fall/www/handins/brh/final/

EigenFaces

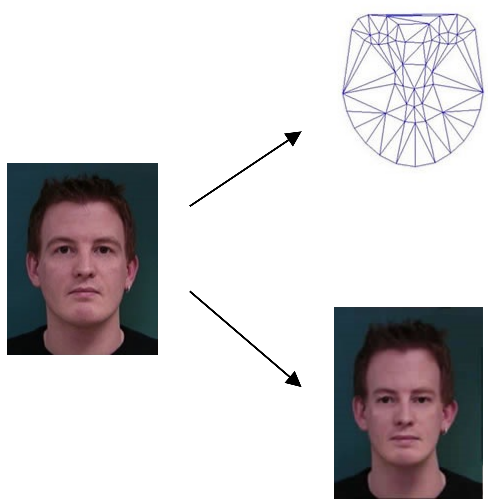
First popular use of PCA on images was for modeling and recognition of faces *[Kirby and Sirovich, 1990, Turk and Pentland, 1991]*

- Collect a face ensemble
- Normalize for contrast, scale, & orientation.
- Remove backgrounds
- Apply PCA & choose the first N eigen-images that account for most of the variance of the data.



The Morphable Face Model

The actual structure of a face is captured in the shape vector **S** = $(x_1, y_1, x_2, \ldots, y_n)^T$, containing the (x, y) coordinates of the n vertices of a face, and the appearance (texture) vector **T** = $(R_1, G_1, B_1, R_2, \ldots, G_n, B_n)^T$, containing the color values of the mean-warped face image.



Shape S

Appearance T

First 3 Shape Basis

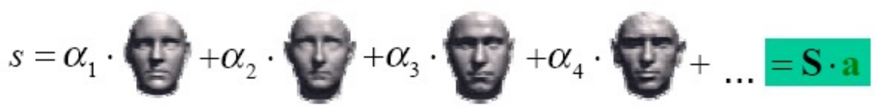
Mean appearance

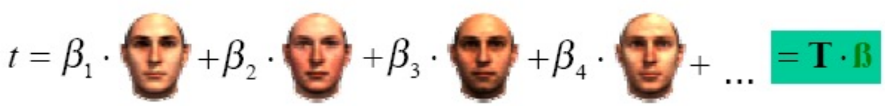
http://graphics.cs.cmu.edu/courses/15-463/2004_fall/www/handins/brh/final/

The 3D Morphable Face Model

Again, assuming that we have ***m*** such vector pairs in full correspondence, we can form new shapes $\mathbf{S}_{model}$ and new appearances $\mathbf{T}_{model}$ as:

$$\mathbf{S}_{model} = \sum_{i=1}^{m} a_i \mathbf{S}_i \qquad \mathbf{T}_{model} = \sum_{i=1}^{m} b_i \mathbf{T}_i$$

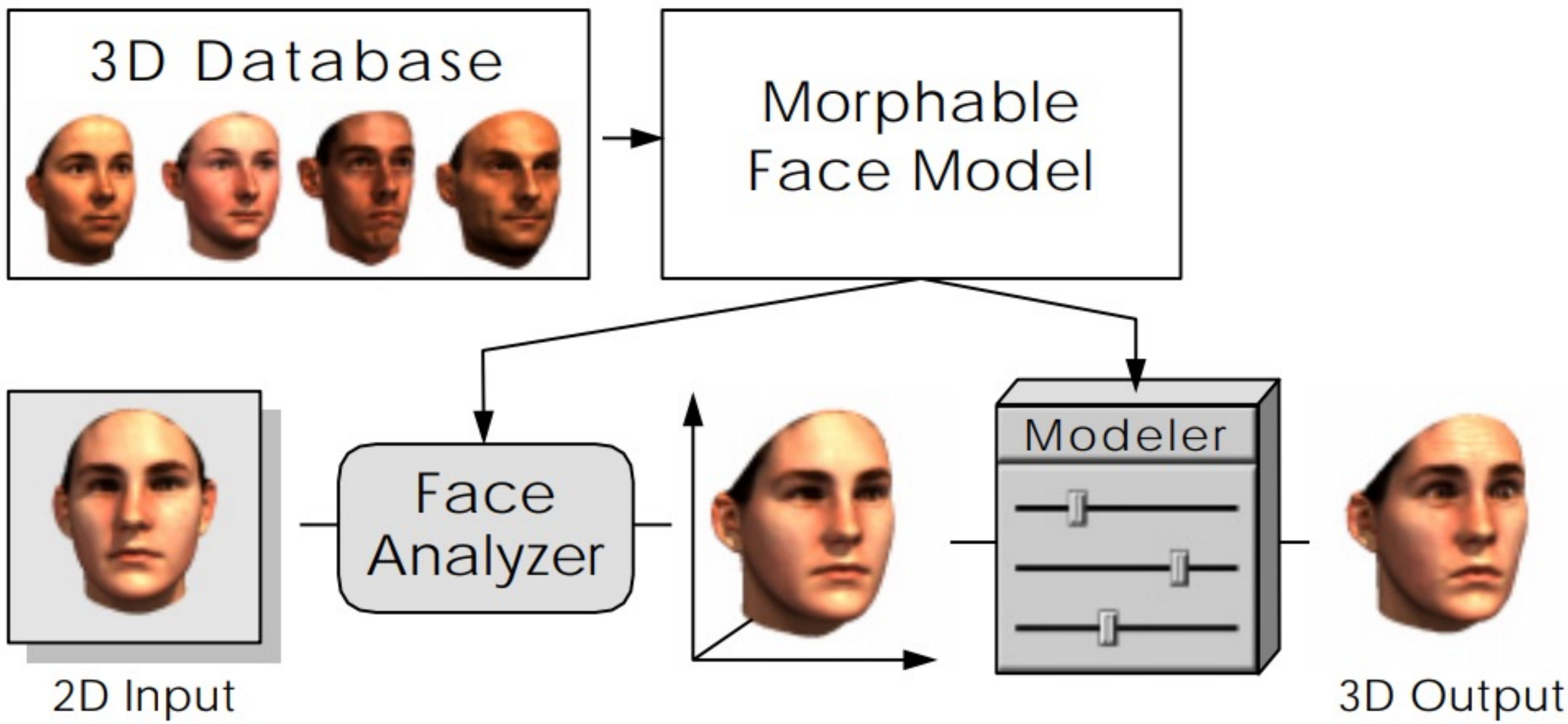




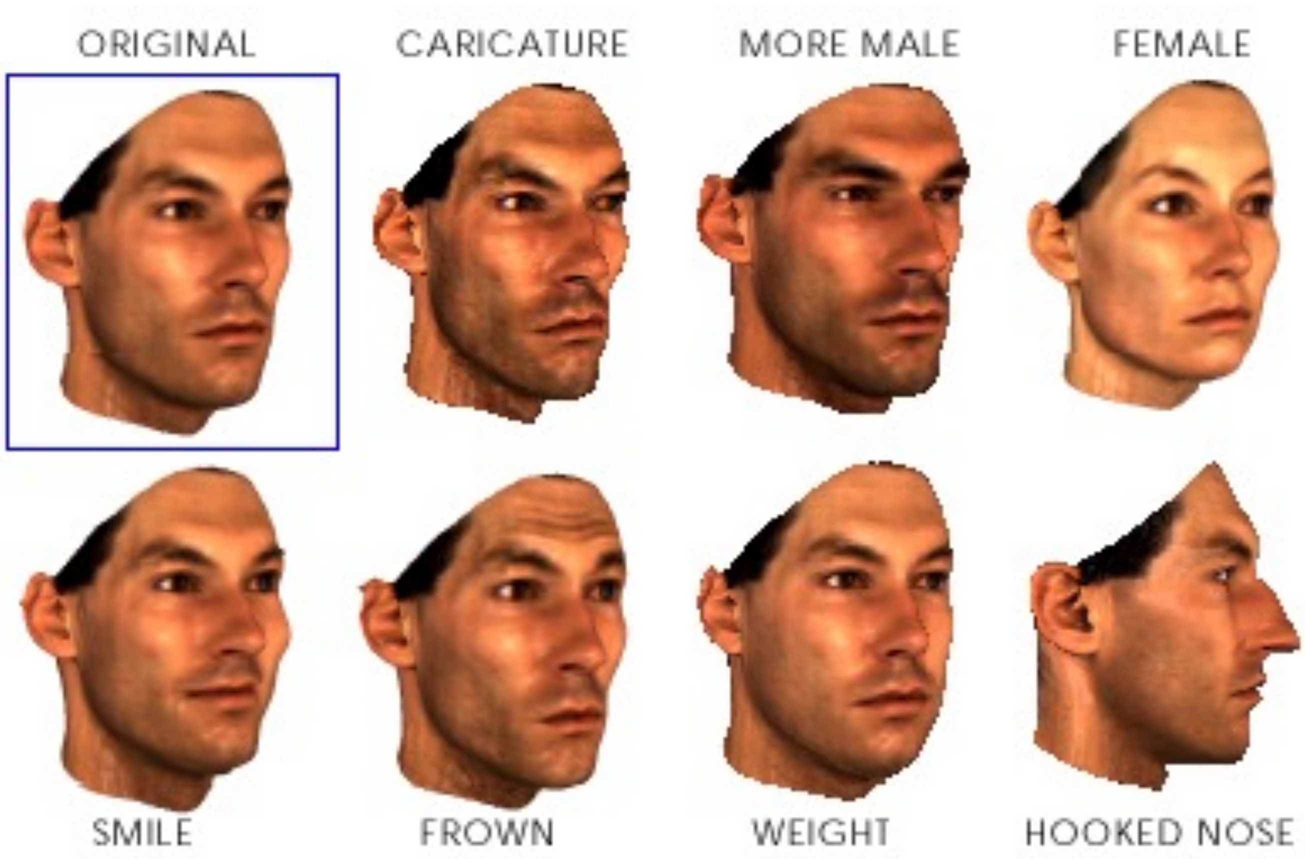
If number of basis faces ***m*** is large enough to span the face subspace then:

<u>Any new</u> face can be represented as a pair of vectors $(\alpha_1, \alpha_2, \ldots, \alpha_m)^T$ and $(\beta_1, \beta_2, \ldots, \beta_m)^T$!

