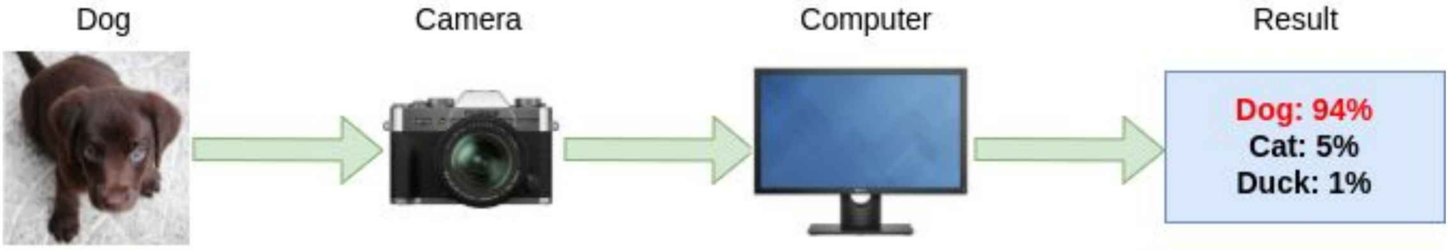
Face Morphing Attacks: MorphGANFormer

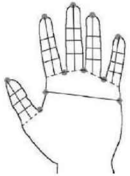
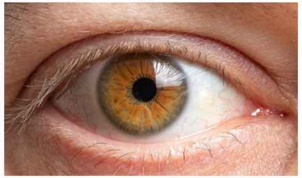
Na Zhang

Computer Vision (CV)

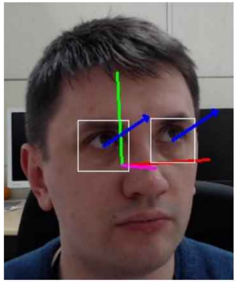
- It enables computers and systems to “see” , observe and understand the content of the inputs, like images, videos, etc.
  - “See”
    - acquire information from the real world
  - Observe
    - derive meaningful information
  - Understand
    - take actions or make decisions based on that information



Biometrics

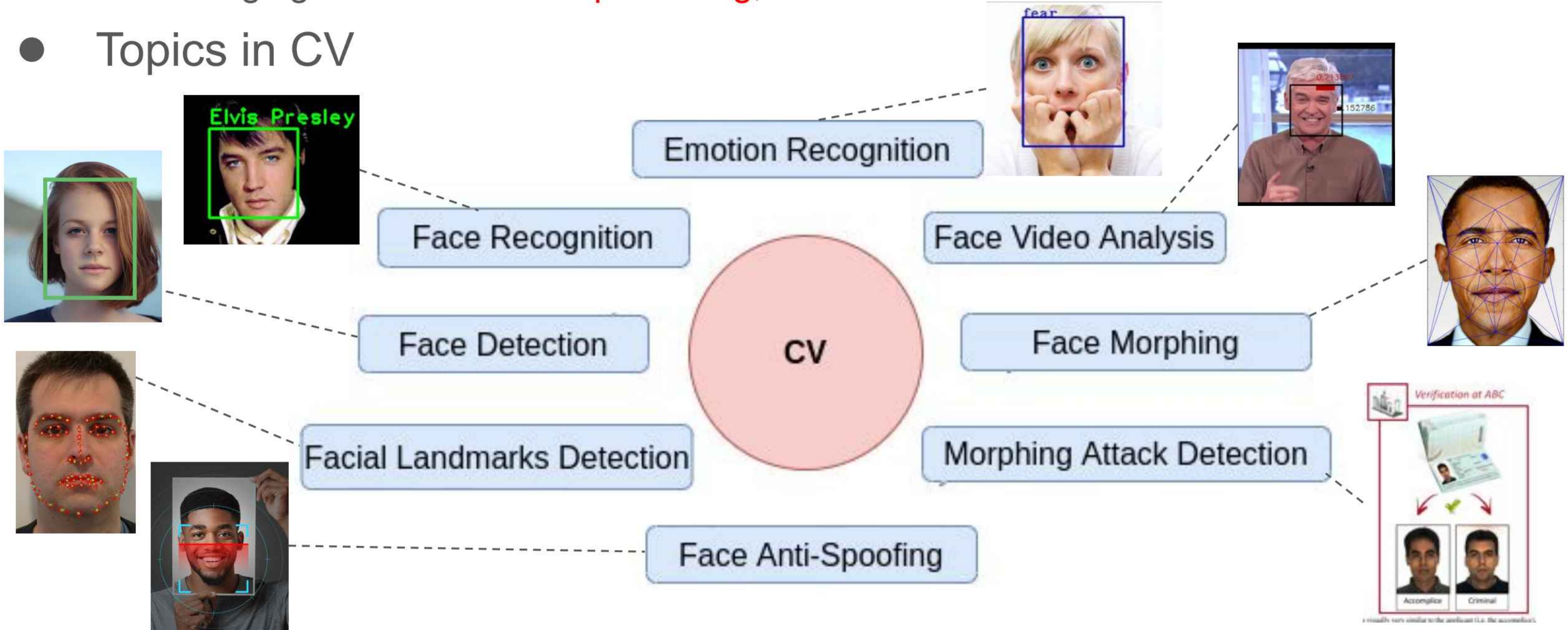
     

- Distinctive and measurable human characteristics
- Used to label / describe individuals
- It combines CV and knowledge of human physiology and behavior
    - Physiological characteristics
        - related to the shape of the body
        - e.g. fingerprint, palm, face, DNA, hand geometry, iris, retina, odor/scent
    - Behavioral characteristics:
        - related to the pattern of behavior of a person
        - e.g. hand gesture, typing pattern, gaze pattern, voice, gait

Face Biometric

- One of the most expressive and informative biometric traits
- Many studies from the perspectives of various different disciplines
  - ranging from CV and deep learning, to neuroscience and biometrics
- Topics in CV

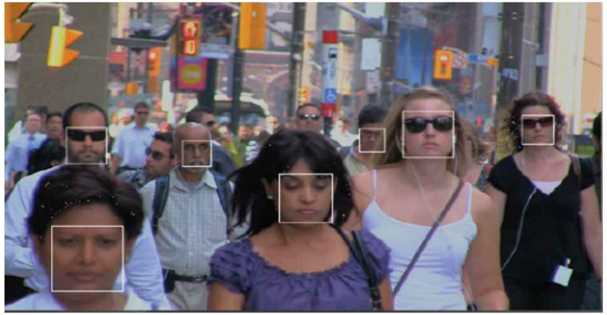


Face Analysis

- With the development of computer hardware and imaging technology, face related applications have been applied widely to daily lives

access control

video surveillance



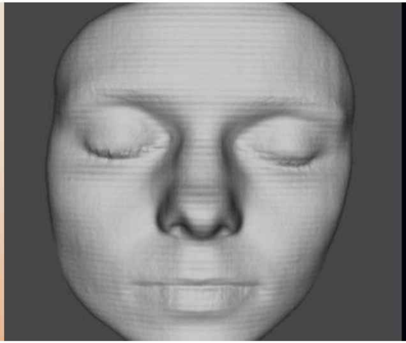
- The demands of face analysis are also growing quickly in recent years
- Automatic face analysis will be one promising tool in many areas in the future

