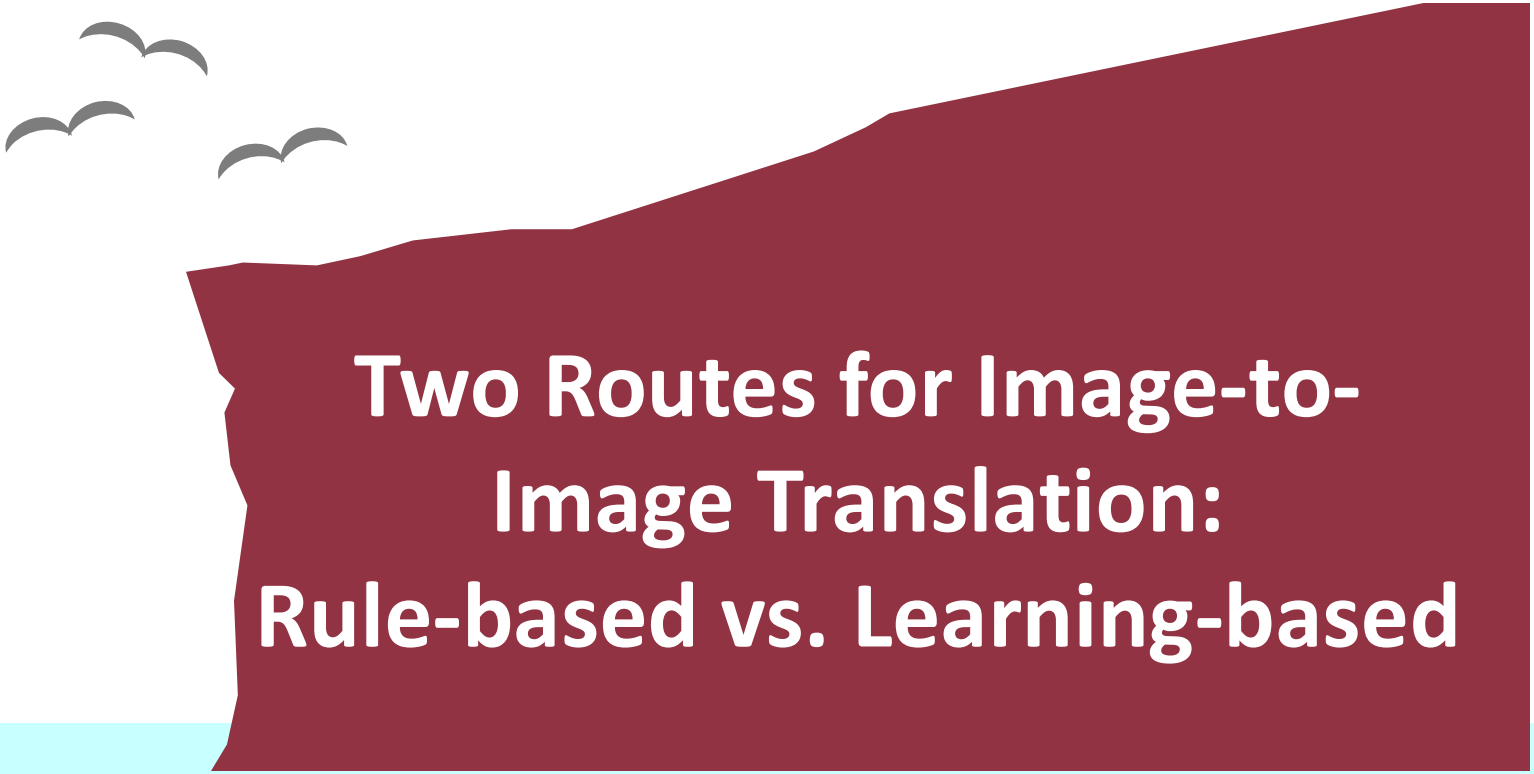


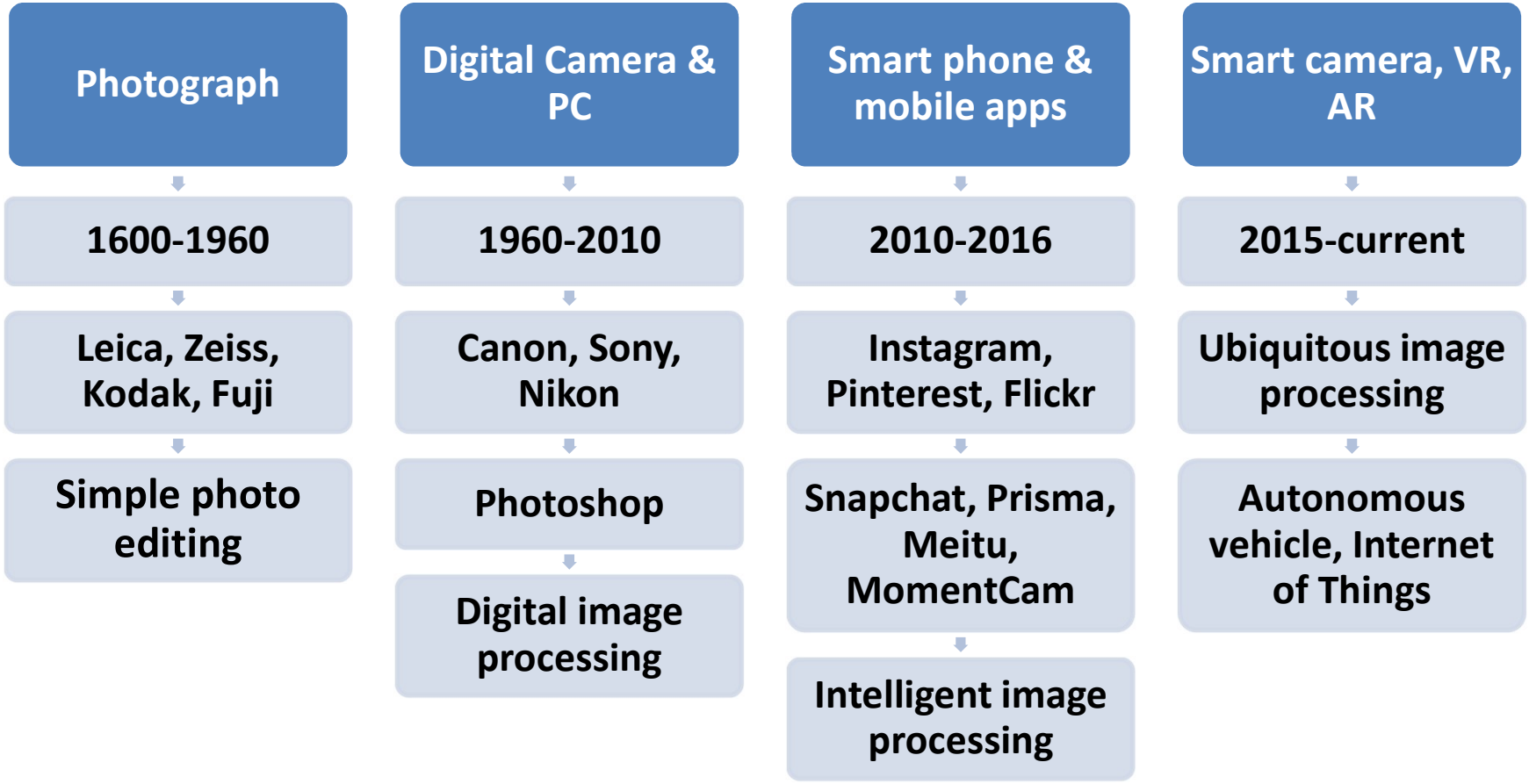
# Two Routes for Image-to-Image Translation: Rule-based vs. Learning-based

Minglun Gong, Memorial Univ.

Collaboration with Mr. Zili Yi

# Outline

- Introduction
  - A brief history of image processing
  - Image-to-Image translation problem
- Two approaching routes
  - A rule-based approach
  - A data-driven approach




# Brief History of Image Processing

- **Many problems in image processing, computer graphics, & computer vision can be posed as translating an input image to a corresponding output image**

| Problem | Input | Output |
|---|---|---|
| Edge detection | Natural image | Edge map |
| Image segmentation | Natural image | Label map |
| Saliency detection | Natural image | Saliency map |
| Image colorization | Grayscale image | Color image |
| Image stylization | Natural image | Stylized image |
| Sketch-to-photo synthesis | Sketch | Natural image |

# Image-to-Image Translation

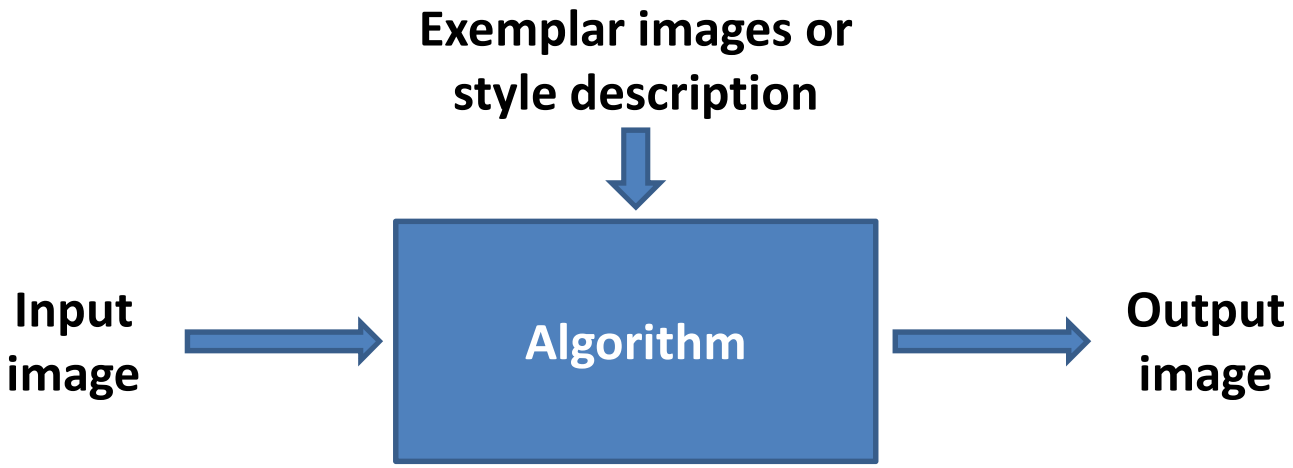
# Two Approaching Routes

- **Rule-based:**
  - **Different tasks call for different algorithms**
  - **Algorithms are customized for the tasks**
- **Learning-based:**
  - **Different tasks can be solved under the same framework**
  - **Data-driven**

## Rule-based Approach:

# Face Photo Stylization based on a Single Exemplar

- **Image stylization aims to automatically generate stylized images by manipulating photographs**

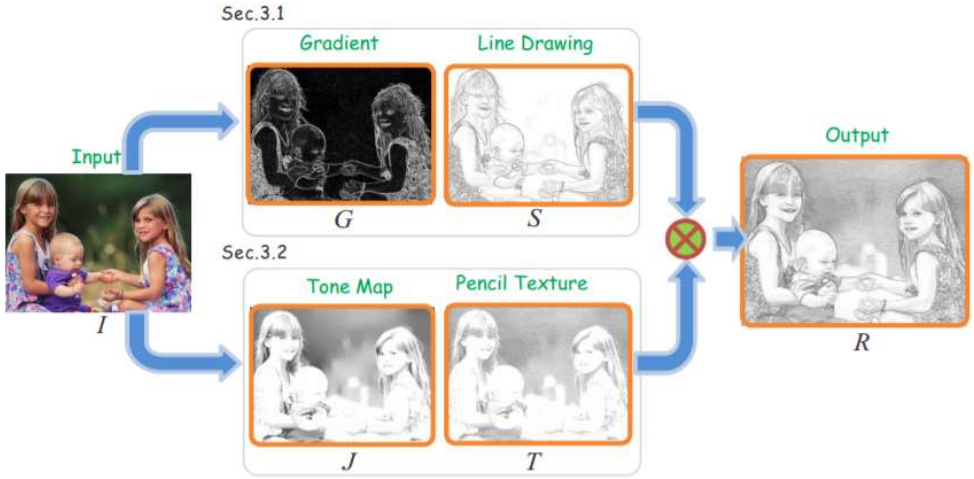



# Problem Statement

- By output type
  - Photorealistic vs. non-photorealistic
- By requirement of external resources
  - Example-based, data-driven, interactive
- By extent of automation
  - Semi-automatic vs. fully-automatic
- By level of understanding/processing
  - Low-level, middle-level, high-level stylization

# Categorization of Previous Methods

- **Combining Sketch and Tone for Pencil Drawing Production [Liu et al., NPAR 2012]**
  - **Fully-automatic**
  - **Non-exemplar-based**
  - **Non-photorealistic**