Using 3D Geometry: Blinz & Vetter, 1999

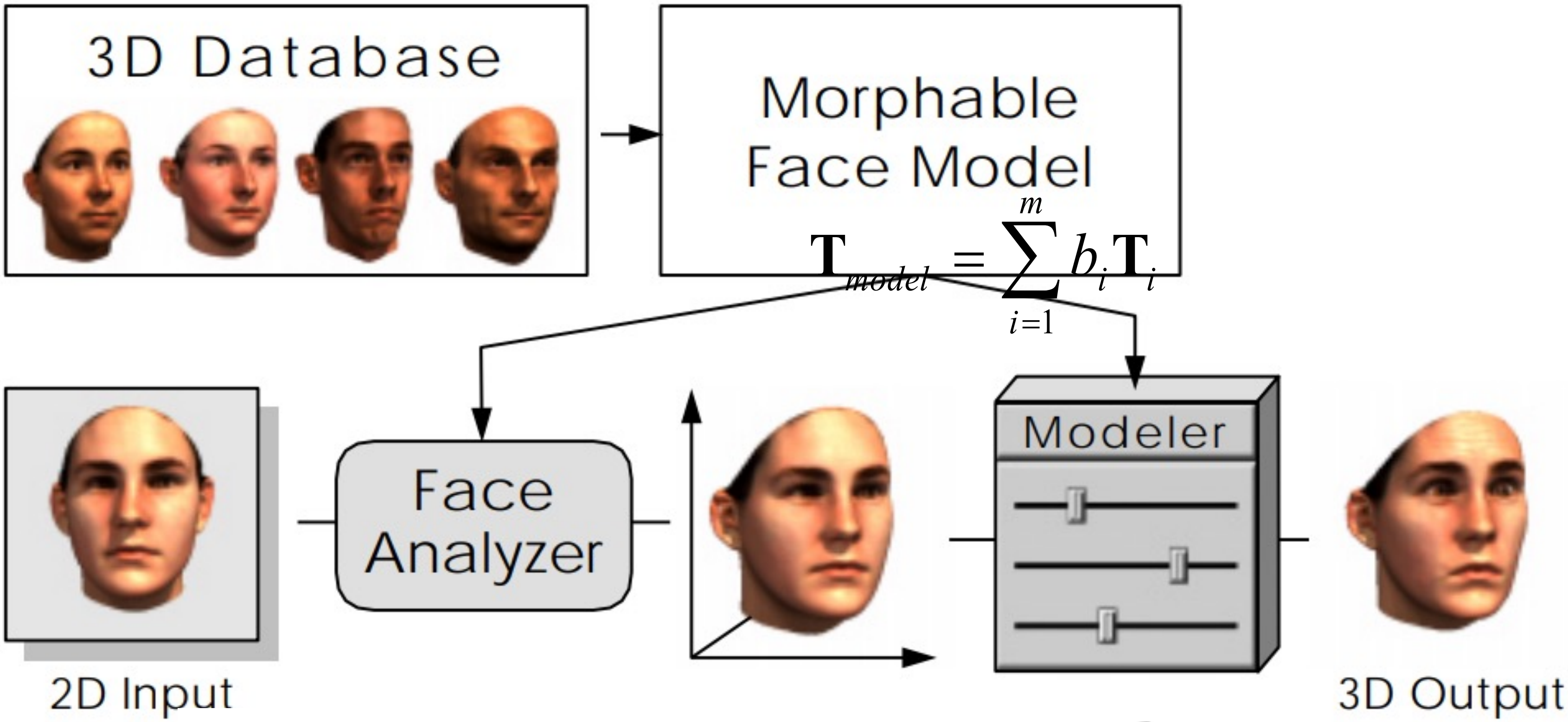


Using 3D Geometry: Blinz & Vetter, 1999



Blanz and Vetter, SIGGRAPH 1999

Using 3D Geometry: Blinz & Vetter, 1999

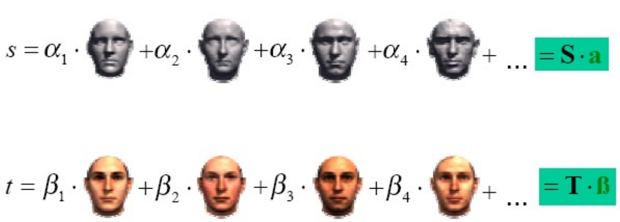


$$\mathbf{T}_{model} = \sum_{i=1}^{m} b_i \mathbf{T}_i$$

$$E_I = \sum_{x,y} \|\mathbf{I}_{input}(x, y) - \mathbf{I}_{model}(x, y)\|^2.$$

Input image

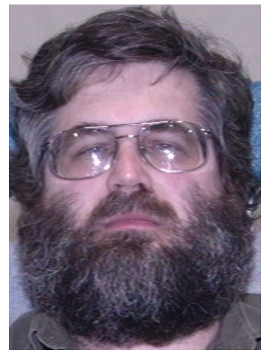
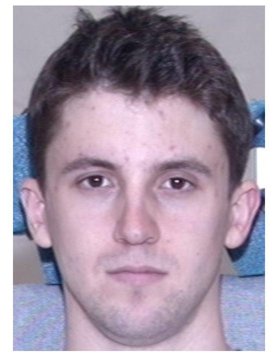
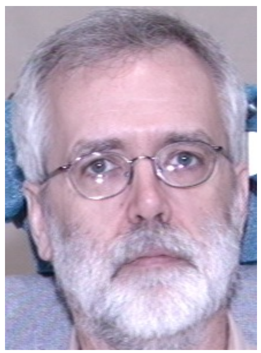
Phong illumination model

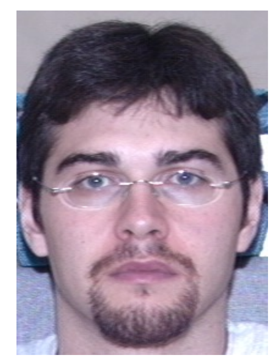
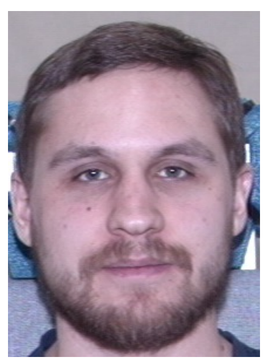