Data Types

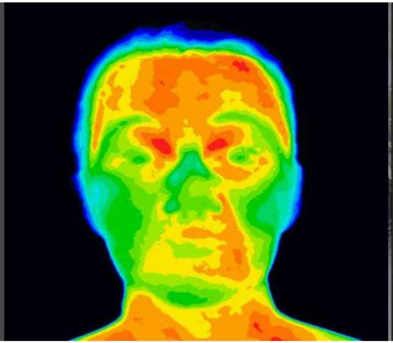
- The types of raw data can be:

RGB Images

Depth Maps



Thermal Images

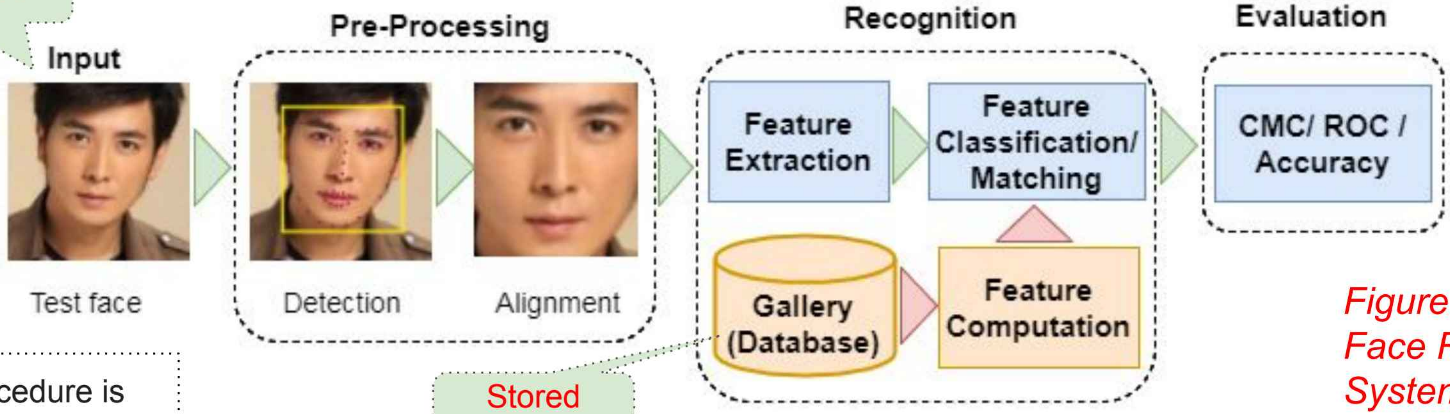


Video

Morphing Attack – Morphed Faces Generation

- Face recognition systems (FRS) have emerged as a popular technique for person identification and verification
- e.g., Automatic Border Control System
  - verify a person's identity with his electronic machine-readable travel document (eMRTD)
  - by comparing the face image of the traveler with a reference in the database

Traveler's Face



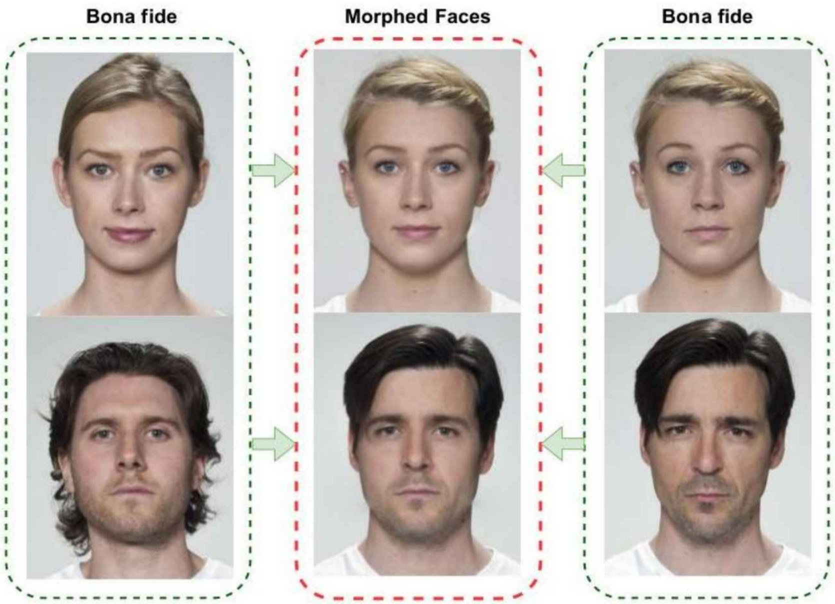
Stored Reference

verification procedure is conducted in test stage

Figure: Pipeline of Face Recognition System (FRS)