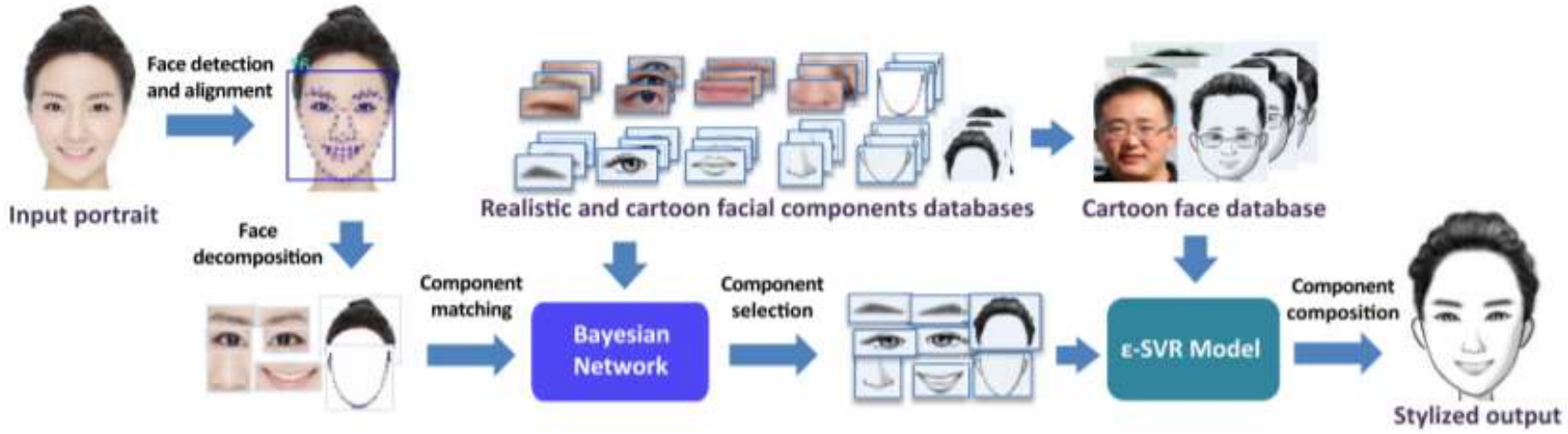



# Pencil Drawing Production

- **Automatic Content-Aware Color and Tone Stylization [Lee et al., CVPR 2016]**
  - **Data-driven**
  - **Fully-automated**
  - **photo-realistic/naturalistic**


# Color and Tone Stylization

- **Data-Driven Synthesis of Cartoon Faces Using Different Styles [Zhang et al., TIP 2017]**
  - **Data-driven**
  - **Fully-automatic**
  - **Non-photorealistic**




# Cartoon Face Synthesis

- **Most are style-specific methods:**
  - **Use heuristically designed algorithm to generate results under a given style**
    - **Pencil drawing, stippling, etc.**
  - **Cannot be applied to general-purpose**
- **Some are data driven methods:**
  - **Require a large number of training data**

# Limitations of Existing Methods

- **Our goal is to achieve general stylization with a single exemplar image**

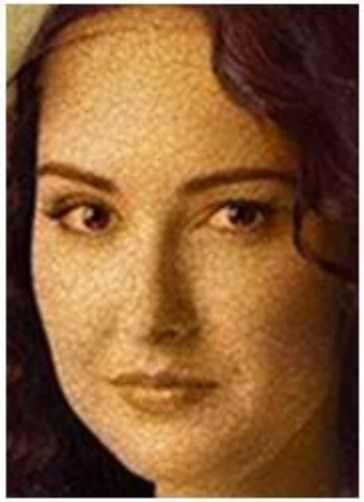
# Our Objective

$I_{input}$

Color Transfer

Texture Transfer

Example



# Two Step Approach

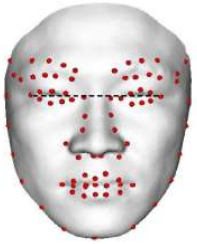
- **Assigns a color for each pixel in the input image by finding its correspondence in the exemplar**
- **The correspondence is found by minimizing a cost function that consists of 3 terms:**
  - **Semantic term**
  - **Geometry term**
  - **Color term**

## Semantics-aware Color Transfer

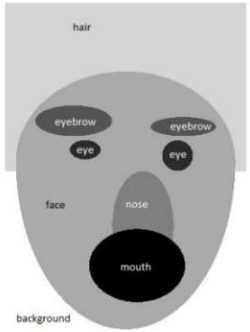
face detection & normalization

320×448 image

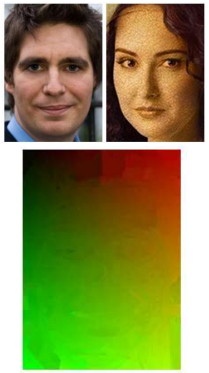
Face landmark



label assignment




dense correspondence



Color transfer

# Semantics-aware Color Transfer

- **Evaluates the incompatibility among different facial parts**
  - **Heuristically defined**
  - **Prevent incompatible mapping, e.g. eyes in input image to nose or mouse in exemplar**

| | Eye | Eyebrow | Nose | Mouth | Face |
|---|---|---|---|---|---|
| Eye | 0 | ∞ | ∞ | ∞ | 1 |
| Eyebrow | ∞ | 0 | ∞ | ∞ | 1 |
| Nose | ∞ | ∞ | 0 | ∞ | 0 |
| Mouth | ∞ | ∞ | ∞ | 0 | 1 |
| Face | ∞ | ∞ | 0 | 1 | 0 |

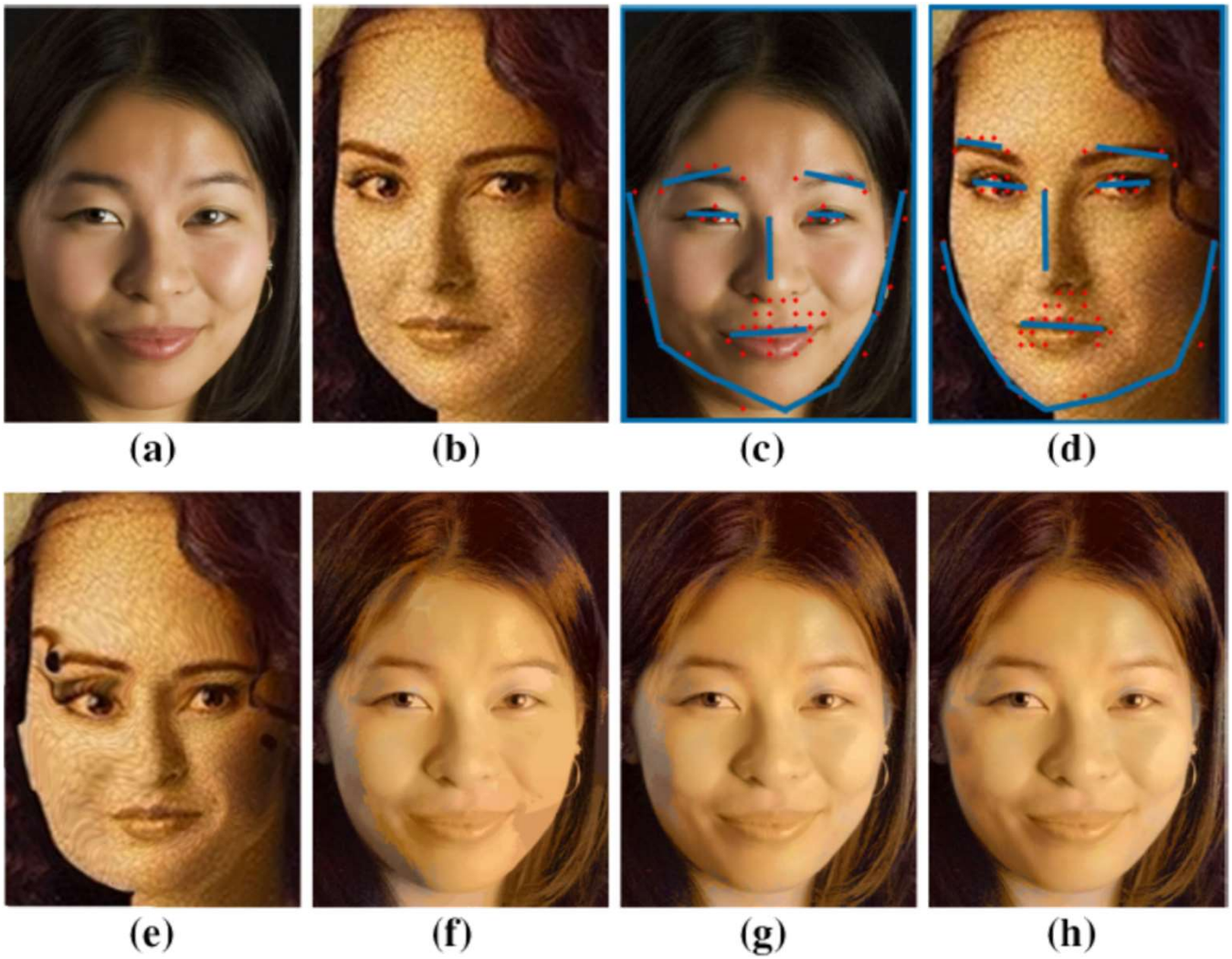
## Semantic Term

- **Measures the geometric cost between the pixel in input image and its correspondence in exemplar image**
- **Directly using the Euclidean distance does not tolerate the pose and shape differences**
  - **Warp the exemplar to align with input face before computing the geometry term**

# Geometry Term

- **Measures the color cost between a pixel in input image and its correspondence in exemplar**
- **To accommodate the overall intensity differences between 2 images, histogram equalization is performed first**

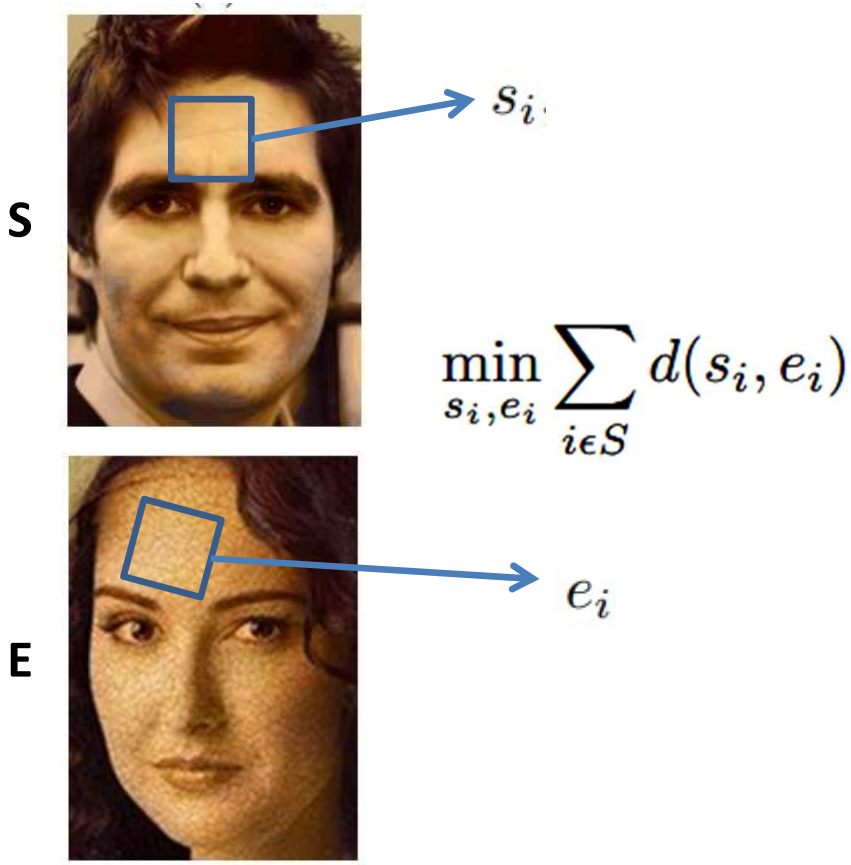
# Color Term




# Effects of Different Terms

- **The generation of paintings and artworks is treated as a texture synthesis problem**
  - **Need to handle texture at multiple scales**
    - **Textures show up many scales and may have distinct characteristics at each scale**
  - **Need to preserve edges**
    - **Chin line, eye/eyebrow boundaries, and mouth/nose curves are essential for keeping face identities**