Blanz and Vetter, SIGGRAPH 1999

Using 3D Geometry: Blinz & Vetter, 1999

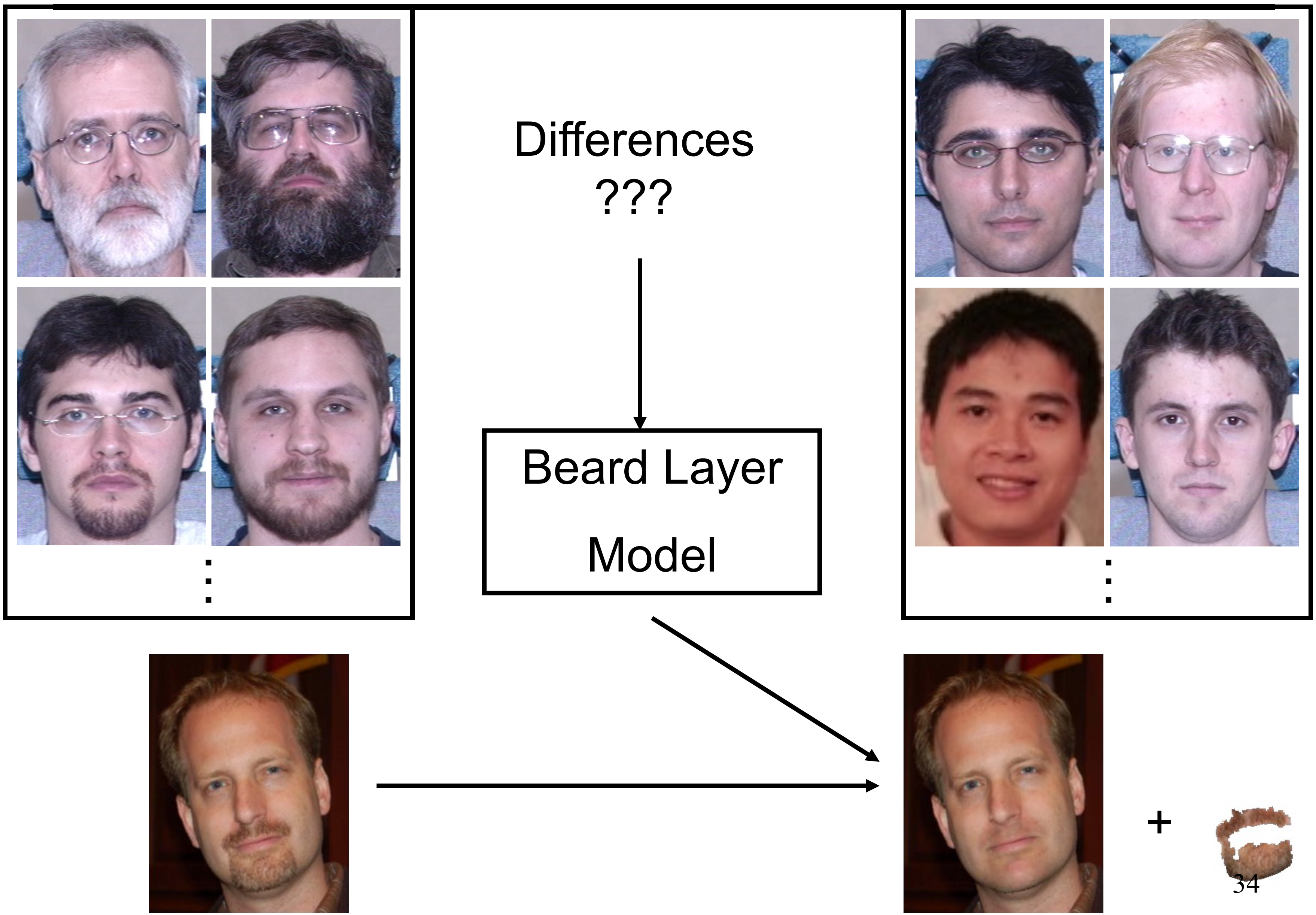
Image-Based Shaving

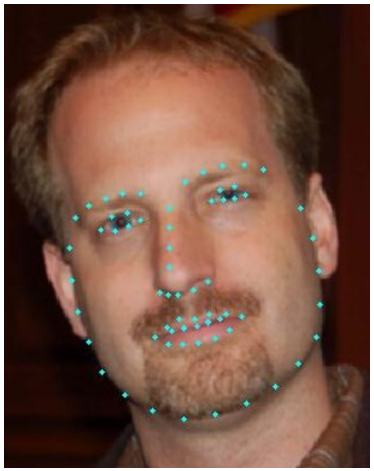




http://graphics.cs.cmu.edu/projects/imageshaving/

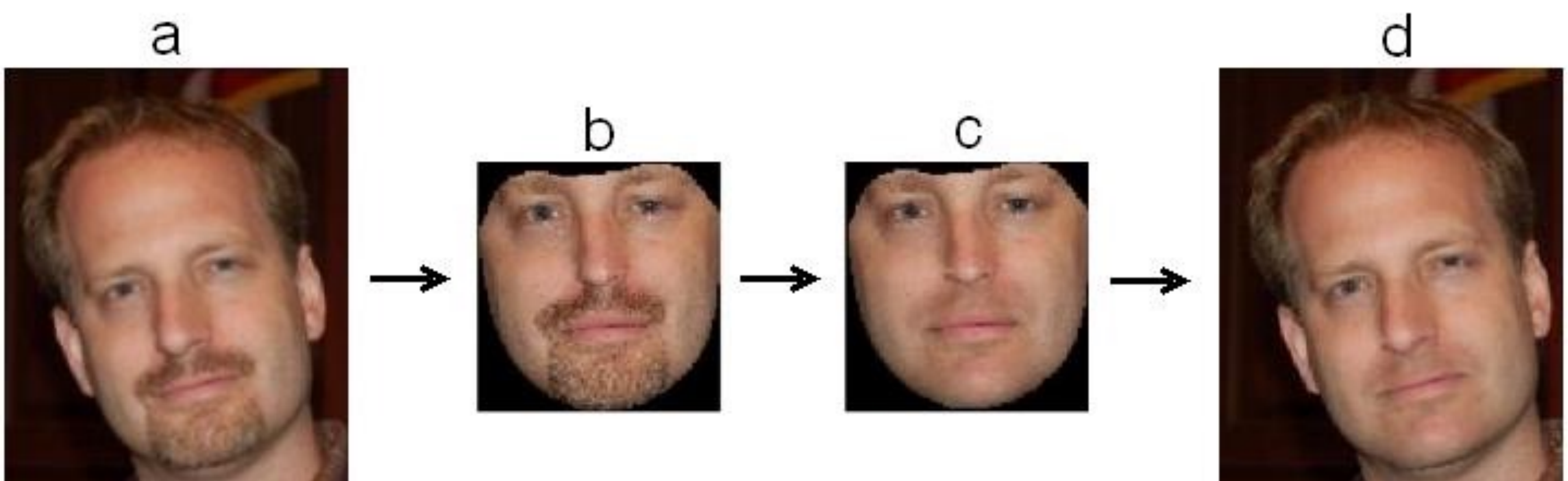
The idea



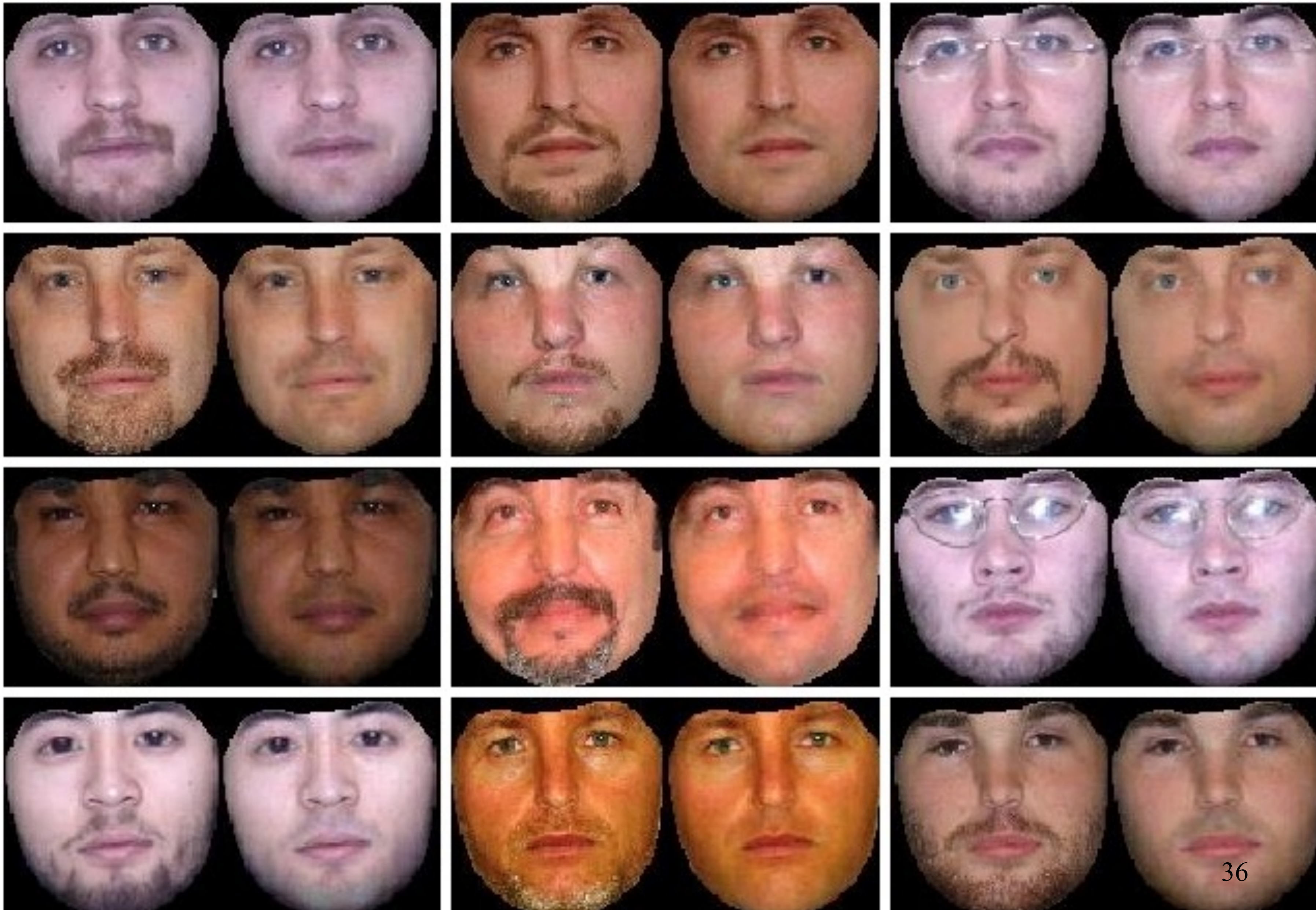
Processing steps



68 landmarks



Some results



Classic Face Pipeline

- Alignment (2D and 3D): 3D is better than 2D.
- Shape + Texture representation.
- Subpopulation mean $\bar{x}$ and deviation Δx
- 3D data and 3D shape representation helps!
  - Easy to change the viewpoint.
- Standard face pipeline:

  Given: Input Image

  Step 1: warp it to canonical pose (2D or 3D)

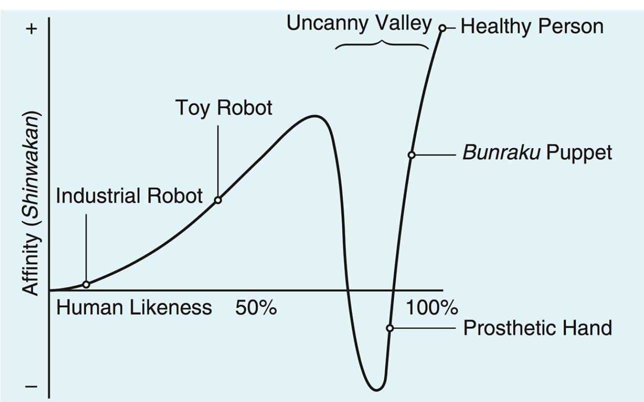
  Step 2: Calculate distances between faces OR apply image manipulation operations.