Vulnerability of FRS

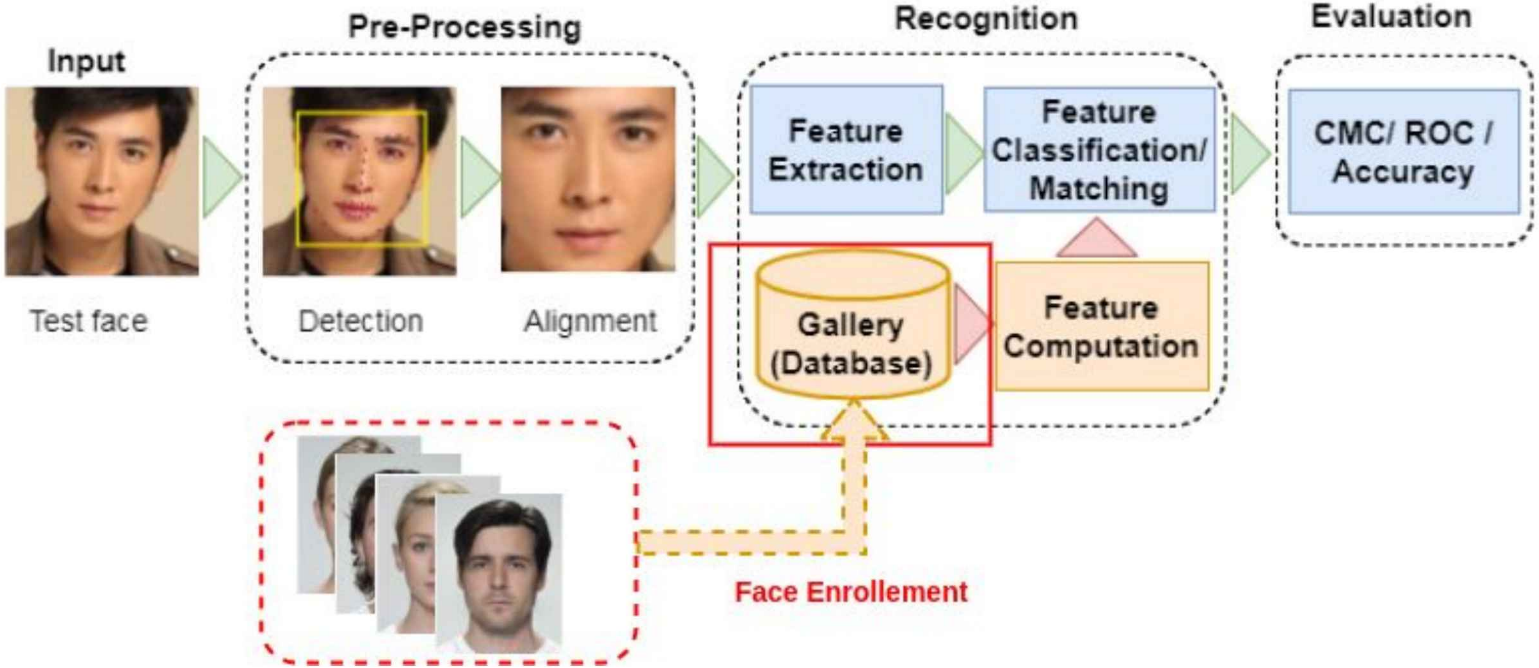
- FRS
  - a popular technique for person identification and verification
- Vulnerable to adversarial attacks
  - although with high accuracy
- Attacks based on morphed faces pose a severe security risk
  - realistic enough to fool human
- Attack vs. Defense



strong visual resemblance to both bona fide faces

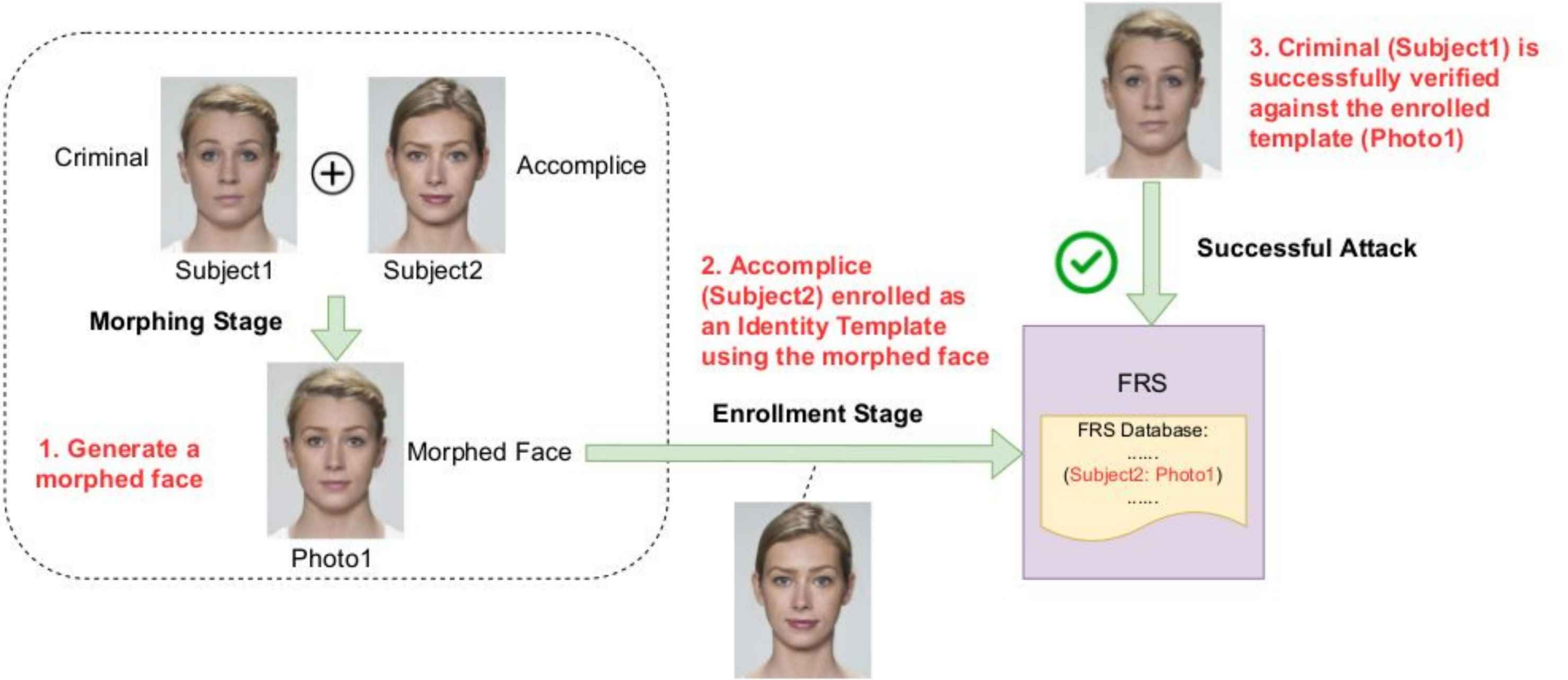
What's Morphing Attack

- Try to interfere with the operation of the FRS by presenting an attack at the time of enrollment