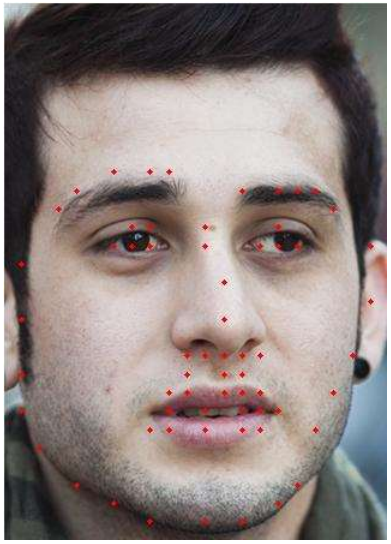
## Edge-preserving Texture Transfer

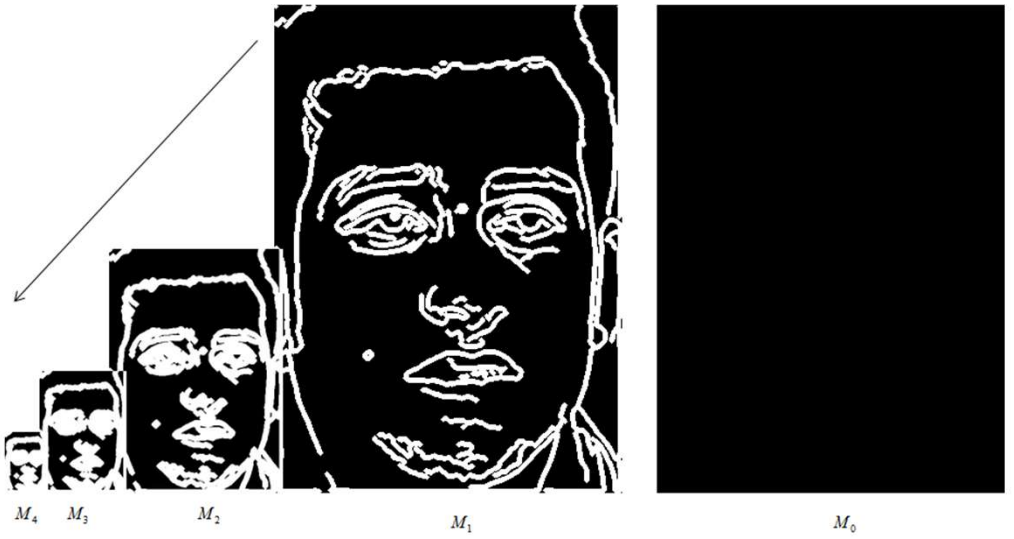



- **First optimize:**
  - **Finding the Nearest Neighbor Field (NNF) using generalized PatchMatch**
- **Then vote:**
  - **Averaging the nearest neighbors**

# Optimization-based Texture Synthesis

- **To preserve edges and structures:**
  - **First compute an edge map for the input image**
  - **Then create its lower resolution versions**
  - **These edge pixels are masked from texture synthesis**



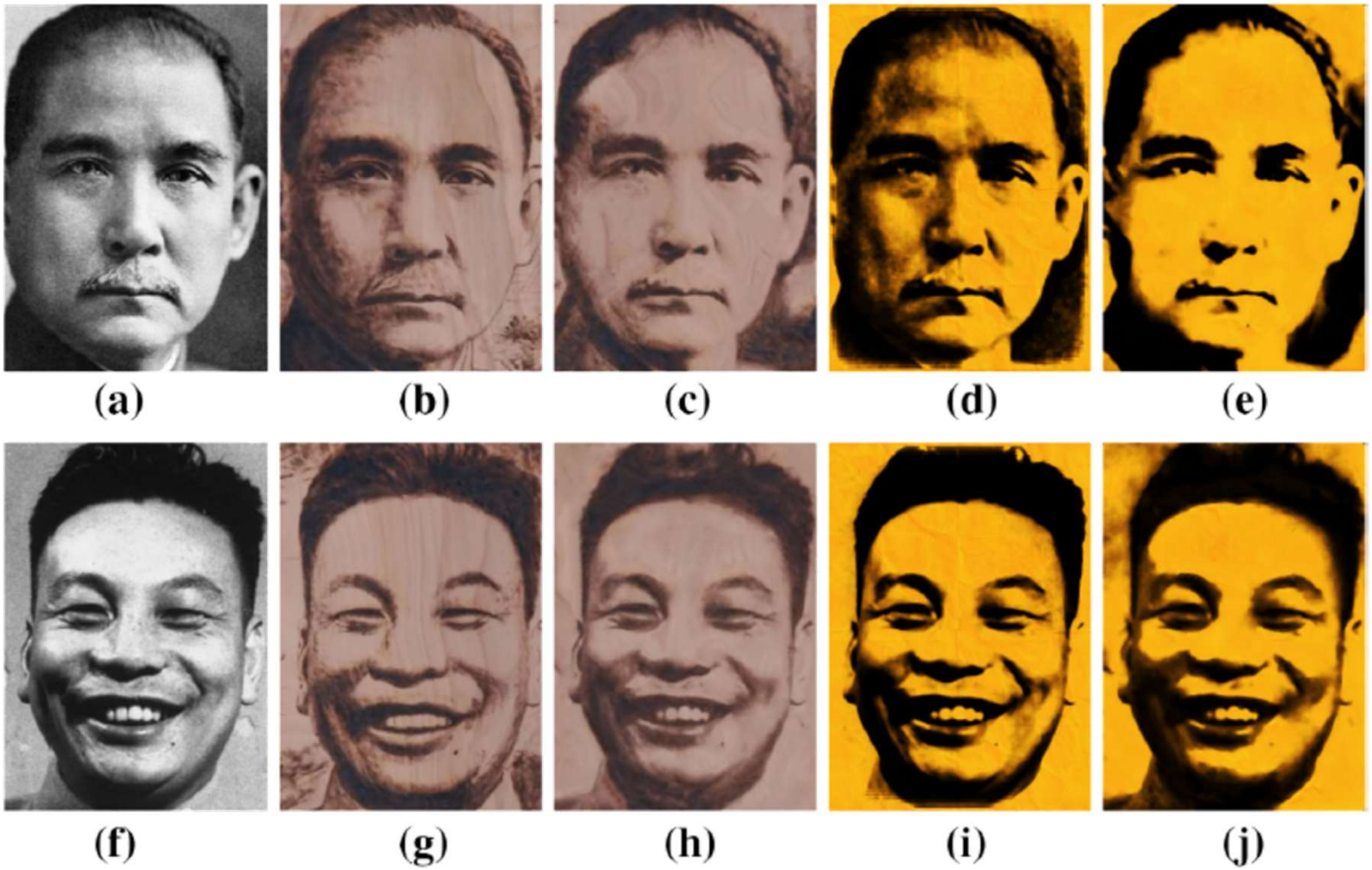



# Edge Masks Generation

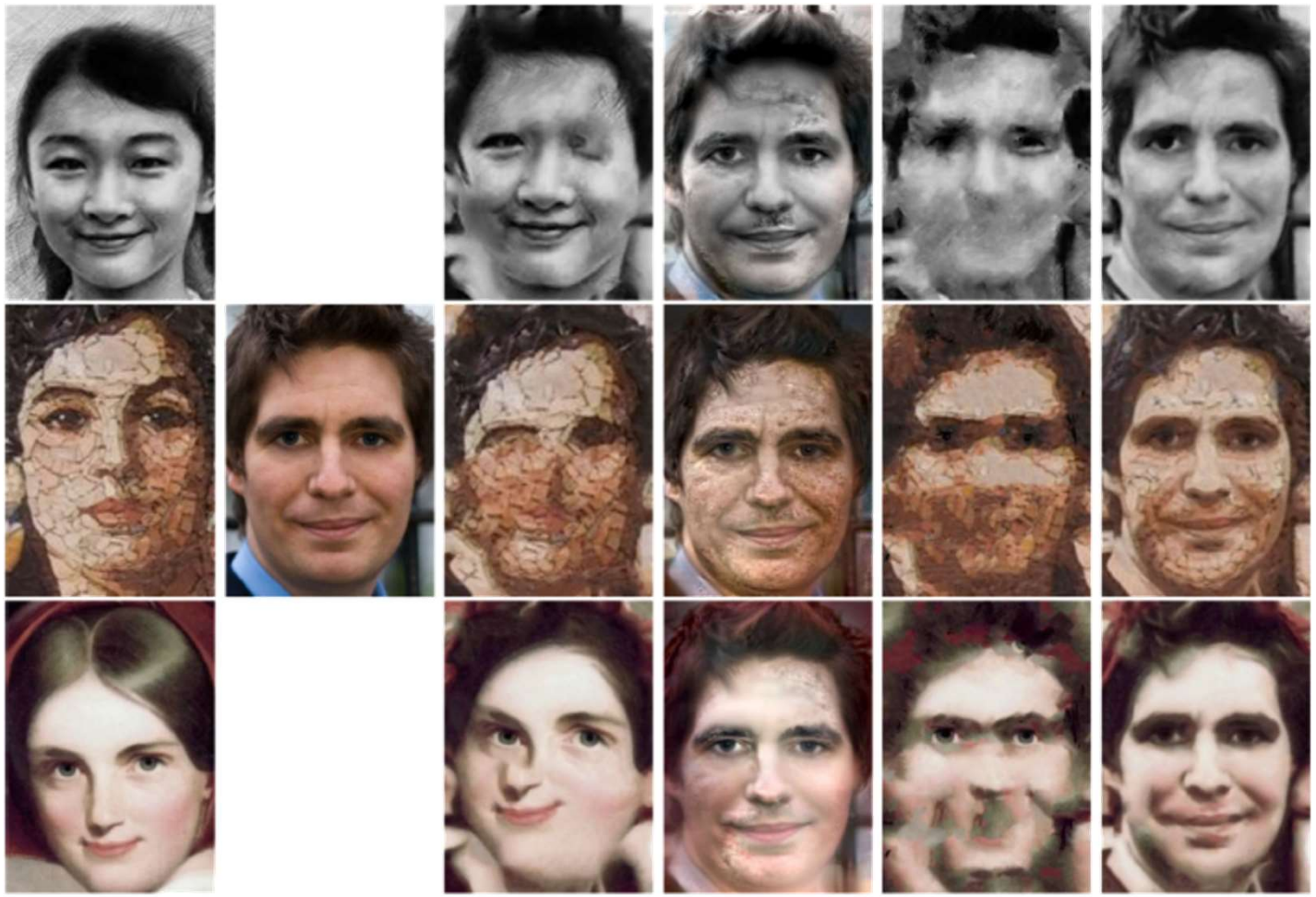
- **The texture is synthesized in a pyramid manner:**
  - **First synthesize the lower resolution version and for unmasked areas only**
  - **Enlarge the result image and repeat texture synthesize**


# Coarse-to-fine Processing

# Experiment Result




# Comparison with Real Artworks



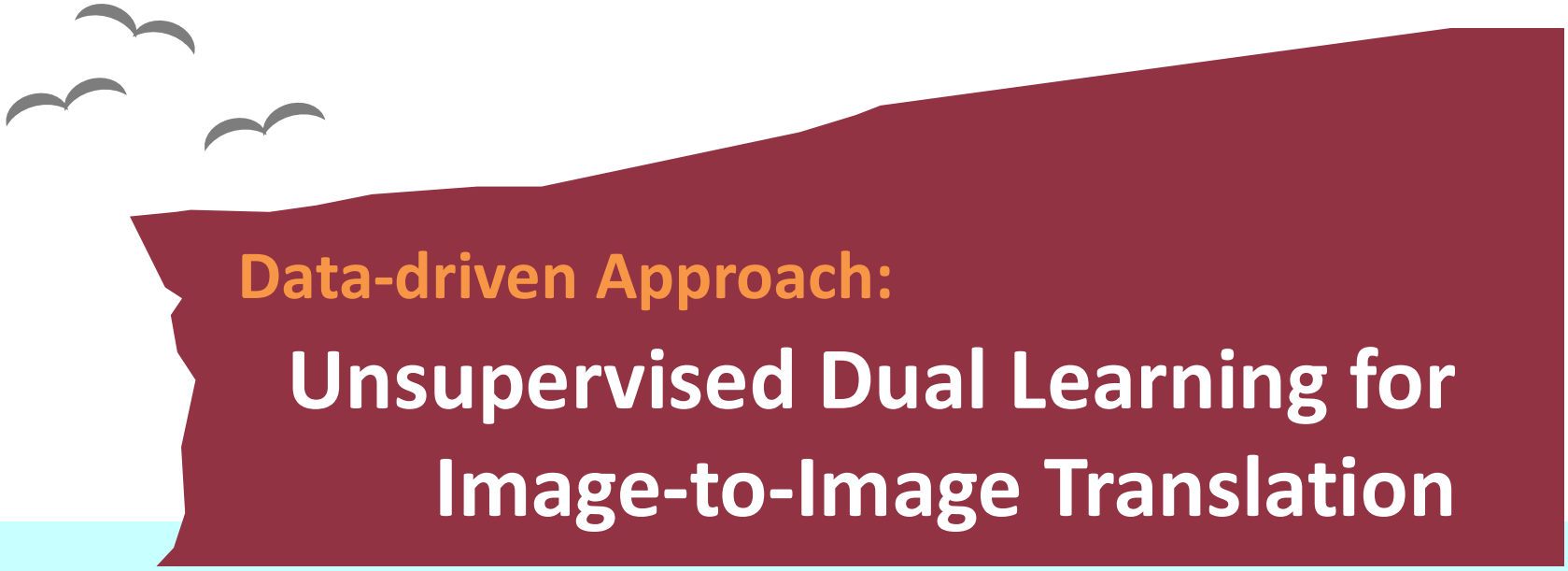

# Comparison with Image Melding & Quilting

- **A general face photo stylization approach that requires a single exemplar**
  - **Able to handle a wide range of face photos**
    - **Gender, skin color, hair style, face accessories, beards, glasses, variation in poses & lightening conditions**
  - **Able to transfer wide varieties of styles**
    - **Pencil drawing, sand drawing, oil painting, mosaic, screening, water color painting, Chinese painting, & Pyrography**

## Conclusions

- **Employed semantic information to guide the stylization**
  - **Only work for face photos**
- **The algorithm is heuristically designed**
  - **Not an end-to-end system**