  Step 3: Unwarp the result back to the original image

  Step 4: Post-processing (e.g., Poisson blending)

Is Face Modeling Easy/Hard?

- Face modeling is easy?
  - Plenty of aligned 3D face data.
  - 2D and 3D computer vision methods.
- Face modeling is hard?
  - Uncanny valley: Human eyes are extremely sensitive to any imperfections on faces.

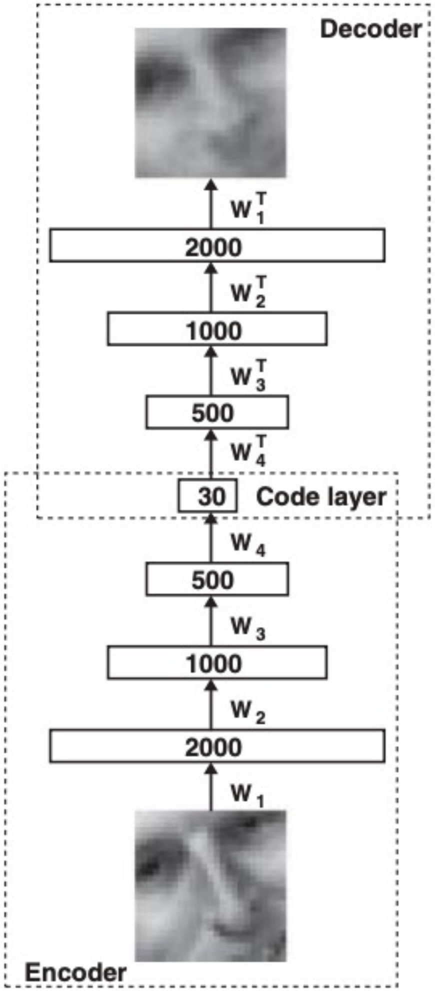


Masahiro Mori

How to Improve the results?

- Using Deep Learning?
- But how?
- Deep learning vision methods:
  - 2D/3D landmark detection
  - 3D pose estimation
  - Face shape reconstruction
- Deep learning graphics models
  - generative models
  - 3D-aware generative models

Autoencoder vs. PCA



Training objective: E encoder, G decoder/generator

$$\arg\min_{E,G} \mathbb{E}_x ||G(E(x)) - x||_2$$

Top: Input. Middle: Autoencoder. Bottom: PCA

GE Hinton, RR Salakhutdinov. Science 2006

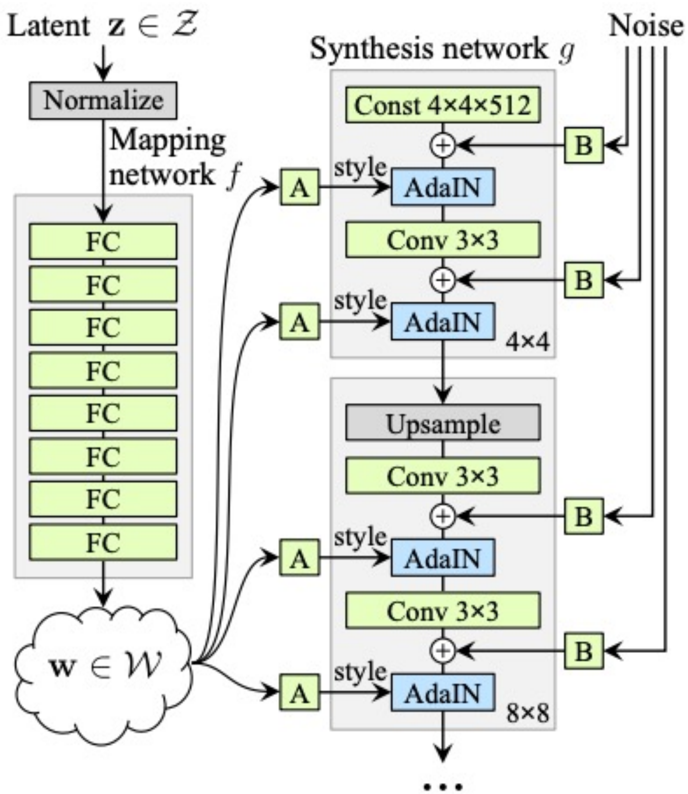
Deep learning method

PCA→ Generative Model

StyleGAN Face Results

StyleGAN [Karras et al., 2019]

Face Editing with GANs Projection



Optimizing the latent code

$$z^* = \arg\min_z \mathcal{L}(G(z), x)$$

Optimizing the style code

$$w^* = \arg\min_w \mathcal{L}(g(w), x)$$

Optimizing the extended style code

$$w_+^* = \arg\min_{w+} \mathcal{L}(g(w_+), x)$$

Image2StyleGAN [Abdal et al., 2019], StyleGAN2 [Karras et al., 2019]

43

Face Editing = latent space editing

Interpolation between two faces in the w+ space

Image2StyleGAN [Abdal et al., 2019], StyleGAN2 [Karras et al., 2019]

Face Editing = latent space editing

Face Editing with GANs Projection

Input Edit Output

Image2StyleGAN++ [Abdal et al., 2020]

Face Editing with GANs Projection

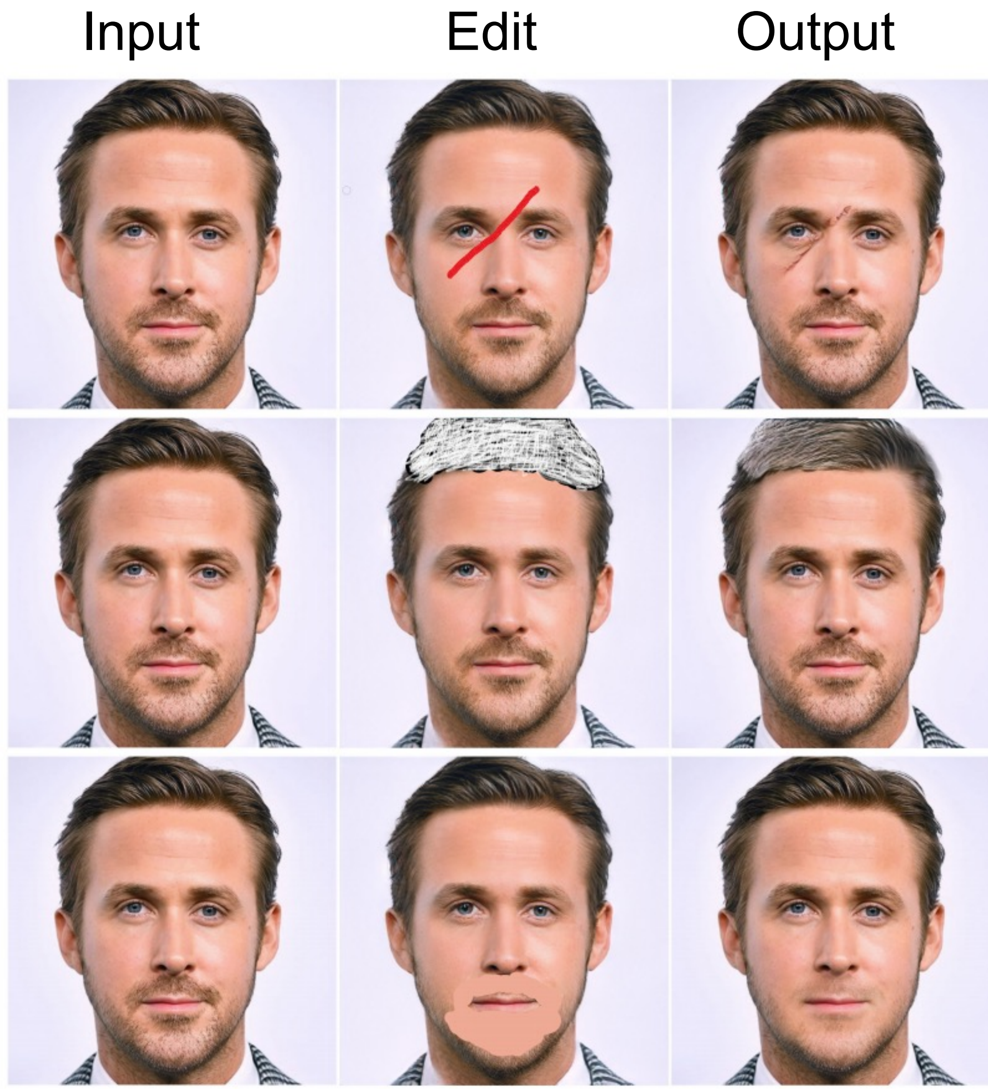
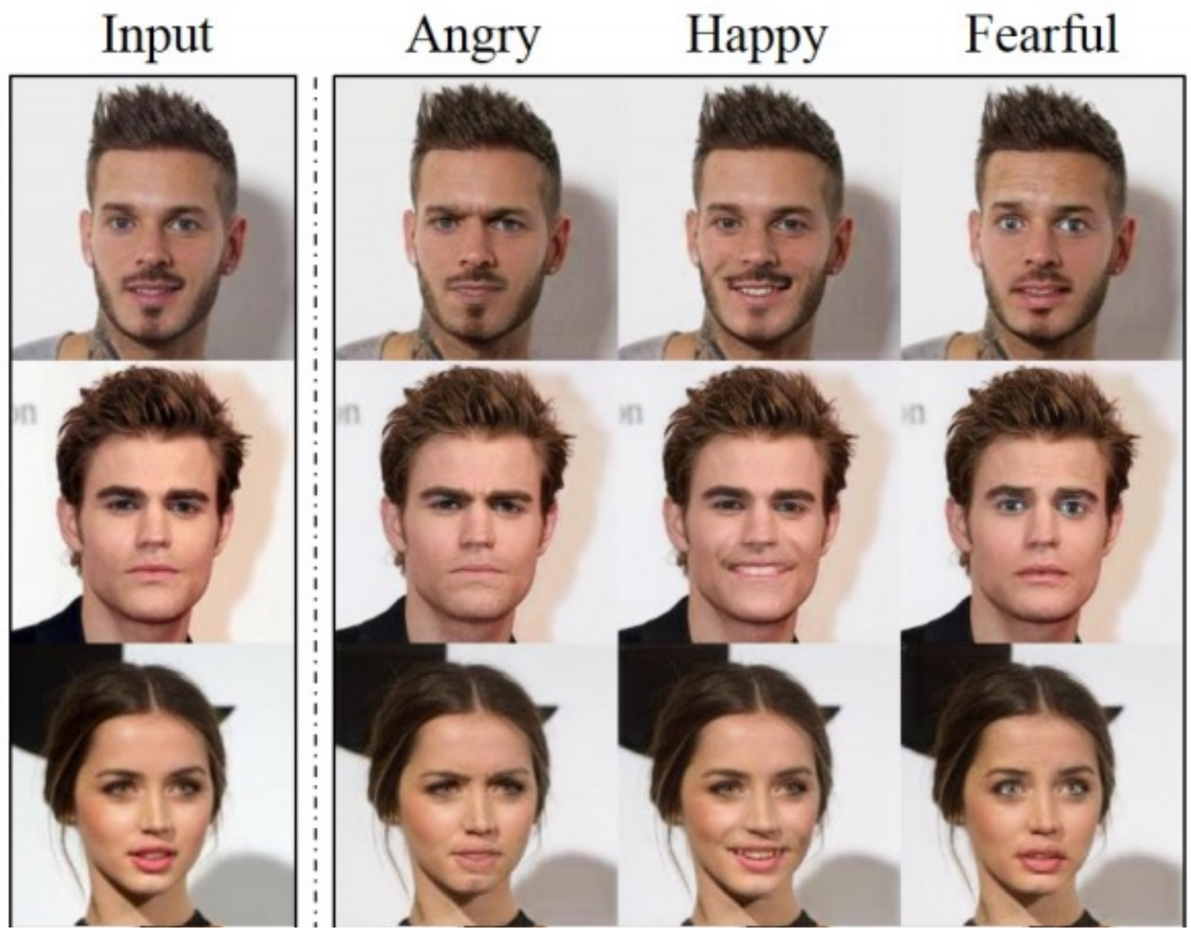