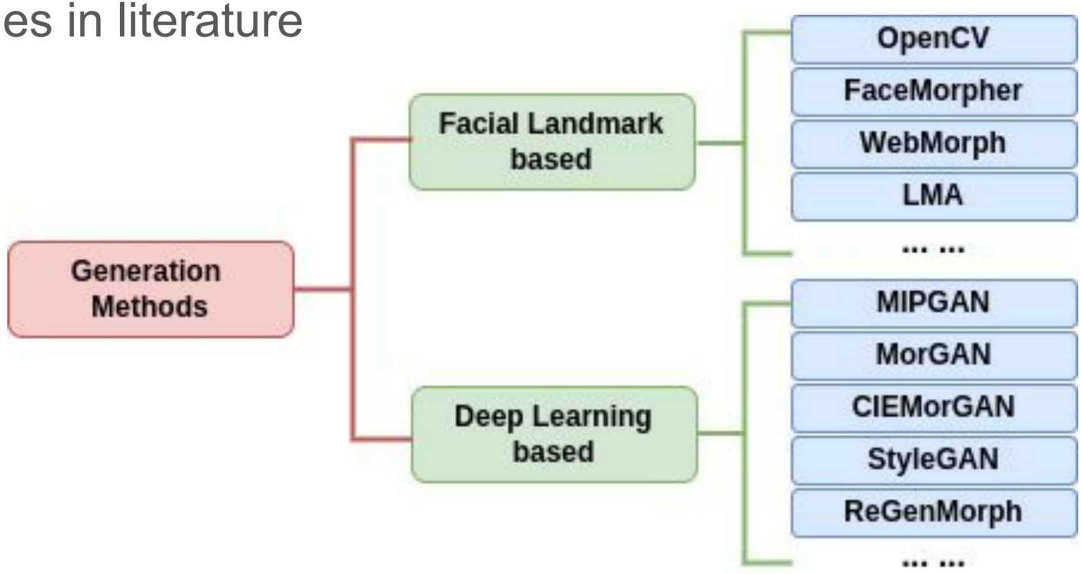
The system treats the criminal as the accomplice, and let him /her pass

Attack Procedure



Existing Morphing Tools/Techniques

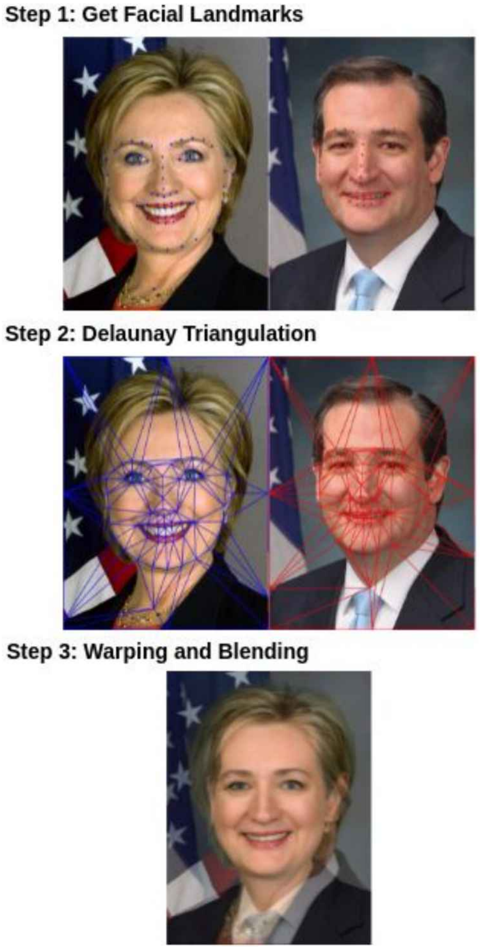
- Numerous easy-to-use morphing tools online
  - e.g., MorphThing, 3Dthis Face Morph, Face Swap Online, Abrosoft FantaMorph, FaceMorpher, MagicMorph
- Techniques in literature



Facial Landmark based

- Works by obtaining landmark points on facial regions
  - e.g., nose, eye, and mouth
- The landmark points obtained from two bona fide faces are warped by moving the pixels to different, more averaged positions
  - e.g. Delaunay triangulation
    - Affine transform
    - Alpha blending
- Post-processing
  - misaligned pixels generating artifacts
  - ghost-like artifacts

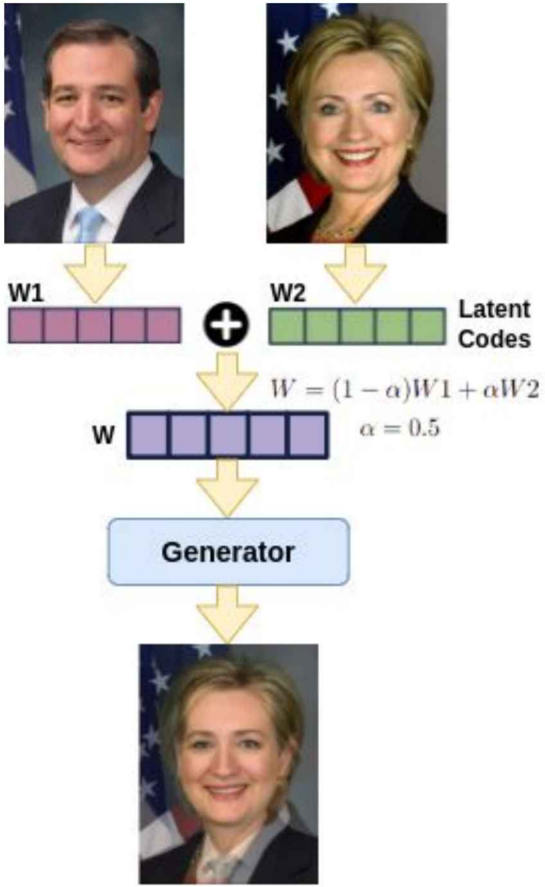
Delaunay triangulation



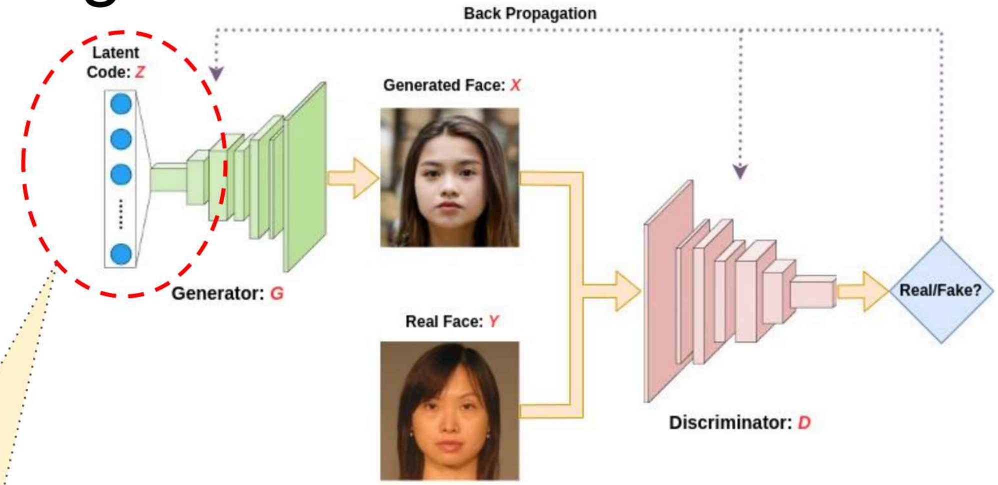
Deep Learning based

- Most are based on Generative Adversarial Networks (GAN)
- Most adopt CNN as basic architecture
- Works by embedding the images in the intermediate latent space
  - e.g. StyleGAN
    - Linear combination
    - Synthesize using Generator
- Post-processing if needed
  - Synthetic-like generation artifacts