# Limitations



## Data-driven Approach:

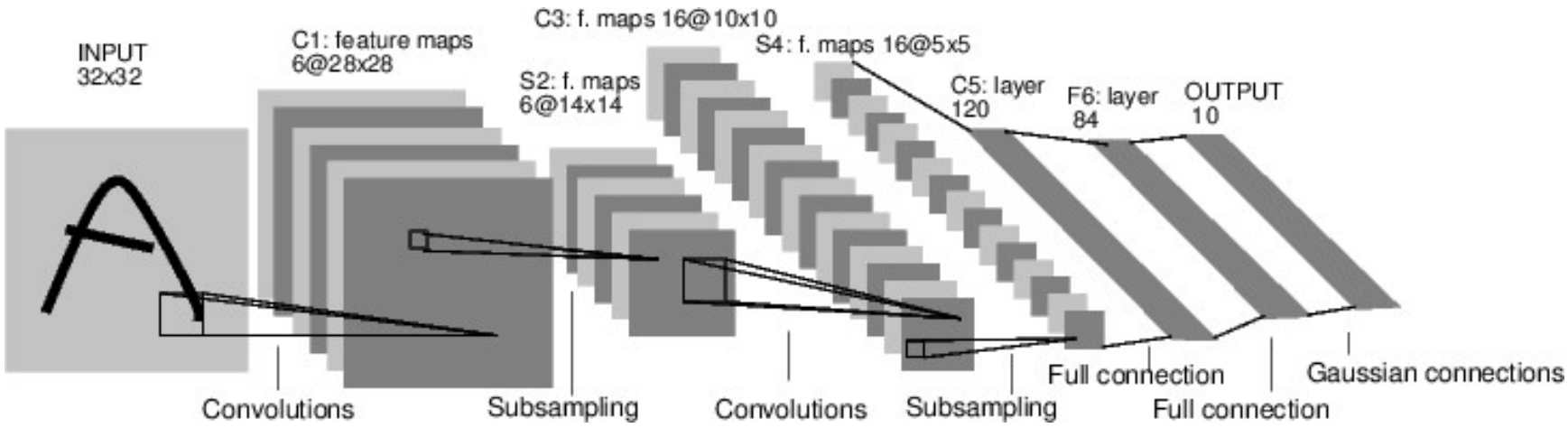
# Unsupervised Dual Learning for Image-to-Image Translation

# Advantages of Deep Neural Network

- **End-to-end training**
  - A DNN can be treated as a black box
  - Only need to fuel the networks with training data
- **General-purpose solution**
  - A solution to one task can often be adapted to another task, as long as new training data are provided
- **Semantic-aware**
  - Utilize the pattern recognition power of DNN

- **Before 2014: traditional methods**
- **2014-2016: Application-specific solution**
  - **FCNs**
  - **GANs**
  - **cGANs**
- **2016: general-purpose solution**
  - **cGAN (supervised learning)**
  - **Cross-domain (pre-trained third-party representations)**

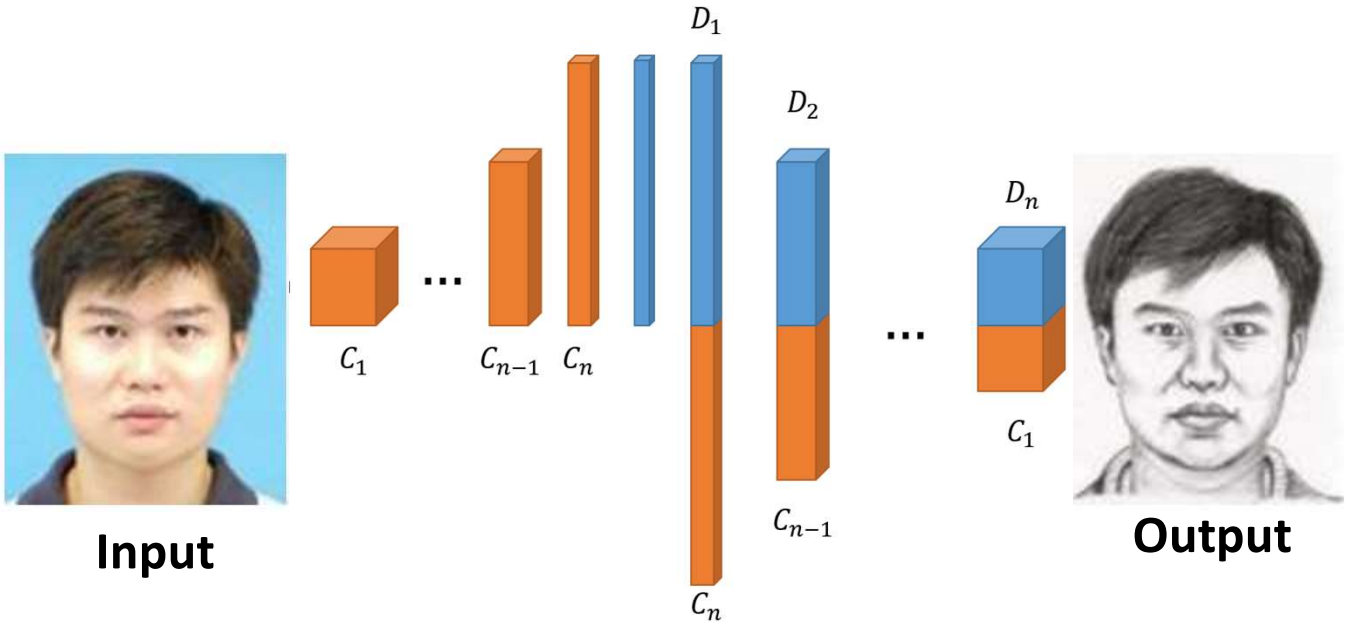
## Related Work




- **LeNet for document recognition [LeCun et al]:**
  - **Combines convolution layers & subsampling layers**
  - **Connections are between local neurons**
    - **Avoid full association**
  - **Sharing of weights**

# Convolutional Neural Network (CNN)

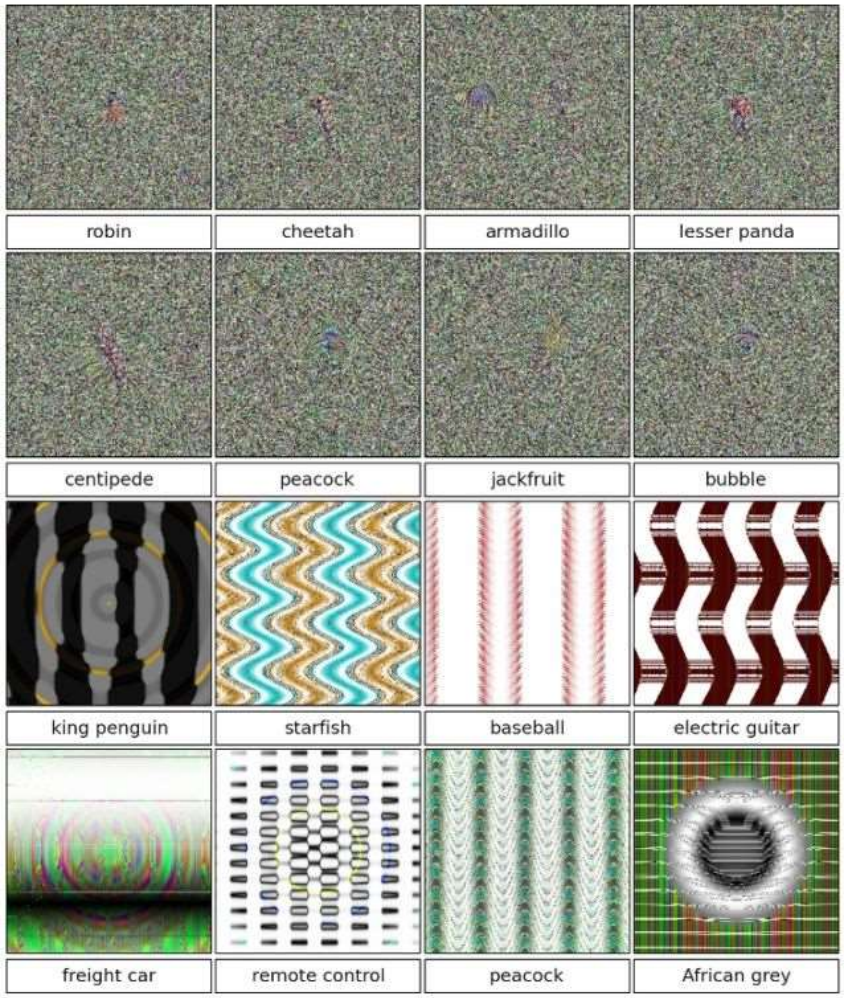
- **DNN can also be used for training generative models**
  - **Learning deconvolution network for semantic segmentation [Noh, et al. ICCV 2015]**




# Fully Convolutional Networks (FCN)

# Adversarial Samples

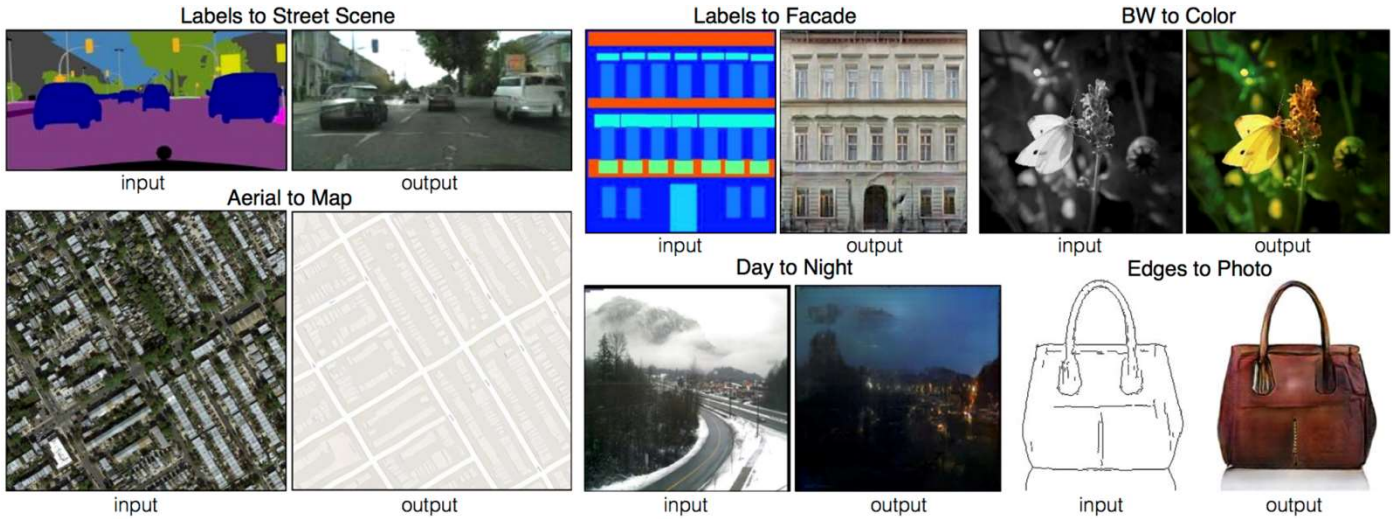
- **It is easy to generate adversarial samples to fool a discriminative DNN**
  - **Adversarial samples can make the discriminative DNN more robust**
  - **More efforts are then needed to generate adversarial samples**



- **GANs extend the idea of using to generative DNN:**
  - **Train Generator & Discriminator together and against each other**
    - **Generator generates fake samples to fool the Discriminator**
    - **Discriminator tries to distinguish between real & fake samples**
  - **Repeat this and we get better Generator and Discriminator**

# Generative Adversarial Networks (GAN)

- **Image-to-image translation with conditional adversarial networks [Isola, et al 2017]**
  - **Learn both the mapping from input image to output image and a loss function to train this mapping**
  - **Makes it possible to apply the same generic approach to different problems**