Image2StyleGAN++ [Abdal et al., 2020]

Deep learning method

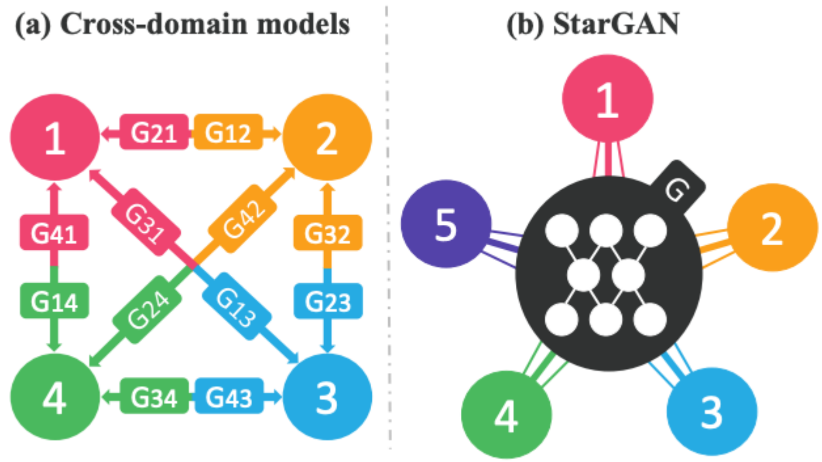
Image-to-Image Translation

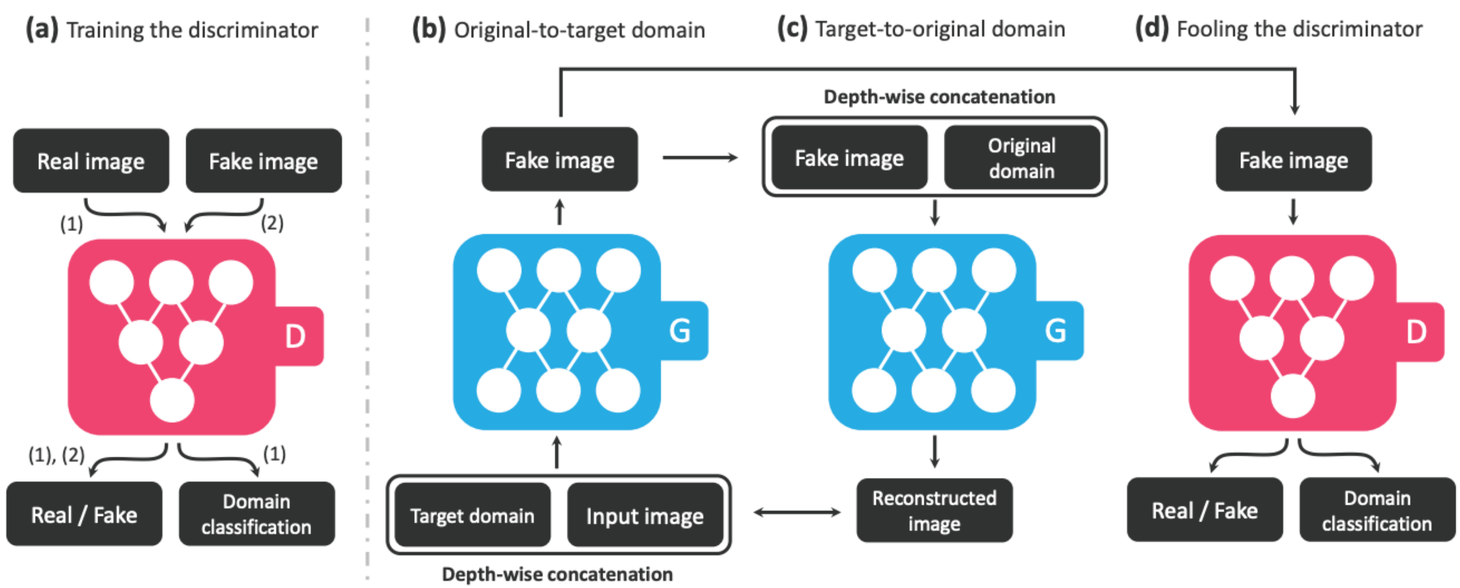
Face Translation with StarGAN



StarGAN [Choi et al., 2018]

Face Translation with StarGAN





StarGAN [Choi et al., 2018]

Face Translation with StarGAN v2

Multi-modal synthesis; supports a reference image

StarGAN v2 [Choi et al., 2020]

3D + Deep Learning

3D representation+ image-to-image

CGI Face Editing

Professional video

CGI Face Editing

Personal video

Video: ©
https://www.youtube.com/watch?v=7Flvkn2quLY

Applications

Original video Pose editing Expression editing

- Editing of head pose, rotation, face expression and eye gaze
- Combination of model-based face capture and CNN

3D + CNN

Model-based face capture and reenactment

Garrido et al., ToG 2016

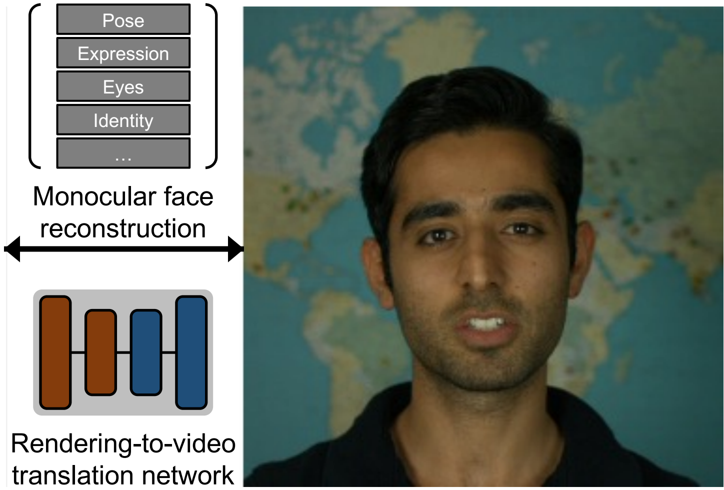
Kemelmacher-Shlizerman et al., ECCV 2010
Shi et al., ToG 2014
Suwajanakorn et al., ICCV 2015
Thies et al., CVPR 2016
Averbuch-Elor et al., ToG 2017
Thies et al., SIGGRAPH 2018

CNN-based methods

Karras et al., ICLR 2018

Goodfellow et al., NIPS 2014
Isola et al., CVPR 2017
Chen and Koltun, ICCV 2017
Tewari et al., ICCV 2017
Olszewski et al., ICCV 2018
Wang et al., CVPR 2018