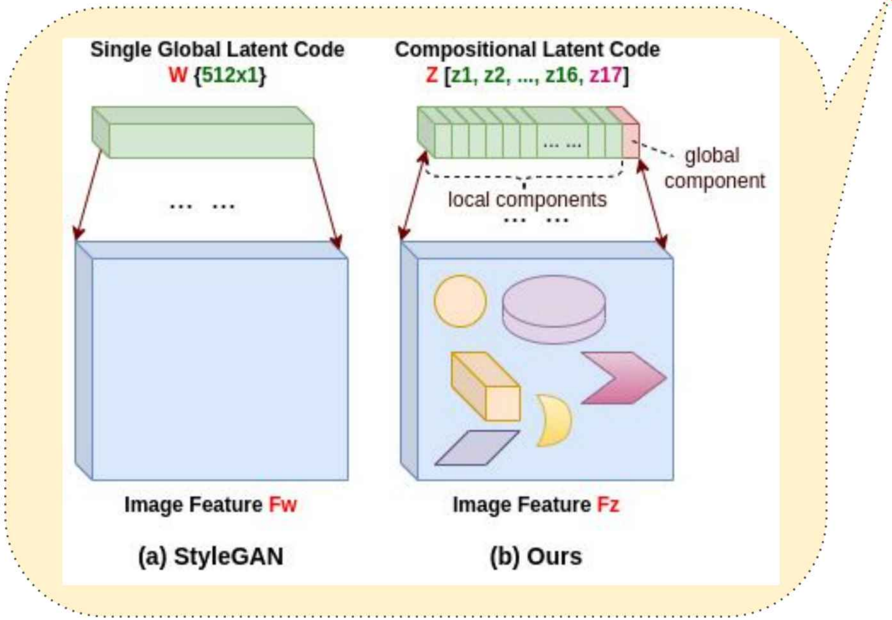
StyleGAN



Transformer based Morphing Attack

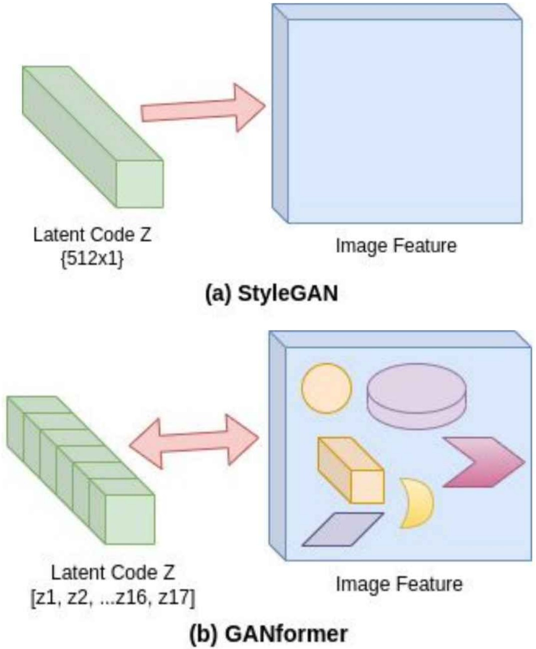


GAN Model



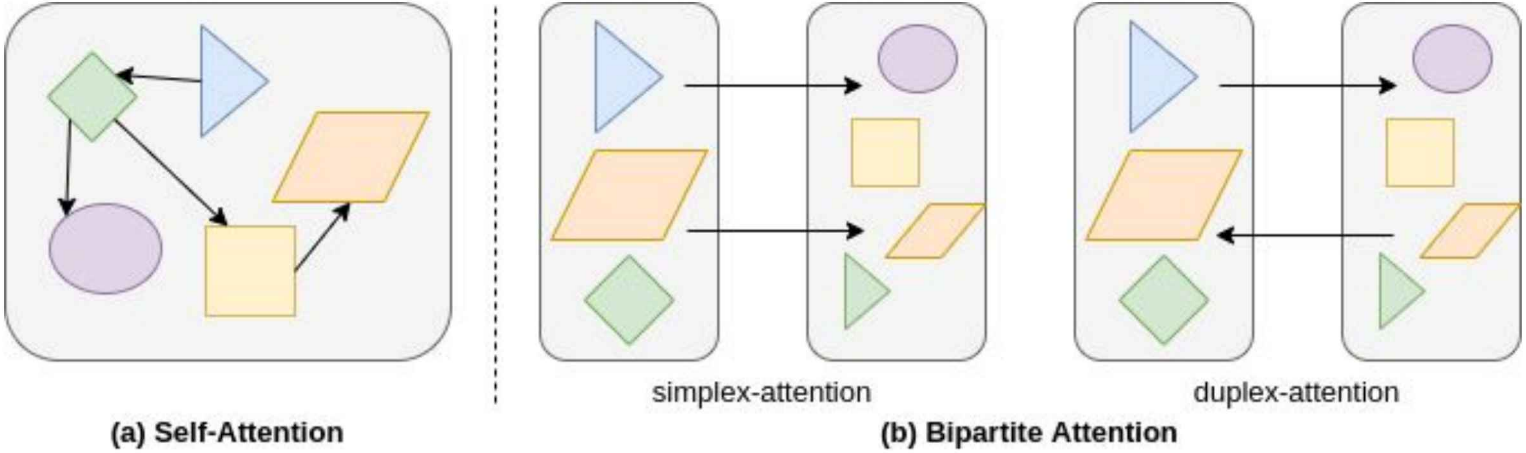
Hudson, Drew A., and Larry Zitnick. "Generative adversarial transformers." In *International conference on machine learning*, pp. 4487-4499. PMLR, 2021.