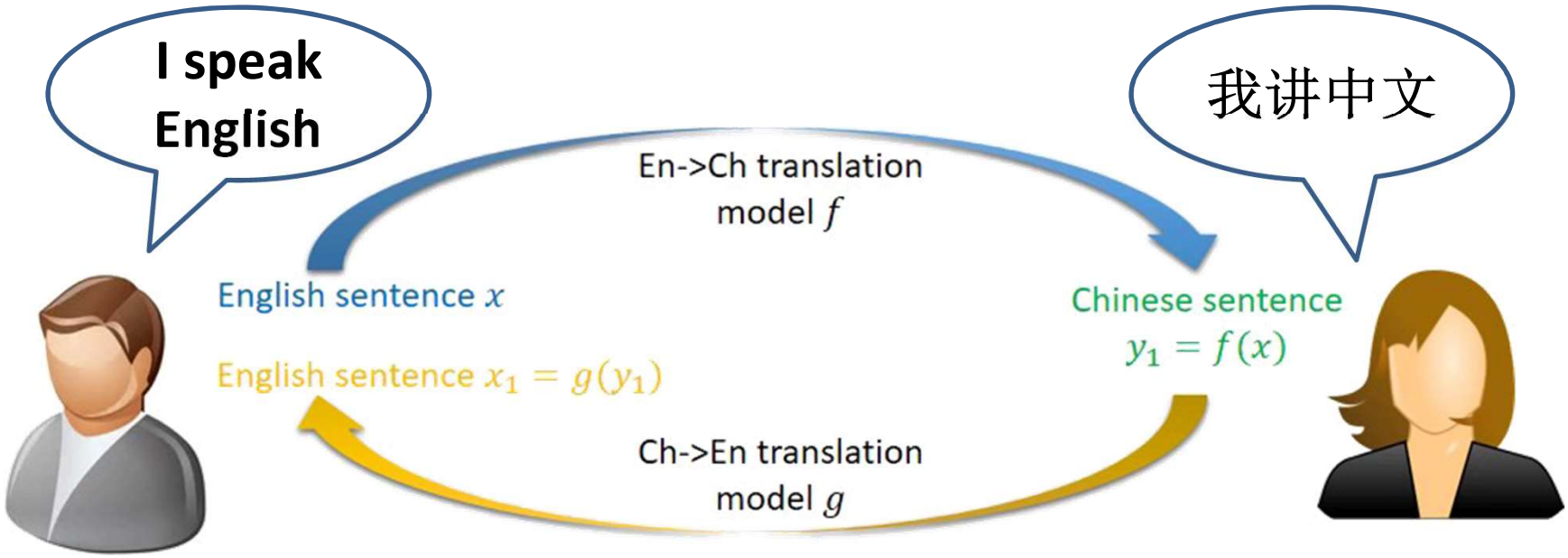



# Conditional GAN (cGAN)

# Limitations of Existing Methods

- **Requires a large number of labeled & matching image pairs**
  - **Data labeling is expensive & sometimes impractical**
  - **Matching image pairs could be misaligned or contain slightly different contents**

- **Dual learning for machine translation [He et. al 2016]**

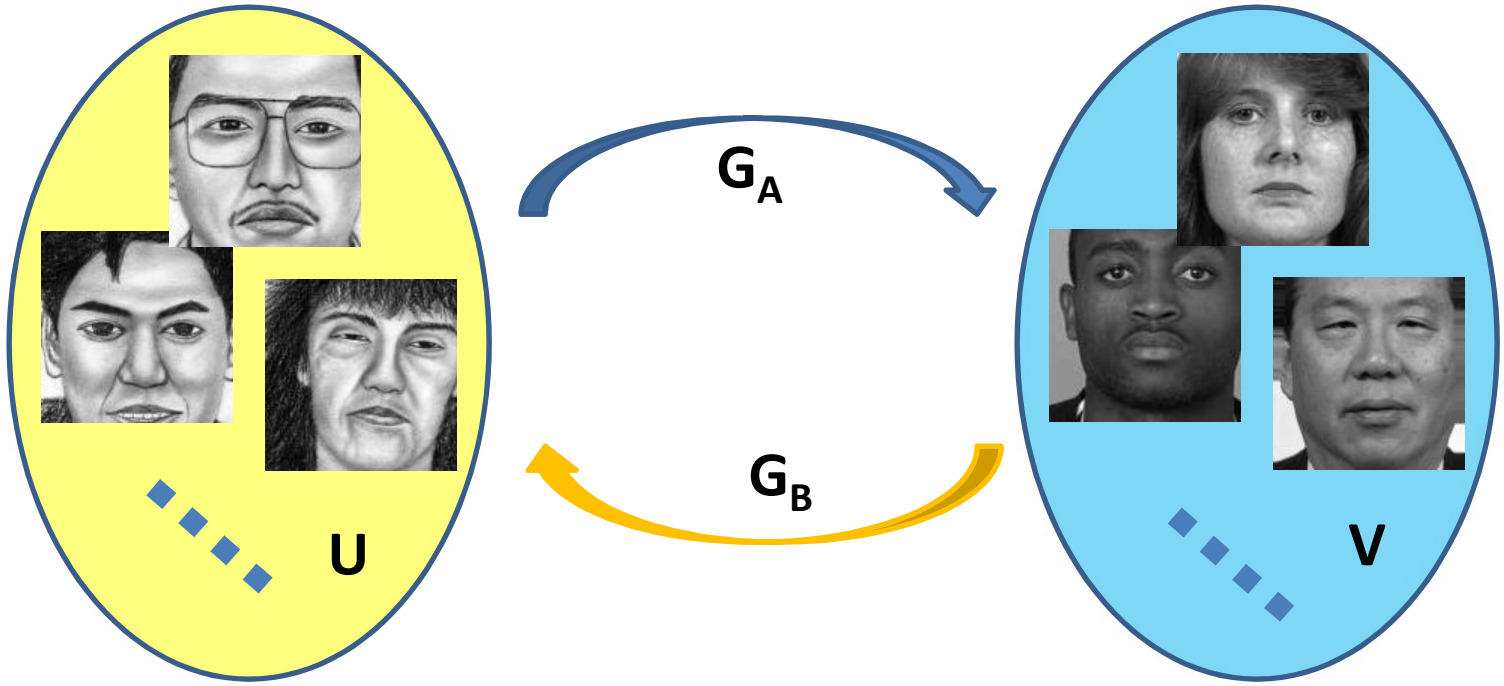



## Dual Learning

- **Develop an unsupervised learning framework**
  - **For general-purpose image-to-image translation**
  - **Only relies on unlabeled image data**

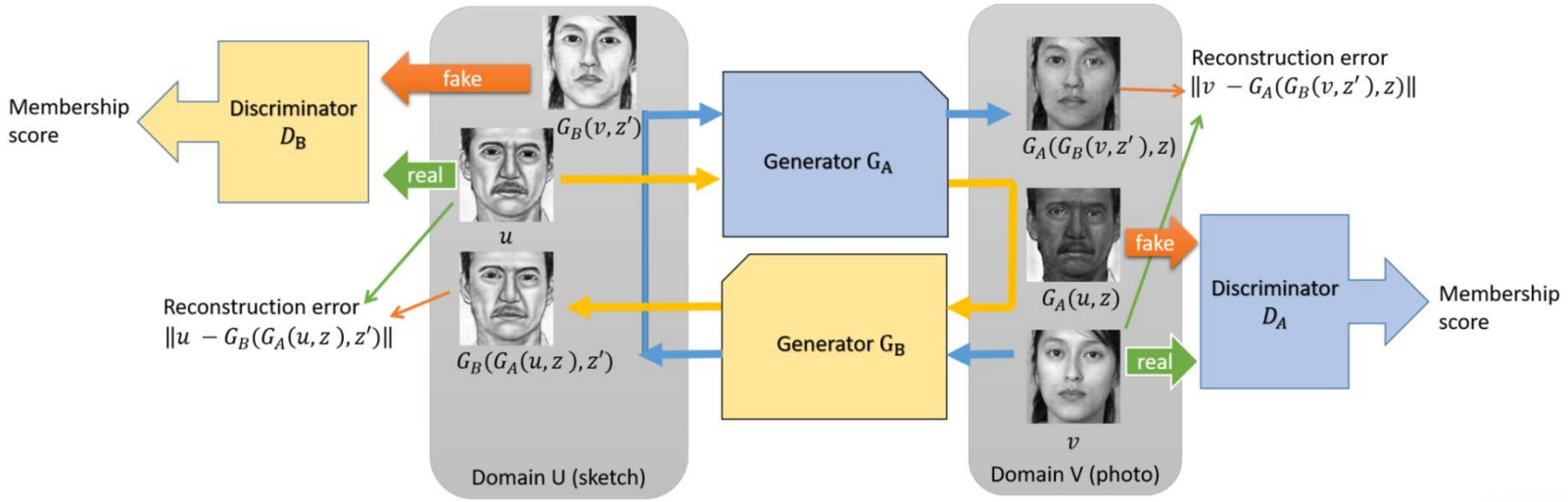
# Our Objective

- **Prime task is to learn a generator $G_A$ : U→V**
  - **Dual task is to train an inverse generator $G_B$ : V→U**




# DualGAN

- **Two GANs:**
  - **Primal GAN learns the generator $G_A$ & a discriminator $D_A$**
  - **Dual GAN learns the generator $G_B$ & a discriminator $D_B$**
- **Reconstruction loss is minimized**




# DualGAN Architecture

- **Loss function for Discriminators**

$$l_A^d(u,v) = D_A(G_A(u,z)) - D_A(v),$$
$$l_B^d(u,v) = D_B(G_B(v,z')) - D_B(u)$$

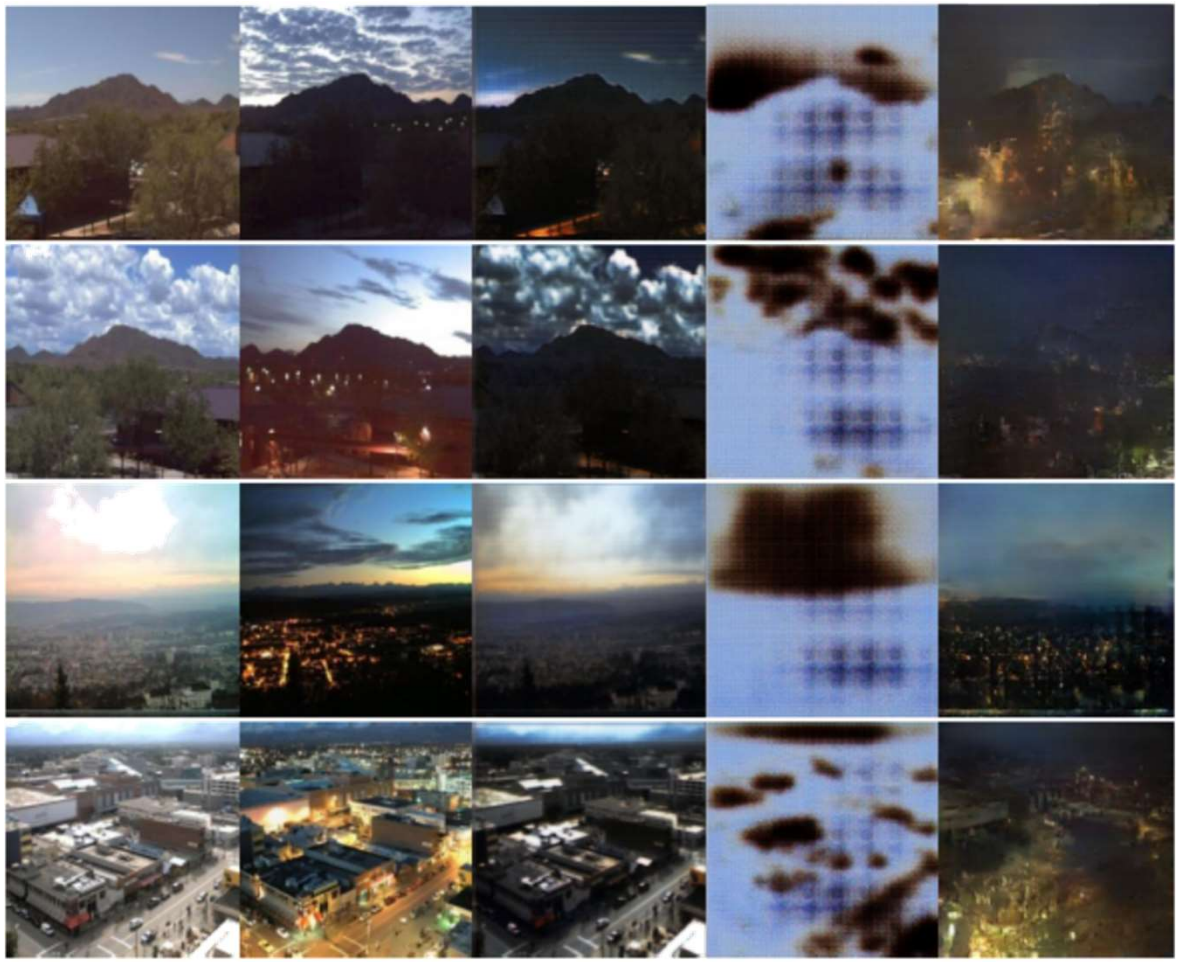
- **Loss function for Generators**

$$l^g(u,v) = \lambda_U \|u - G_B(G_A(u,z),z')\| +$$
$$\lambda_V \|v - G_A(G_B(v,z'),z)\|$$
$$-D_A(G_B(v,z')) - D_B(G_A(u,z)),$$