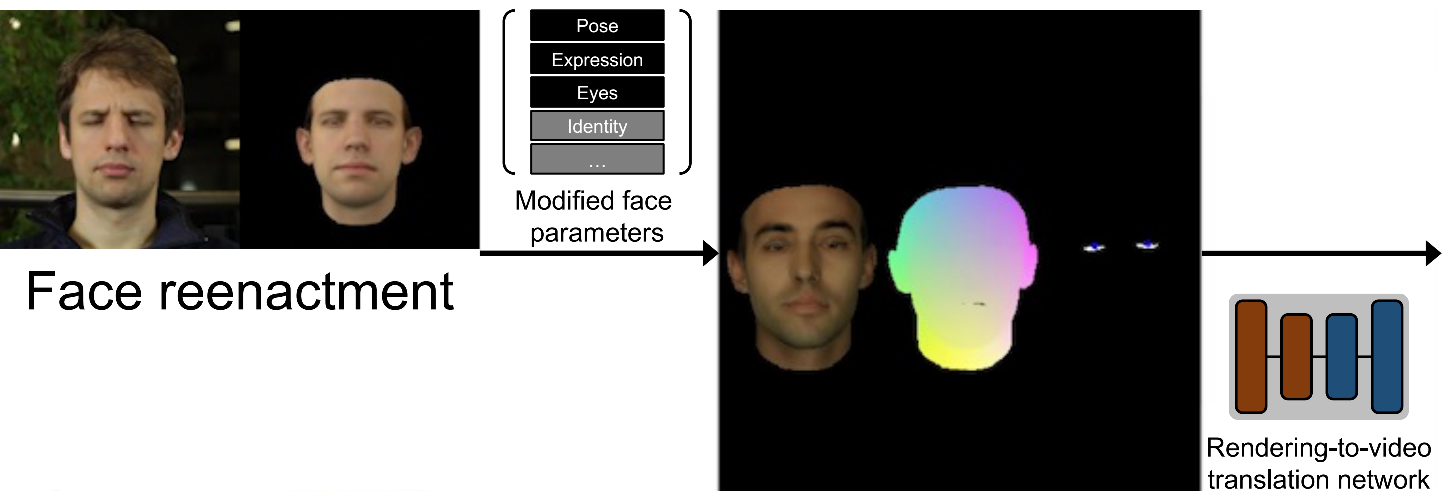
Overview



Training video

Deep video Portrait [Hyeongwoo et al., SIGGRAPH 2018]

Overview



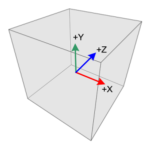
User interaction

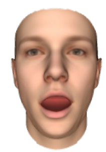
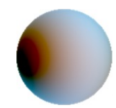
Modified rendering

Deep video Portrait [Hyeongwoo et al., SIGGRAPH 2018]

Monocular 3D Face Reconstruction

- Parametric 3D face model

$p = ($ 

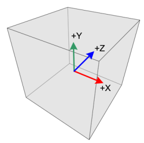
,  ,  ,  $) \in \mathbb{R}^{257}$

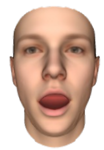
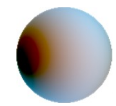
Pose Expression Identity Lighting

$$\min_{p} E\,(p) = E_{\text{photo}}(p) + E_{\text{land}}(p) + E_{\text{reg}}(p)$$

Monocular 3D Face Reconstruction

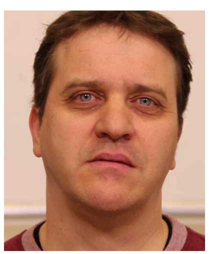
- Parametric 3D face model

$p = ($ 

,  ,  ,  $) \in \mathbb{R}^{257}$

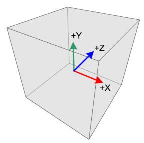
Pose Expression Identity Lighting

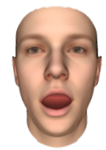
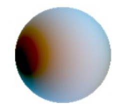
$$\min_{p} E\,(p) = E_{\text{photo}}(p) + E_{\text{land}}(p) + E_{\text{reg}}(p)$$

$\|$  $-$  $\|_2^2$

Monocular 3D Face Reconstruction

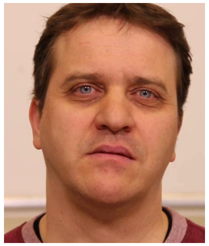
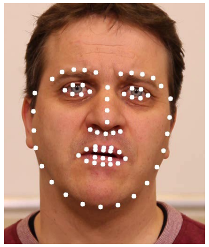
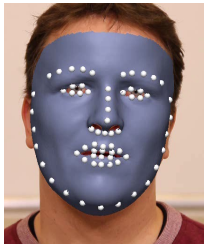
- Parametric 3D face model

$p = ($ 

,  ,  ,  $) \in \mathbb{R}^{257}$

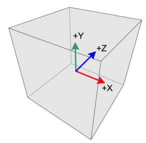
Pose Expression Identity Lighting

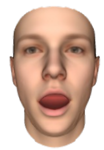
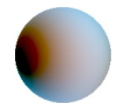
$$\min_p E(p) = E_{\text{photo}}(p) + E_{\text{land}}(p) + E_{\text{reg}}(p)$$

$\|$  $-$  $\|_2^2$ $\|$  $-$  $\|_2^2$

Monocular 3D Face Reconstruction

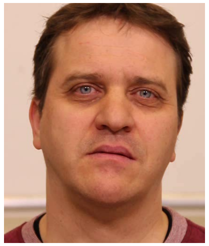
- Parametric 3D face model

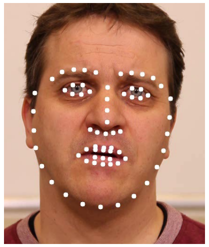
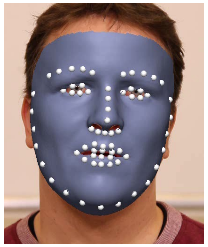
$p = ($ 

$,$  $,$  $,$  $) \in \mathbb{R}^{257}$

Pose Expression Identity Lighting

$$\min_{p} E\,(p) = E_{\text{photo}}(p) + E_{\text{land}}(p) + E_{\text{reg}}(p)$$

$\|$  $-$  $\|_2^2$

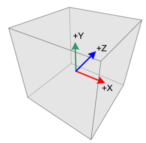
$\|$  $-$  $\|_2^2$

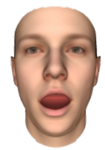
Statistical and temporal regularization

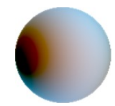
Garrido et al., ToG 2016

Monocular 3D Face Reconstruction

- Parametric 3D face model

$p = ($


,

,

,

$) \in \mathbb{R}^{257}$

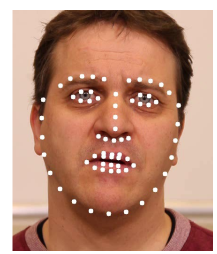
Pose Expression Identity Lighting

$$\min_{p} E\,(p) = E_{\text{photo}}(p) + E_{\text{land}}(p) + E_{\text{reg}}(p)$$