- Generative Adversarial Transformer (GANformer) [1]
- StyleGAN
  - Monolithic latent space
  - Single global style latent code
  - Modulate whole scene uniformly
  - In one direction
- GANformer
  - Compositional latent space
  - Multiple local style latent components
  - Impact different regions in the image
  - Spatially finer control
  - In both directions



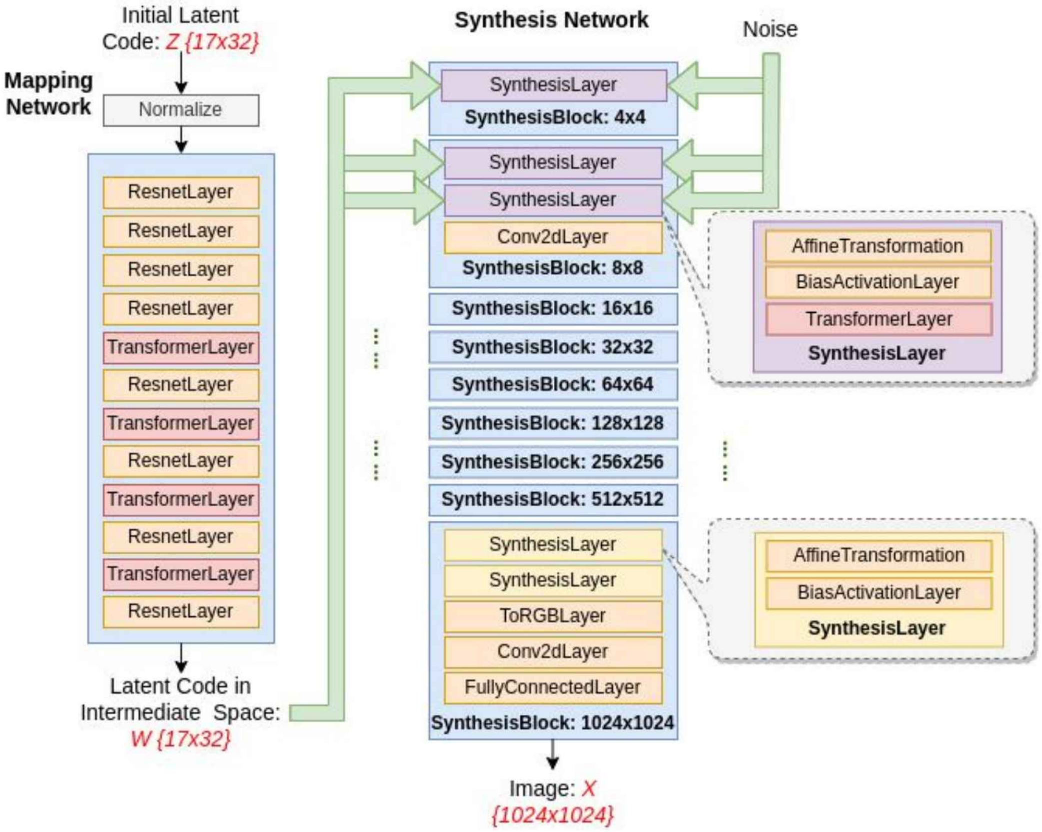
[1] Hudson, Drew A., and Larry Zitnick. "Generative adversarial transformers." International Conference on Machine Learning. PMLR, 2021.

Bipartite Transformer



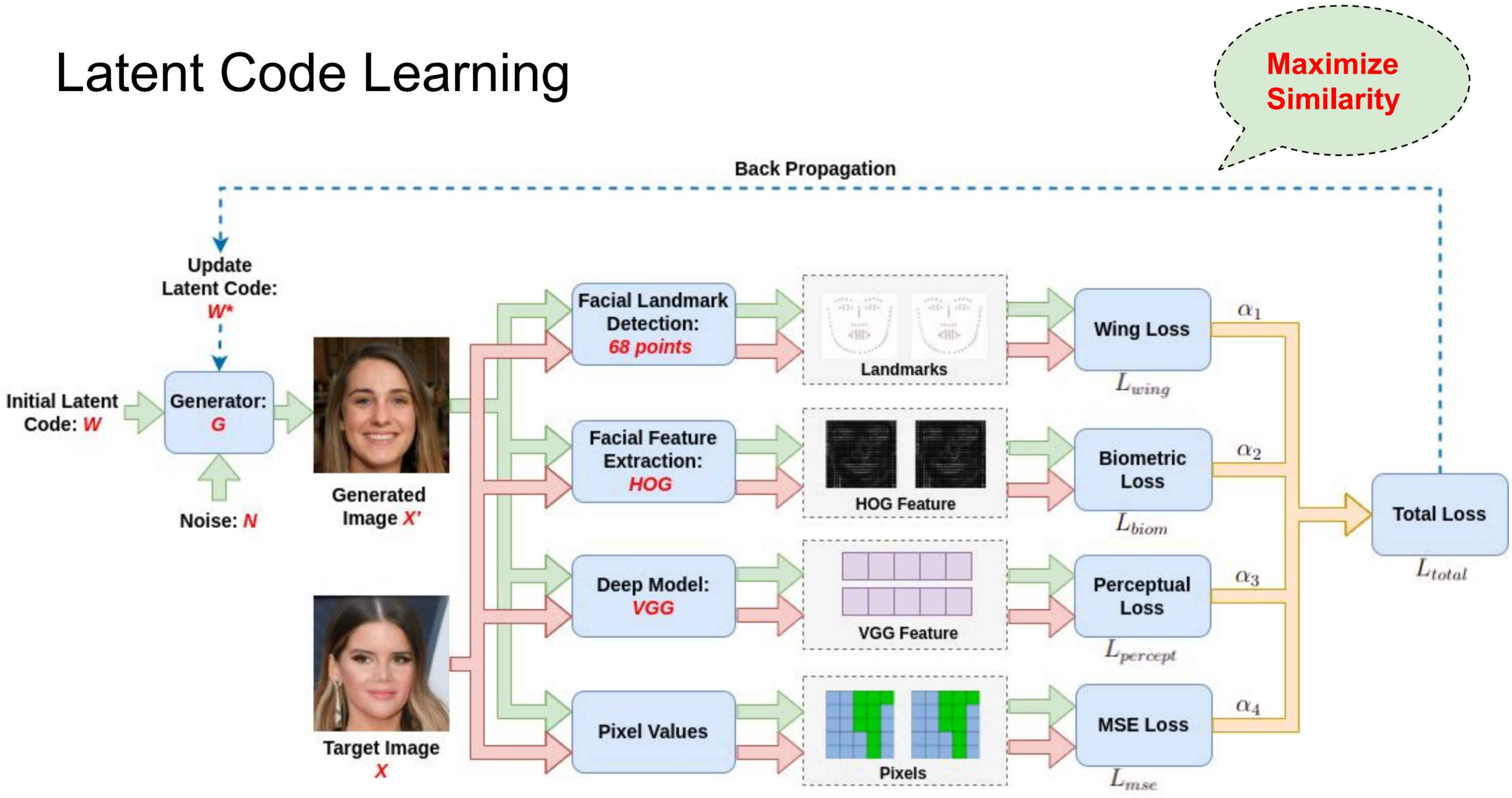
- **Traditional Transformer**
  - Self-attention with pairwise connectivity
  - Highly-adaptive
  - Around relational attention & dynamic interaction
  - Quadratic operation
- **Bipartite Transformer**
  - Two types
    - Simplex-Attention: one direction
    - Duplex-attention: bidirectional
  - Iteratively propagates information
  - Computation of linear efficiency