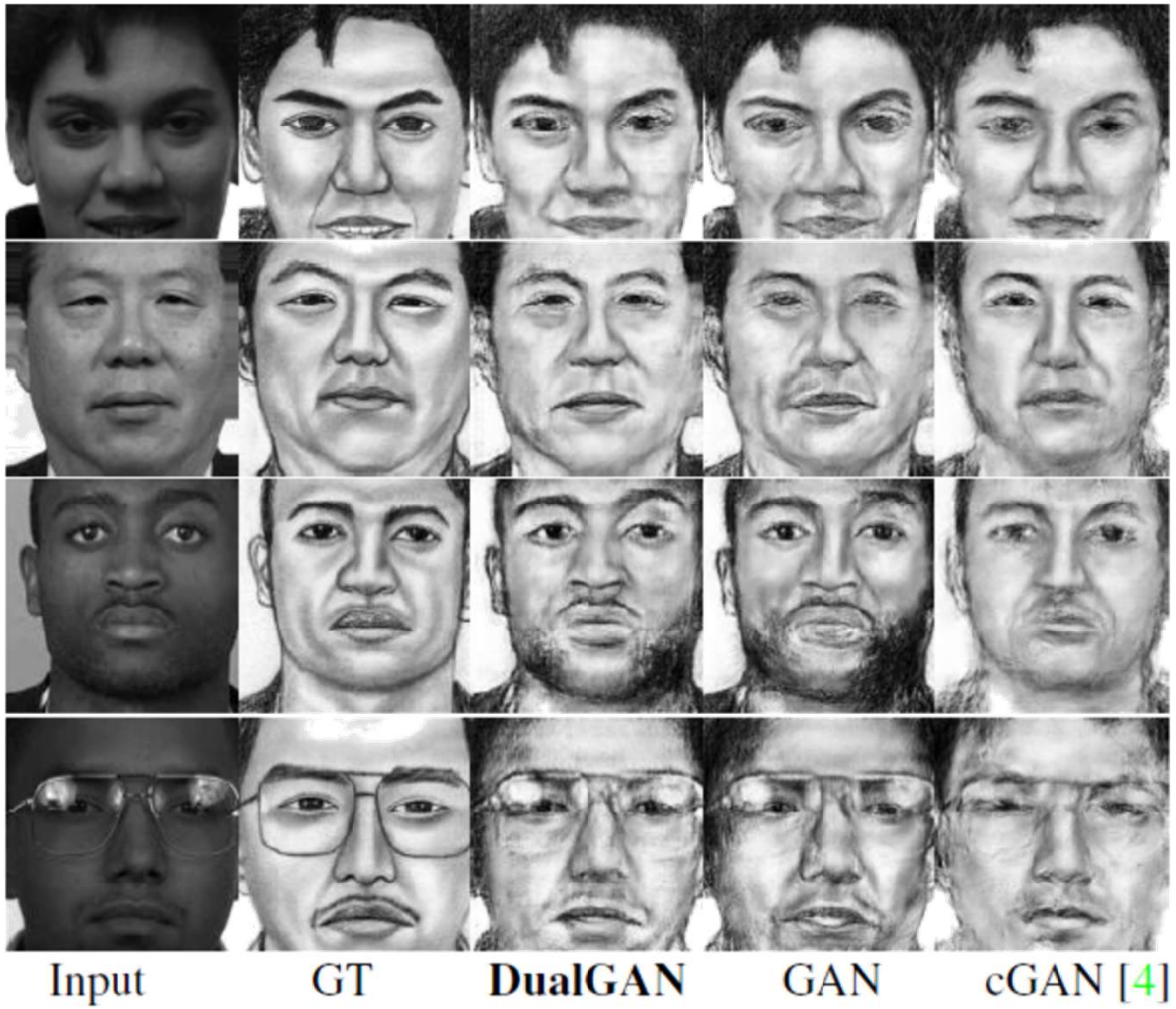
## Loss Functions

| Dataset | Image size | # Training samples | Labeled | Features |
|---|---|---|---|---|
| PHOTO-SKETCH | 128*128 | ~1000 | Yes | Minor misalignment |
| AERIAL-MAPS | 256*256 | ~2000 | Yes | Slight misalignment |
| FACADES-LABEL | 256*256 | ~400 | Yes | Slight misalignment |
| DAY-NIGHT | 512*512 | ~100 | Yes | Slight content difference |
| CHINESE-OIL | 512*512 | ~1000 | No | |
| MATERIAL | 512*512 | ~100 | No | |