- Eye model

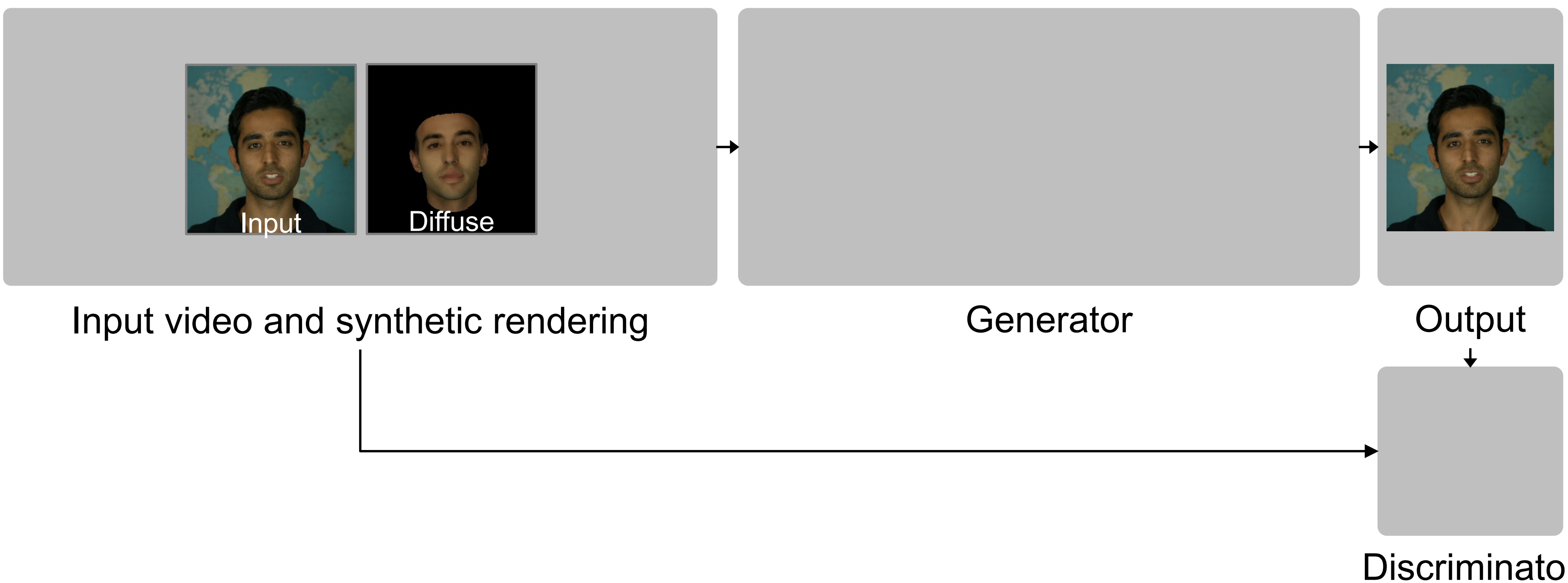
$e = ($

$) \in \mathbb{R}^{4}$



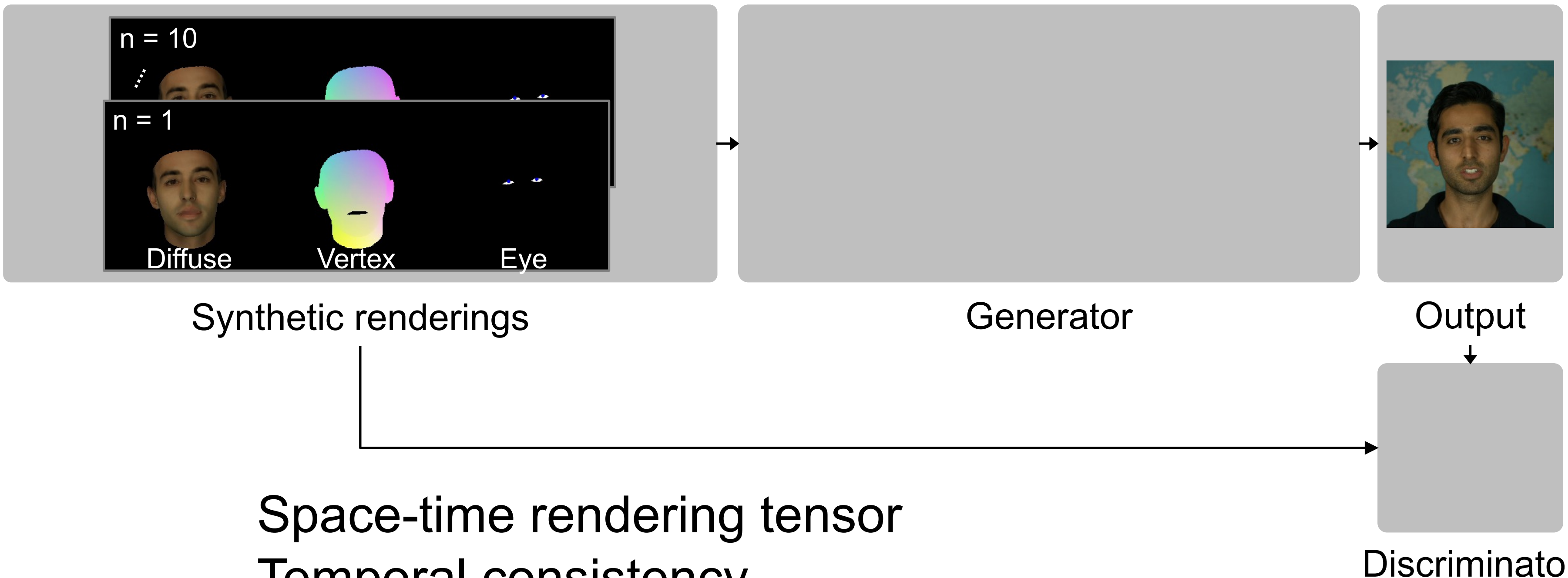
Saragih et al.,
FG 2011

Rendering-to-Video Translation Network



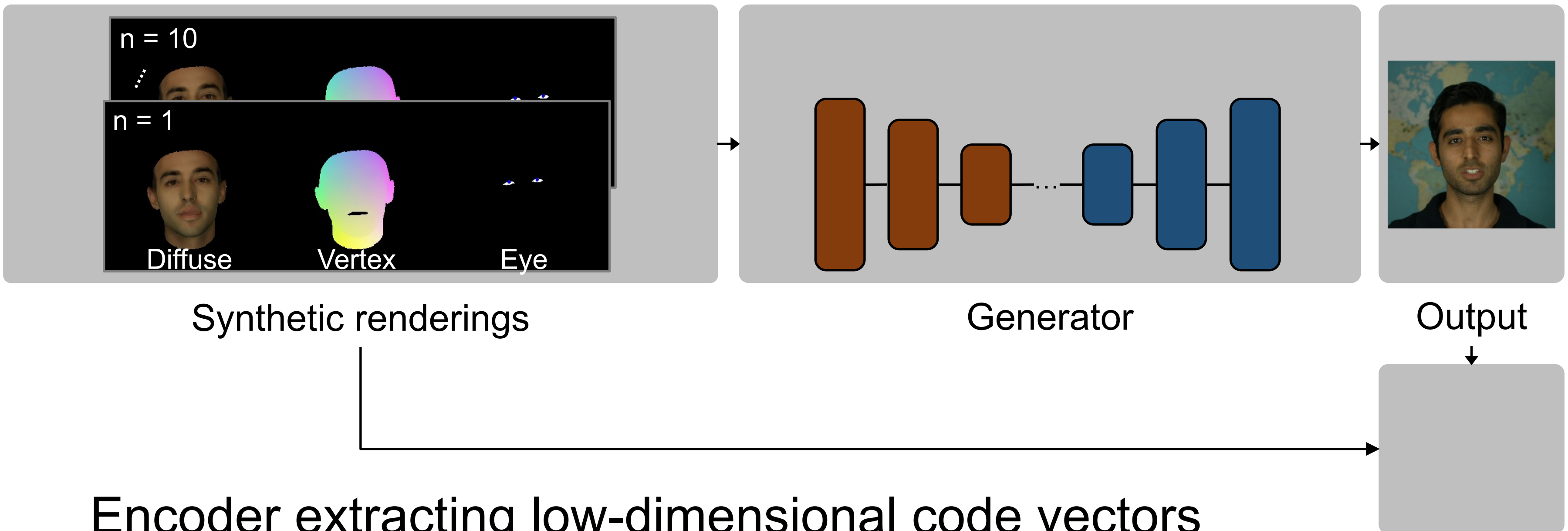
Pix2pix [Isola et al., ICCV 2017]