Architecture of the Generator



- Mapping network
  - feed-forward layers
  - receive a randomly sampled vector Z
  - output an intermediate vector W
- Synthesis network
  - W interacts directly with each transformer layer with added Gaussian noise to **modulate the evolving image features** X
  - Finally, W is transformed into an image X as the output of the synthesis network
  - 9 stacked synthesis blocks
  - from 4x4 grid to 1024x1024

Latent Code Learning



Loss Function

- Total loss

$$L_{total} = \alpha_1 L_{wing} + \alpha_2 L_{biom} + \alpha_3 L_{percept} + \alpha_4 L_{mse}$$

❖ Wing Loss

$$L_{wing} = \begin{cases} \beta ln(1 + |x|/\epsilon) & if|x| < \beta \\ |x| - C & otherwise \end{cases}$$

|x|: means the magnitude of the **gradients between the landmark points** of generated and target images

❖ Biometric Loss

$$L_{biom} = 1 - \frac{HOG_{G(w)} \cdot HOG_I}{\|HOG_{G(w)}\| \|HOG_I\|}$$

The distance between two faces is computed using the **cosine similarity score based on HOG features**

❖ Perceptual Loss

$$L_{percept}(G(w), I) = \sum_{j=1}^{4} \frac{\lambda_j}{N_j} \|F_j(G(w)) - F_j(I)\|_2^2$$

measure the high-level similarity between images perceptually based on Fj – – the **output feature of VGG-16 in layers**: conv1_1, conv1_2, conv3_2 and conv4_2, respectively. Nj is the number of scalars in the j-th layer output

❖ MSE

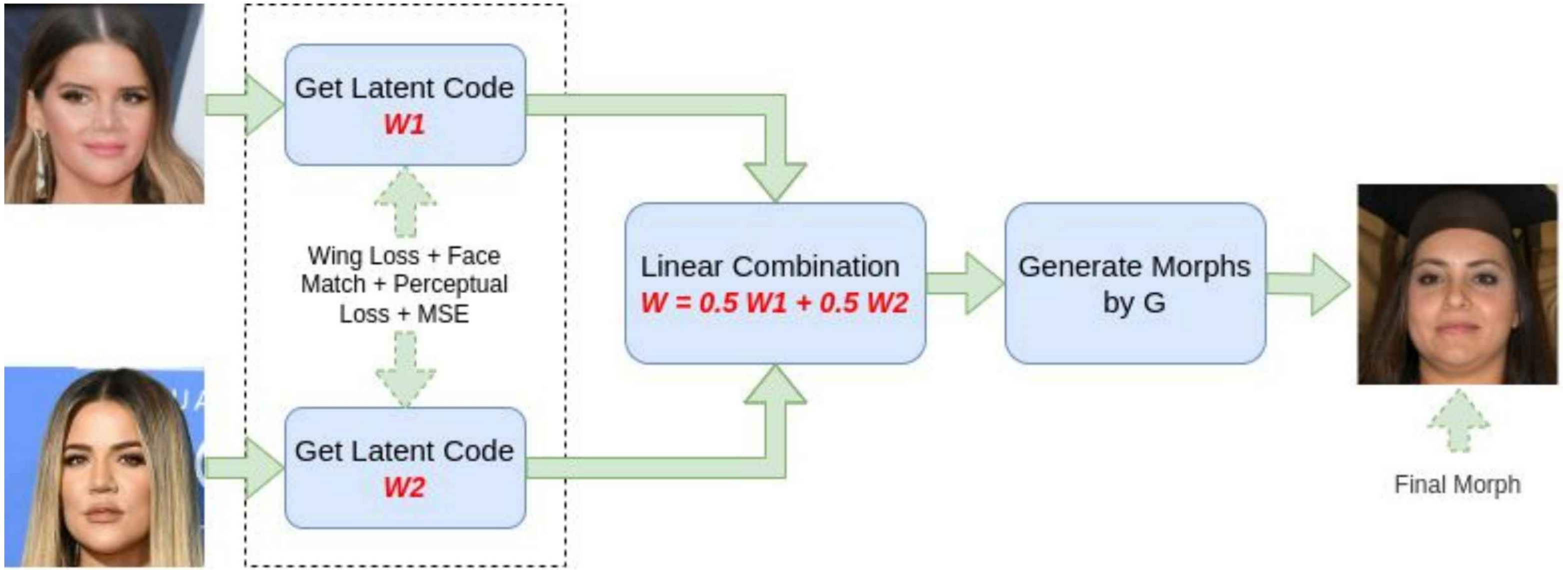
$$L_{mse}(G(w), I) = \frac{1}{N} \|G(w) - I\|_2^2$$

Pixel-level Mean square error.
N is the number of scalars of the image

Face Morphing

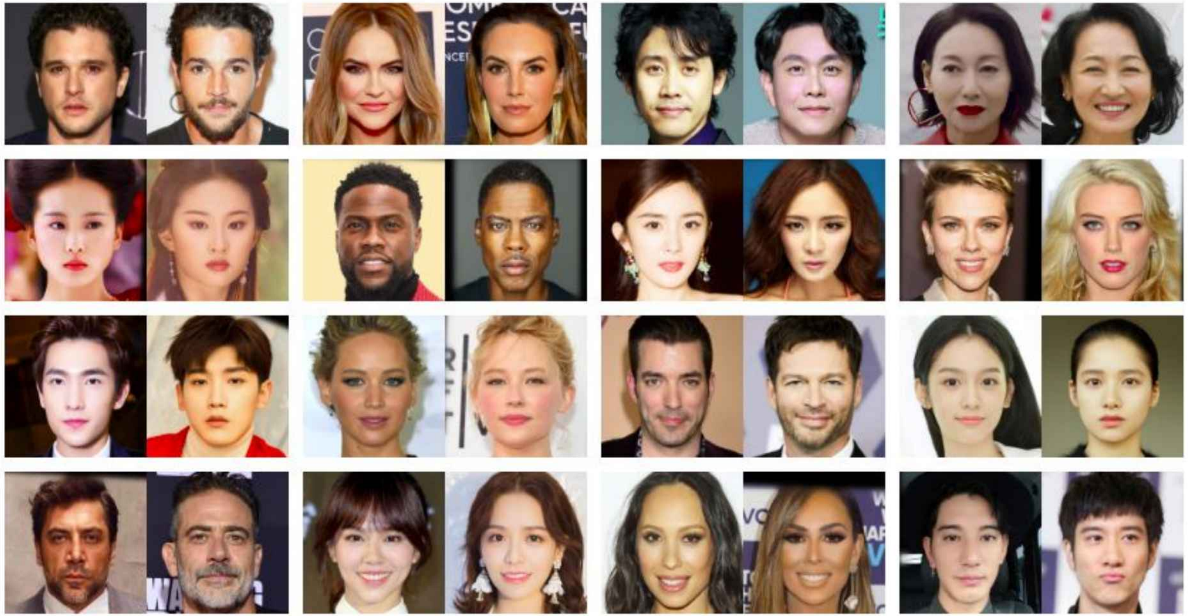
- Given two face images I_1 and I_2 , with their respective latent vectors W_1 and W_2
- Face morphing is performed by a linear interpolation:

$$W = \lambda W_1 + (1 - \lambda) W_2, \lambda \in (0, 1)$$

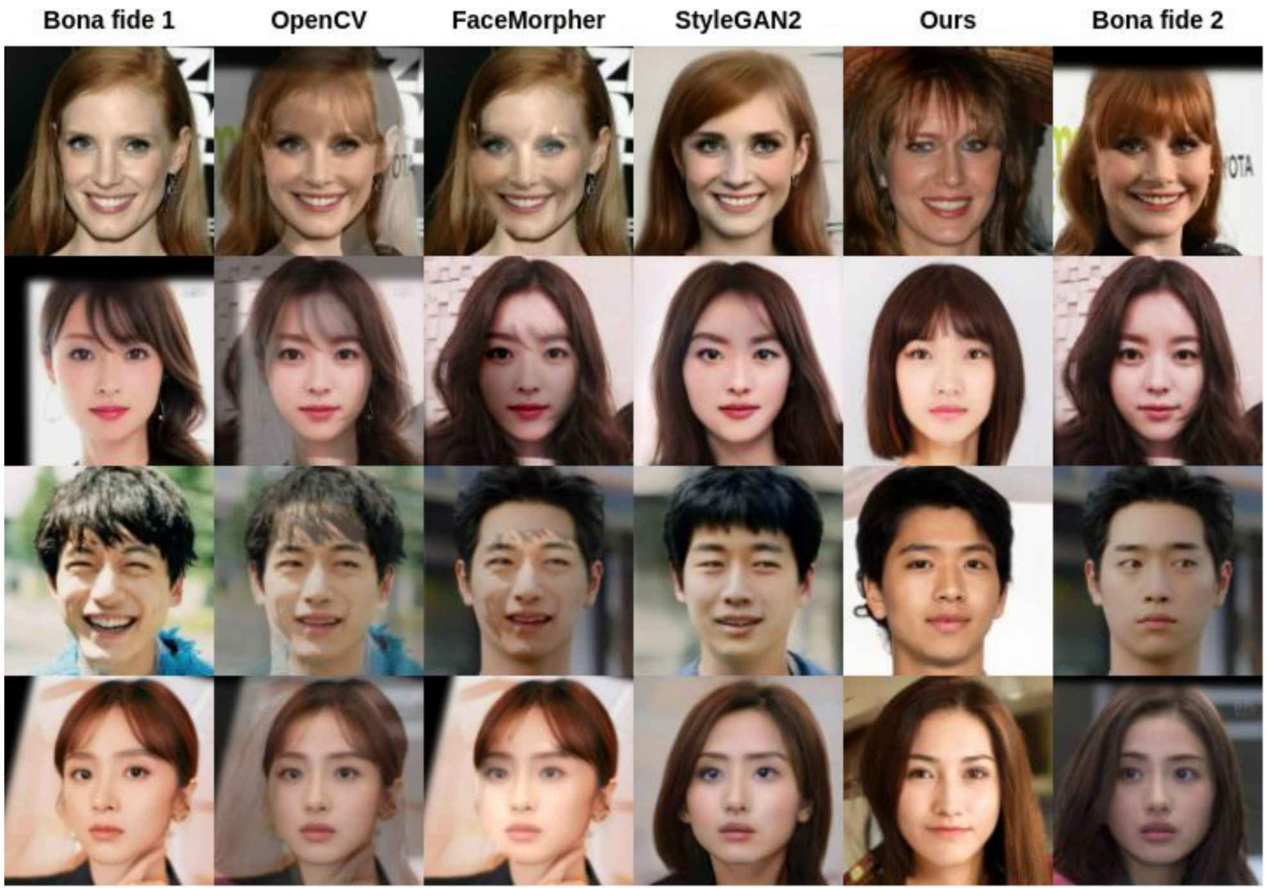


Bona Fide Faces

- Doppelgänger Face Pairs
  - Celebrities that appear similar
  - Same gender and ethnicity
  - 153 pairs
  - 1024x1024



Morphed Result



- OpenCV/ FaceMorpher:
  - misaligned pixels generating artifacts
  - ghost-like artifacts
- StyleGAN2
  - Synthetic-like generation artifacts
- **Ours**
  - More visibly realistic
  - More natural

Vulnerability Test

- On 3 FR models
- Ideally, a strong morphing attack will have a high similarity score to the target identities
- Ours
  - Have same or even better ability to preserve the characteristic of identities
  - Also can generated visually realistic and natural faces

Mated Morphed Presentation Match Rate (MMPMR) - (%) at FMR=0.1%

Method	ArcFace	FaceNet	LBP
OpenCV	94.73	82.23	87.50
FaceMorpher	81.21	73.83	87.92
StyleGAN2	84.21	70.65	85.52
our-FaceNet	56.58	50.53	82.11
our-ArcFace	53.29	47.24	80.79
our-LBP	50.66	43.95	90.00
our-Percept	53.29	43.95	78.82
our-Percept+Wing	82.24	59.08	88.68
our-Percept+Wing+MSE	84.87	62.37	89.34
our-HOG	77.63	45.92	86.71
our-HOG+Percept	86.18	59.74	88.03
our-HOG+Percept+Wing	85.53	61.05	88.03
our-HOG+Percept+Wing+MSE	90.08	70.92	89.77

Limitations

- Local minimum of loss
  - Not all the optimization can lead to good results
  - Sometimes the learning converges on local minimum
- Time of learning latent code
  - Around 8 minutes with 1500 gradient descent steps per image