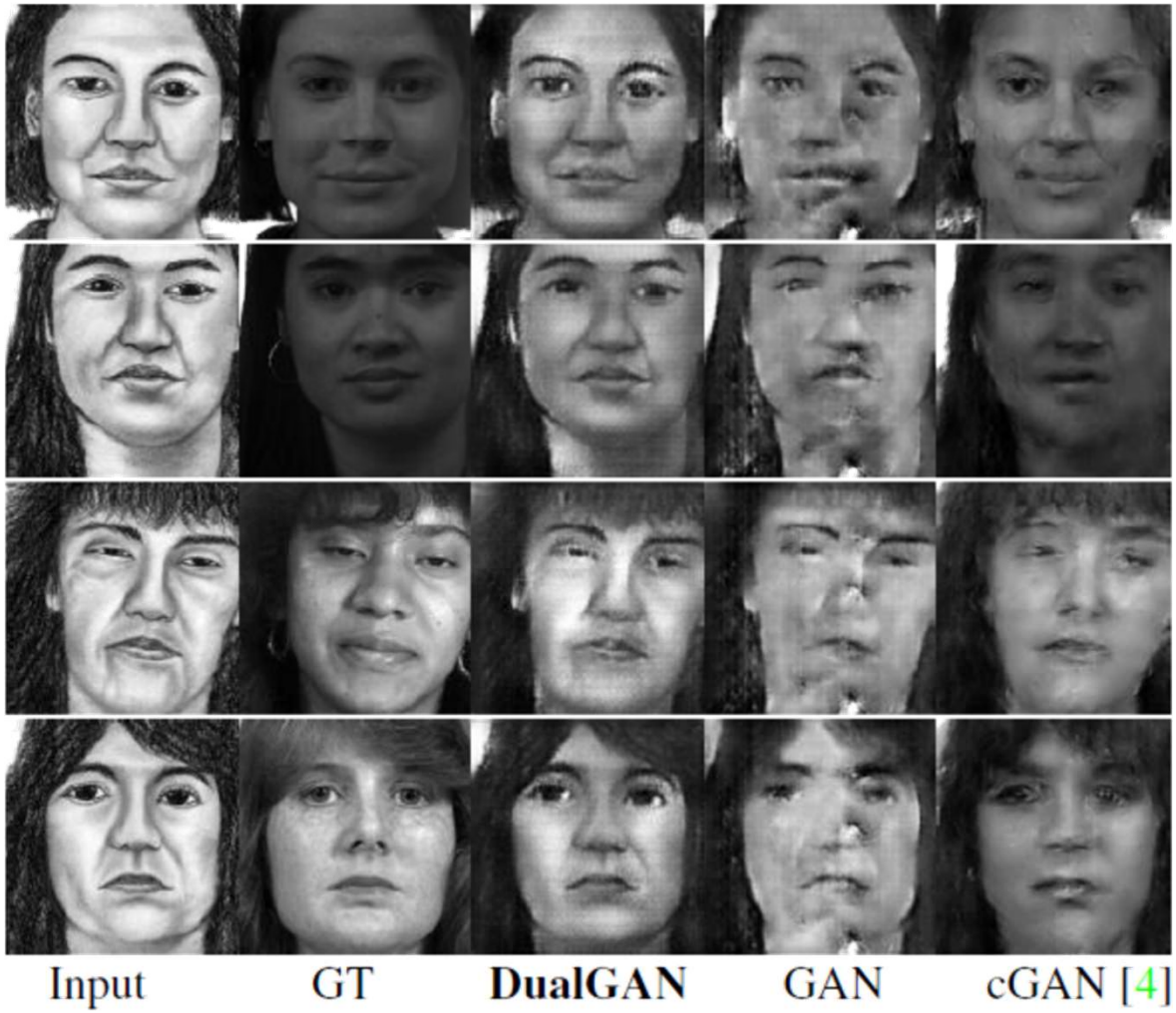
# Experiments



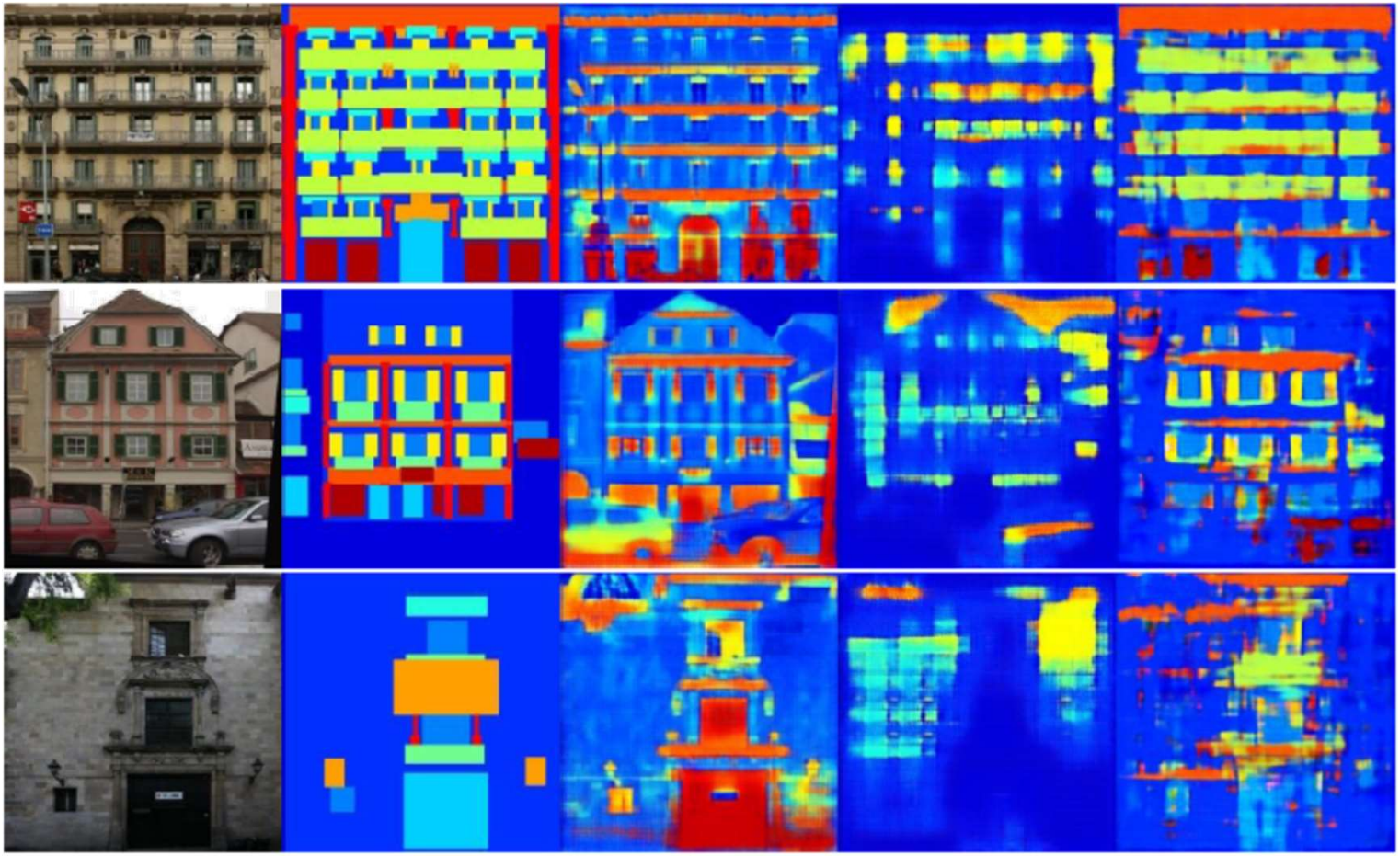

# Day → Night Translation



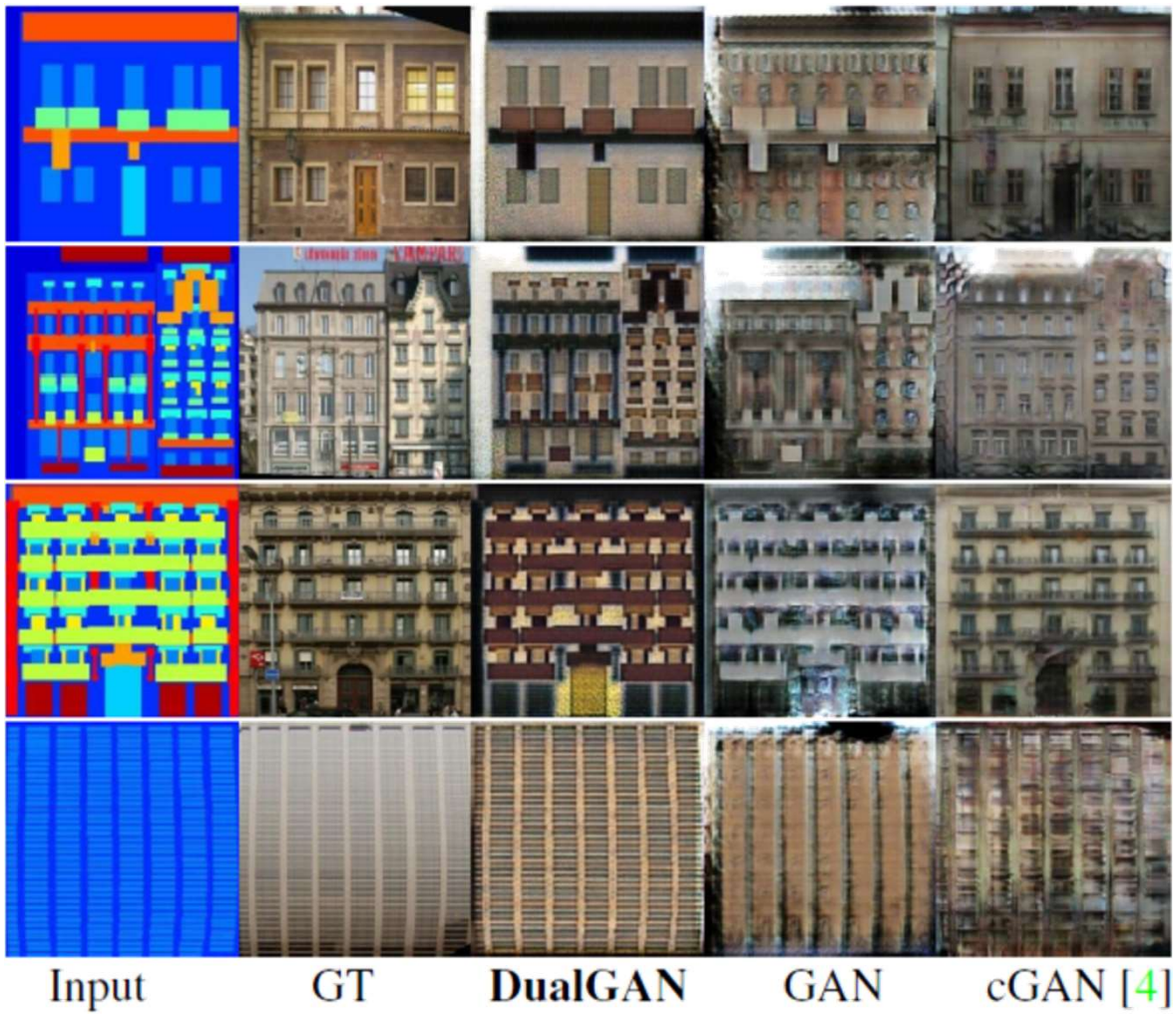

# Photo → Sketch Translation



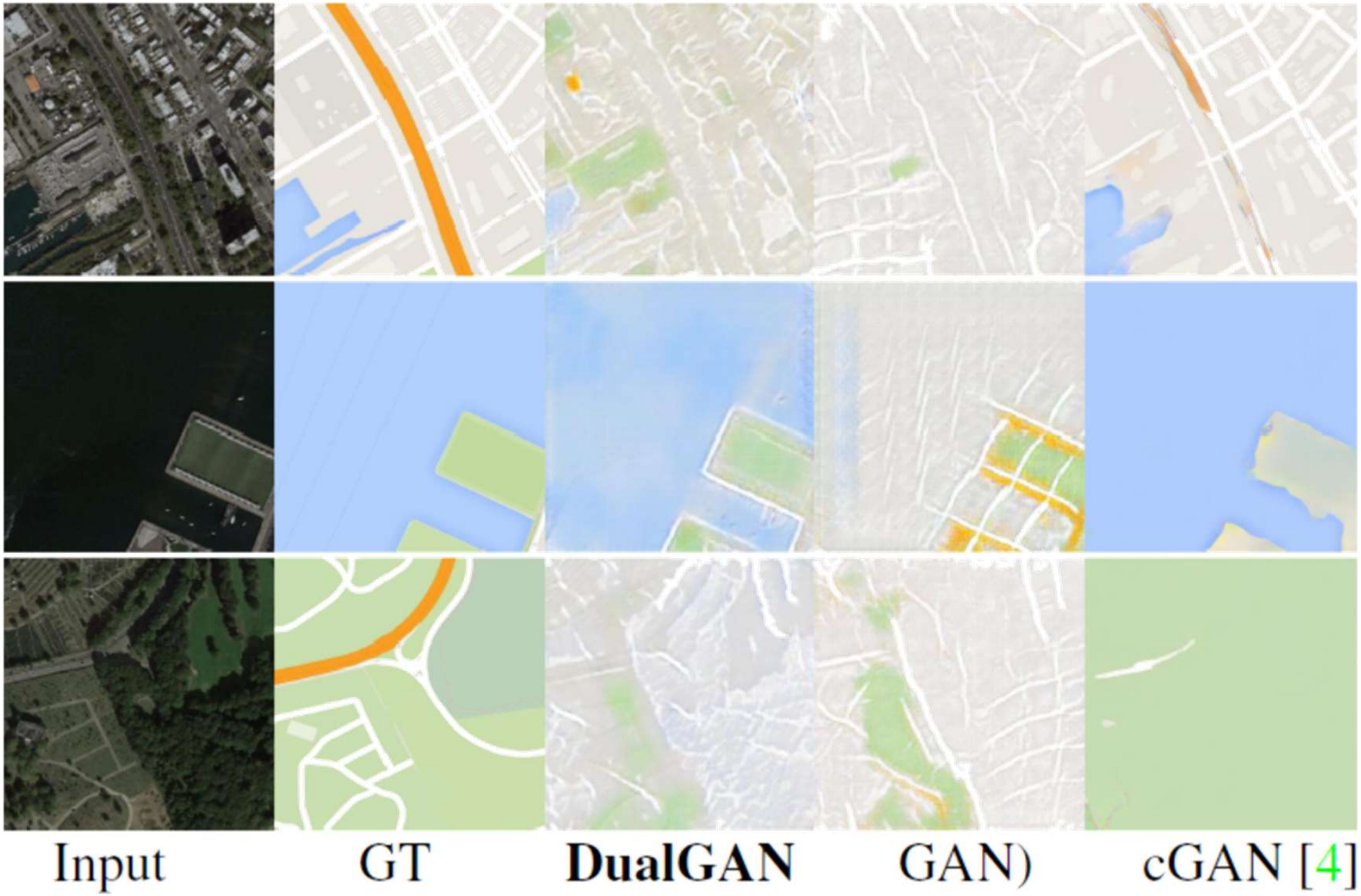

# Sketch → Photo Translation



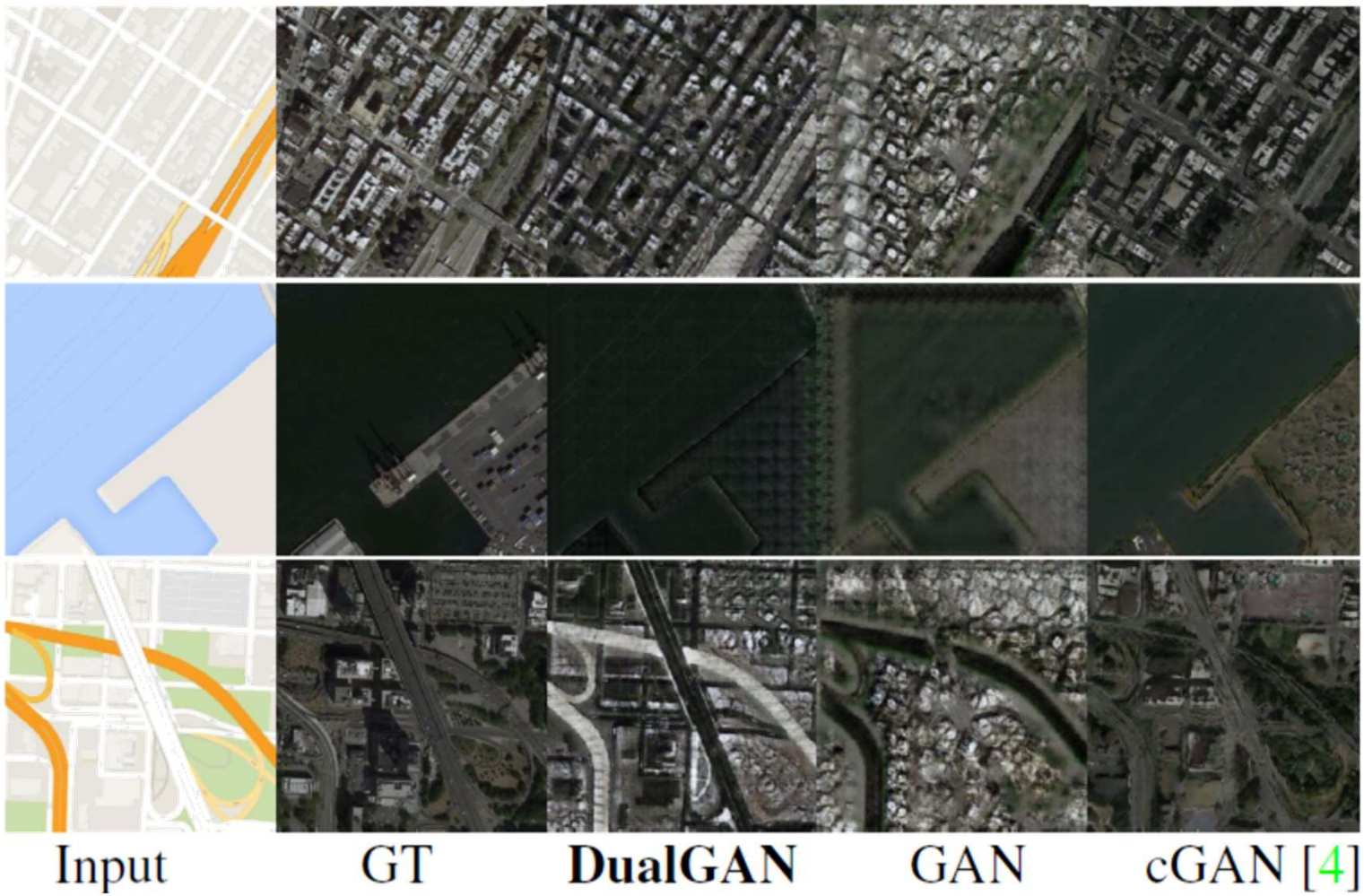

# Facades → Label Translation



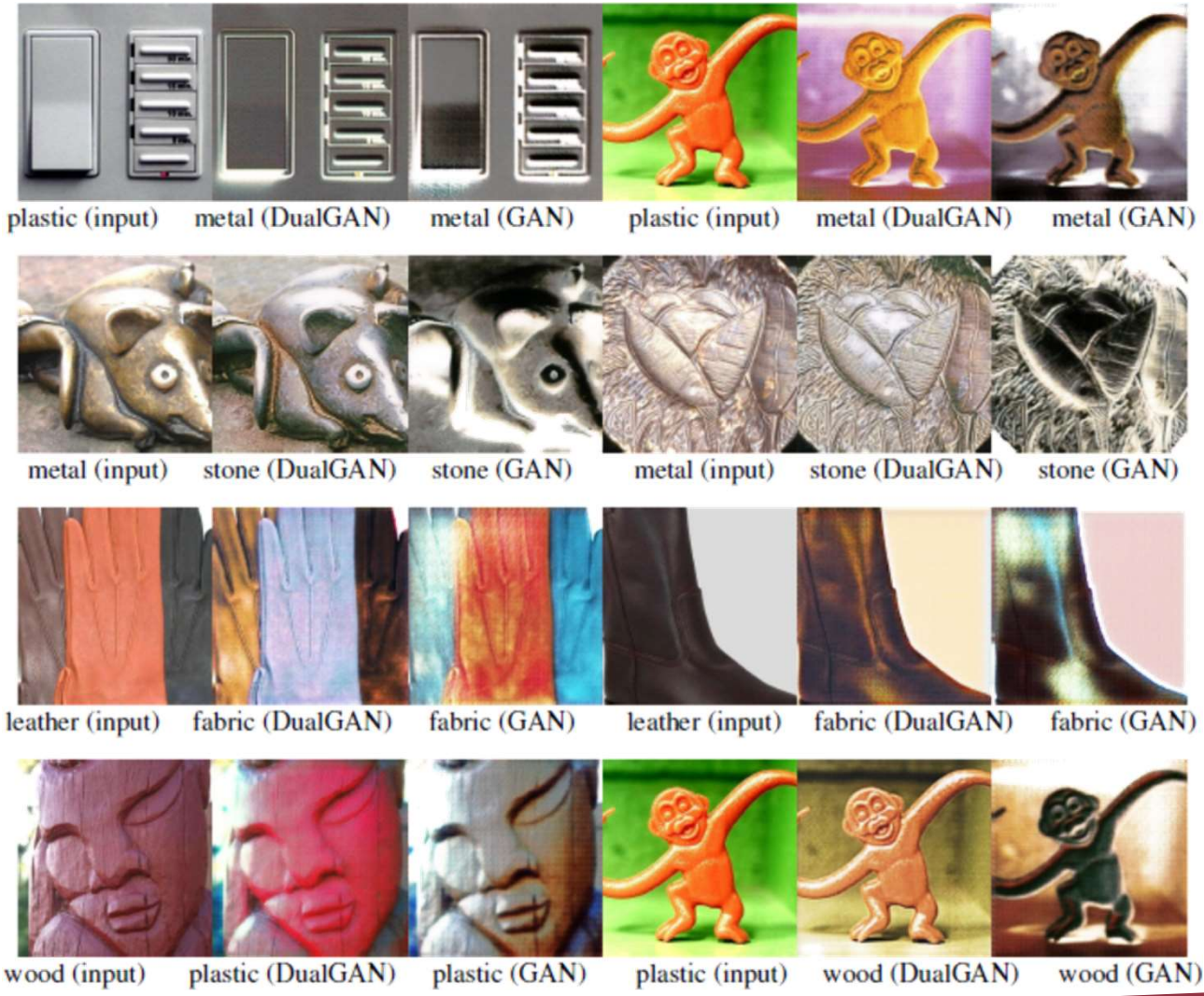

# Label → Facade Translation




# Aerial Photo → Map Translation




# Map → Aerial Photo Translation




# Material Transfer

- **Two user studies were conducted through Amazon Mechanical Turk (AMT)**
  - **AMT is an online platform where a number of Human Turkers are hired to evaluate experimental results or label data**
- **Quantitative evaluation is also performed on image segmentation results**