Rendering-to-Video Translation Network



Space-time rendering tensor
Temporal consistency

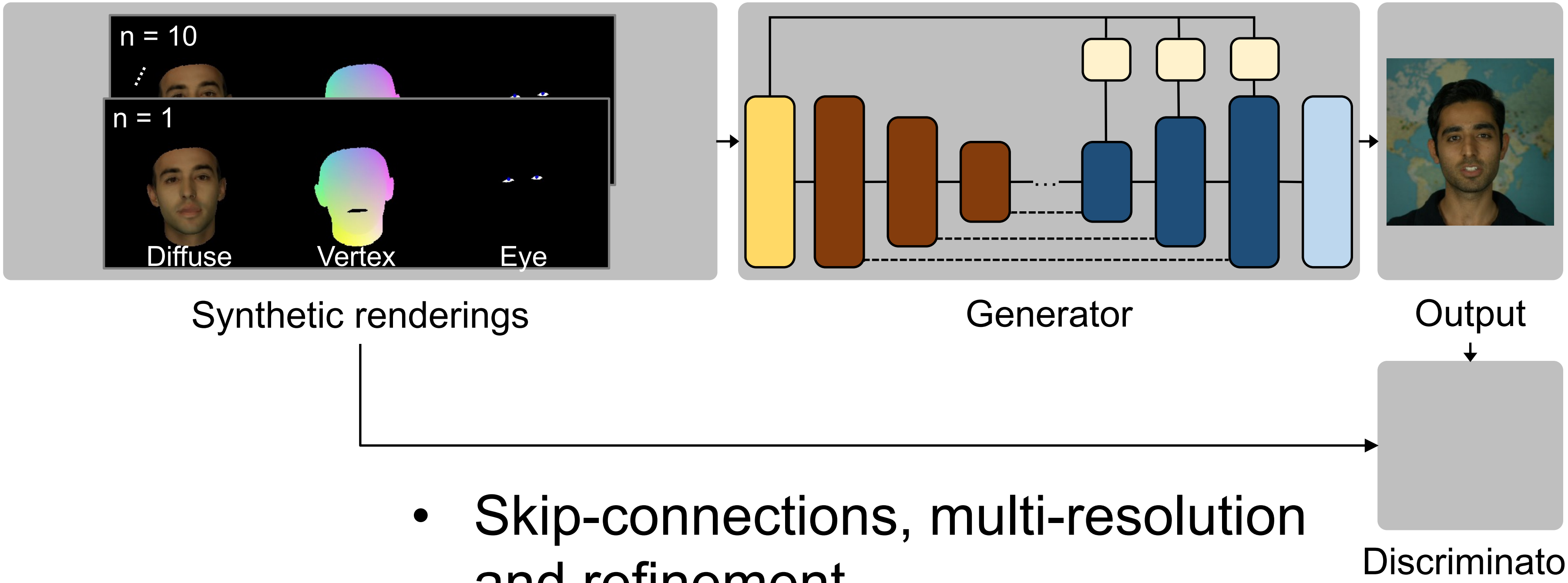
Rendering-to-Video Translation Network



Encoder extracting low-dimensional code vectors

Decoder/generator reconstructing images

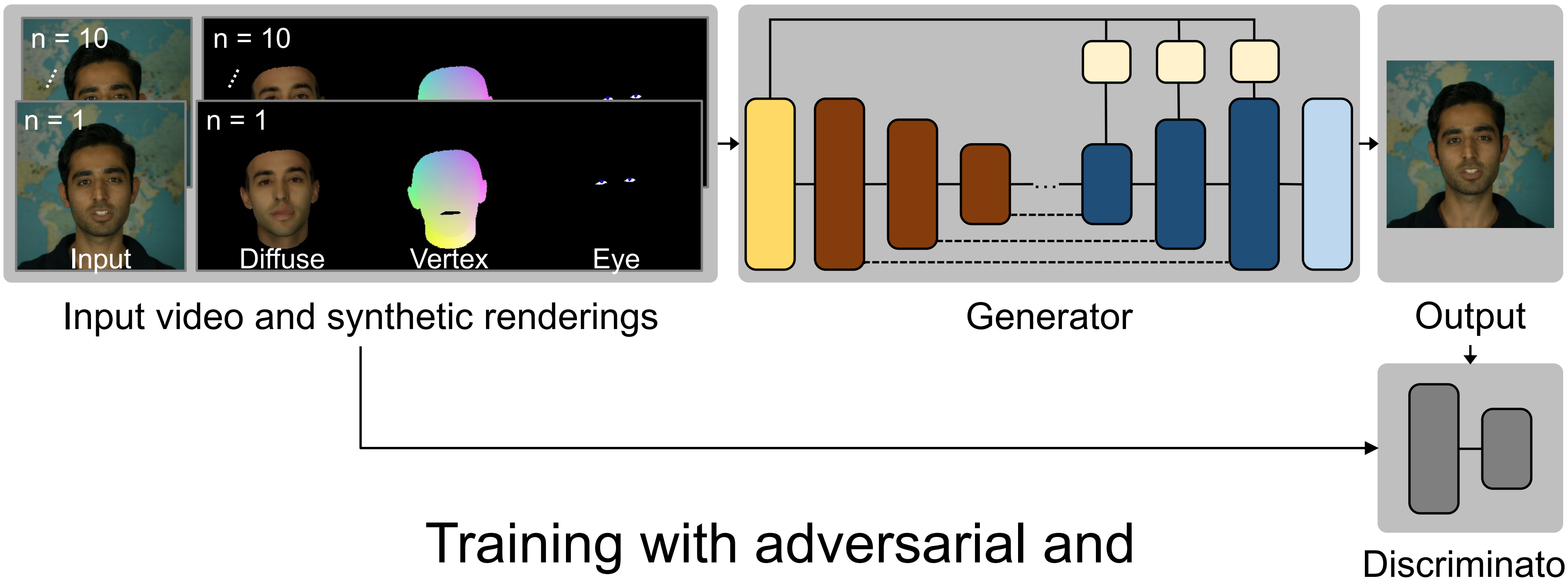
Rendering-to-Video Translation Network



- Skip-connections, multi-resolution and refinement
- Fine-scale details

U-Net [Ronneberger et al., MICCAI 2015]
CRN [Chen and Koltun, ICCV 2017]

Rendering-to-Video Translation Network



Training with adversarial and L_1 losses

GANs [Goodfellow et al. NIPS 2014]
Pix2pix [Isola et al. ICCV 2017]

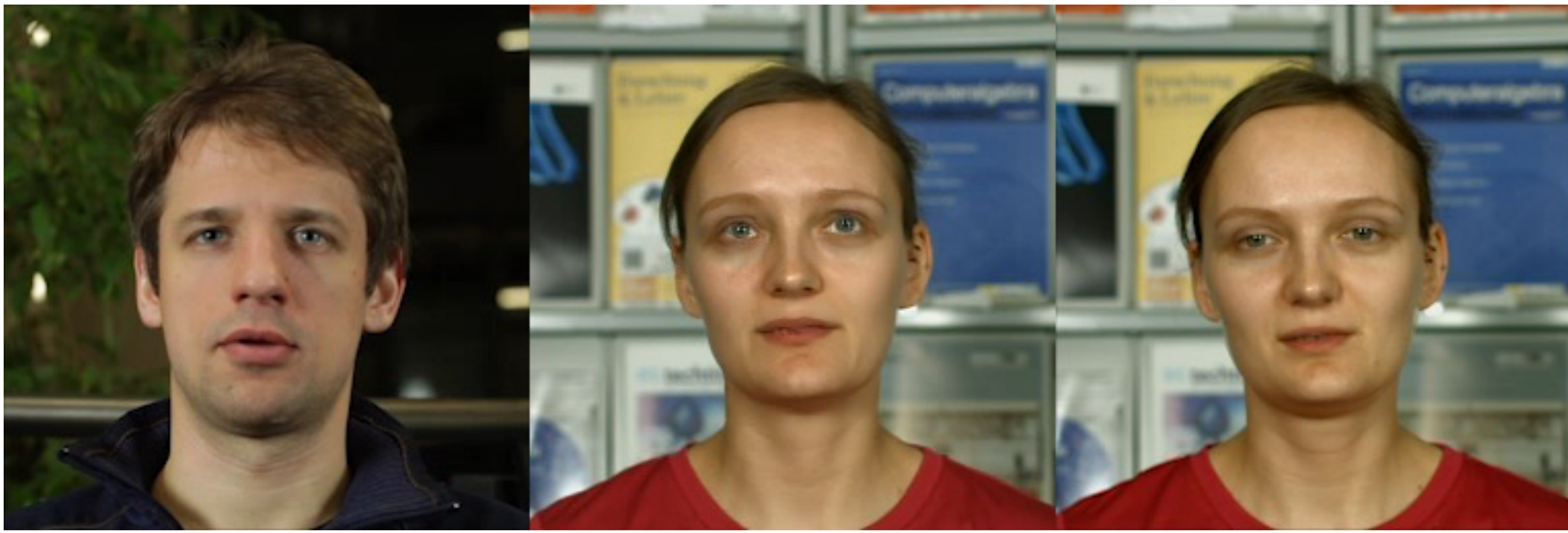
Result: Facial Reenactment

Retargeting portraits videos from source to target

Source Result

Result: Facial Reenactment

Full reenactment of head pose, head rotation, face expression and eye gaze



Source | Result | Face2Face (Thies et al., 2016)

Result: Facial Reenactment

Source Target Result

Video: courtesy of the White House
(public domain)

Result: Visual Dubbing

Visual discomfort due to the discrepancy between video and audio tracks

Dubbing actor video Original video

Result: Visual Dubbing

Modification of mouth motion to match audio tracks

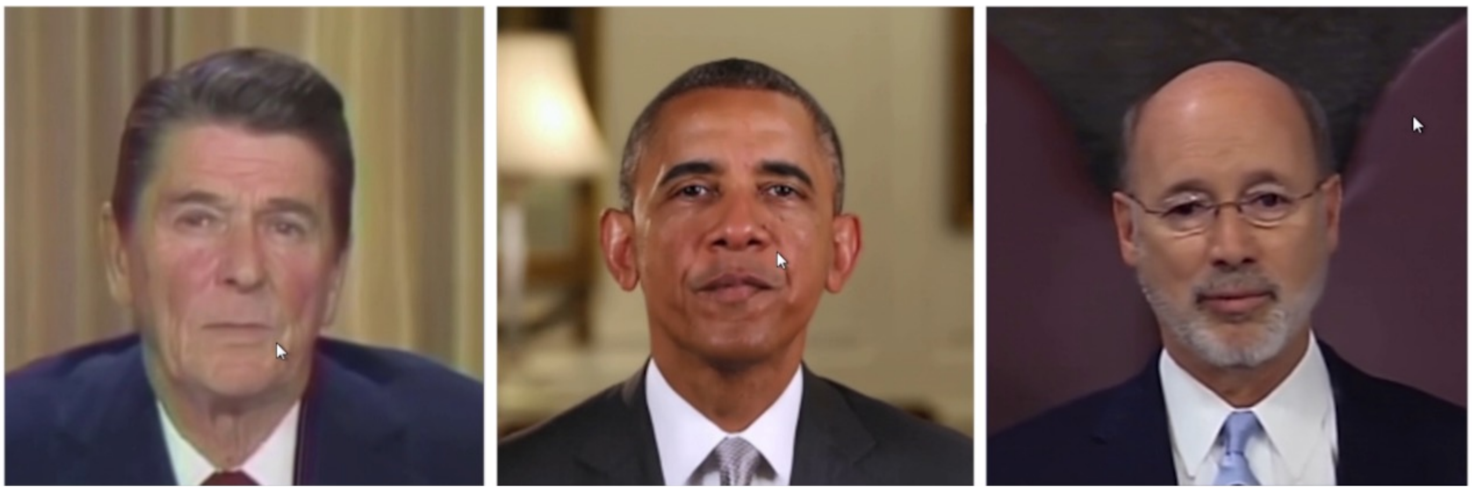
Dubbing actor video　Dubbed video　Garrido et al., 2015

Result: Interactive Editing

Pose　　Expression　　Shape

Approximately 9 fps

Result: Interactive Editing



YouTube videos 2× speed

Approximately 9 fps

Reagan video courtesy of NARA (public domain)
Obama video courtesy of the White House (public domain)
Wolf video courtesy of Tom Wolf (CC BY)

Result: Post-Production

Face reshaping Subtle expression editing

The Curious Case of Benjamin Button
video courtesy of Lola Visual Effects

Result: Pose Correction in Teleconferencing

Modification of head pose to match camera views

Setup

Camera view

Rotating up

Result: Multi-View Teleconferencing

Rotating up + side to side

Model-based video coding: 31 KB/s
h.264 (e.g., Skype): 192 KB/s