# Evaluation

- **Run AMT test on 4 translation results:**
  - **sketch→photo, label map→facades, maps→aerial photo, day→night**
- **Randomly shuffle real photos & outputs from GAN, cGAN, and DualGAN**
  - **Each image is shown to 20 Turkers, who score the "realness" of the image**
  - **Score ranges from: 0 (totally missing), 1 (bad), 2 (acceptable), 3 (good), to 4 (compelling)**

## Image Realness Test

| Task | DualGAN | cGAN[4] | GAN | Real Photo |
|---|---|---|---|---|
| sketch→photo | 1.87 | 1.69 | 1.04 | 3.56 |
| day→night | 2.42 | 1.89 | 0.13 | 3.05 |
| label→facades | 1.89 | 2.59 | 1.43 | 3.33 |
| map→aerial | 2.52 | 2.92 | 1.88 | 3.21 |

# Average Realness Scores

- **Evaluates the material transfer results**
  - **Mix the outputs from all material transfer tasks**
    - **A total of 176 output images**
  - **Each image was evaluated by 10 Turkers**
    - **Turkers choose the best match based on which material they believe the objects in the image are made of**
  - **A output image is rated as successful if at least 3 Turkers selected the target material type**

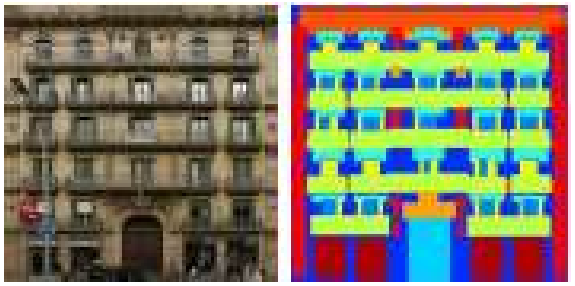
# Material Perceptual Test

| Task | DualGAN | GAN |
|---|---|---|
| plastic→wood | 2/11 | 0/11 |
| wood→plastic | 1/11 | 0/11 |
| metal→stone | 2/11 | 0/11 |
| stone→metal | 2/11 | 0/11 |
| leather→fabric | 3/11 | 2/11 |
| fabric→leather | 2/11 | 1/11 |
| plastic→metal | 7/11 | 3/11 |
| metal→plastic | 1/11 | 0/11 |

# Successful Material Transfer Rates

## Accuracy for the facades→label

| | DualGAN | cGAN | GAN |
|---|---|---|---|
| Per-pixel acc. | 0.27 | 0.54 | 0.22 |
| Per-class acc. | 0.13 | 0.33 | 0.1 |
| Class IOU | 0.06 | 0.19 | 0.05 |



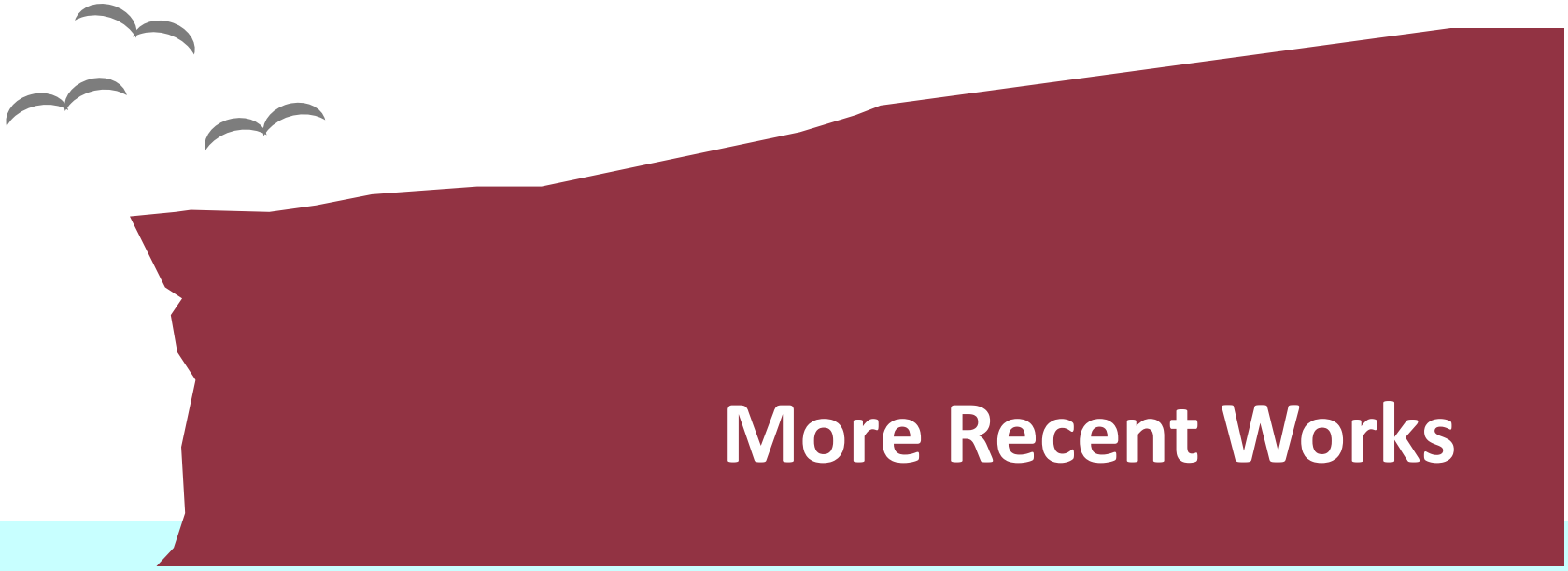
## Accuracy for the aerial→map

| | DualGAN | cGAN | GAN |
|---|---|---|---|
| Per-pixel acc. | 0.42 | 0.7 | 0.41 |
| Per-class acc. | 0.22 | 0.46 | 0.23 |
| Class IOU | 0.09 | 0.26 | 0.09 |

# Segmentation Accuracy

- **A novel unsupervised solution to general-purpose image-to-image translation**
  - **The unsupervised feature enables many more real world applications**
  - **Improves the outputs of vanilla GAN for various image-to-image translation tasks**
  - **With unlabeled data only, DualGAN can generate comparable or even better outputs than conditional GAN that can only be trained with labeled data**

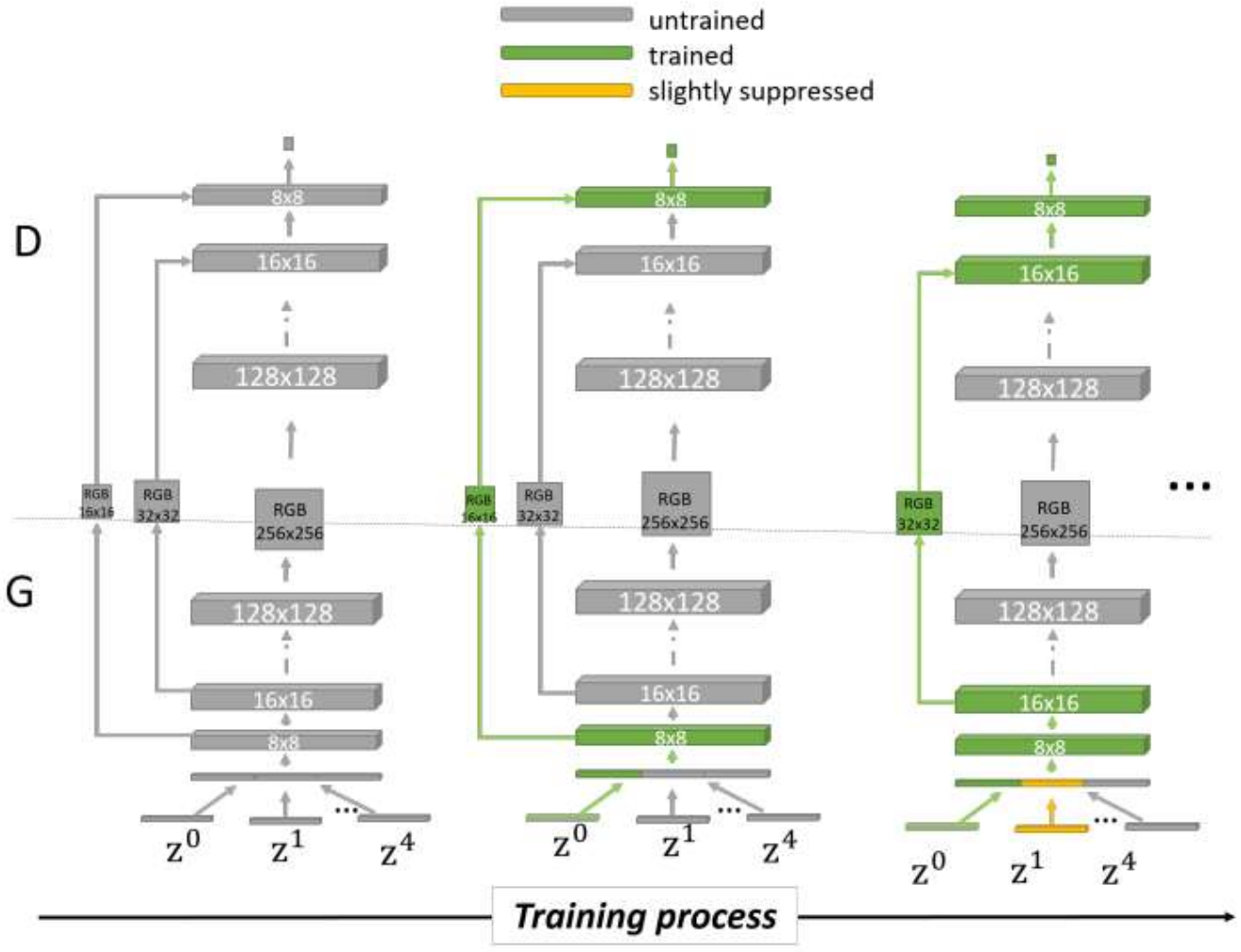
## Conclusions



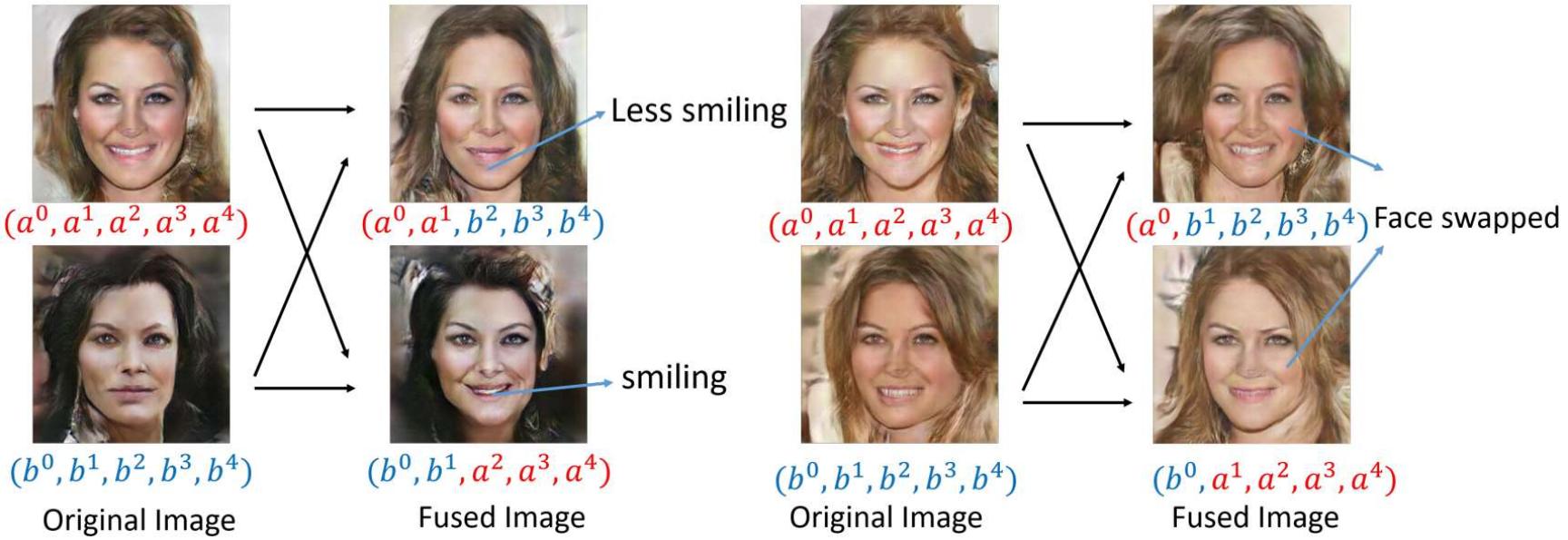
# More Recent Works

- **Objectives:**
  - **Learn scale-disentangled image representations/codings with GAN**
  - **Generate images in more controllable manner**

## Scale-Aware Image Fusion



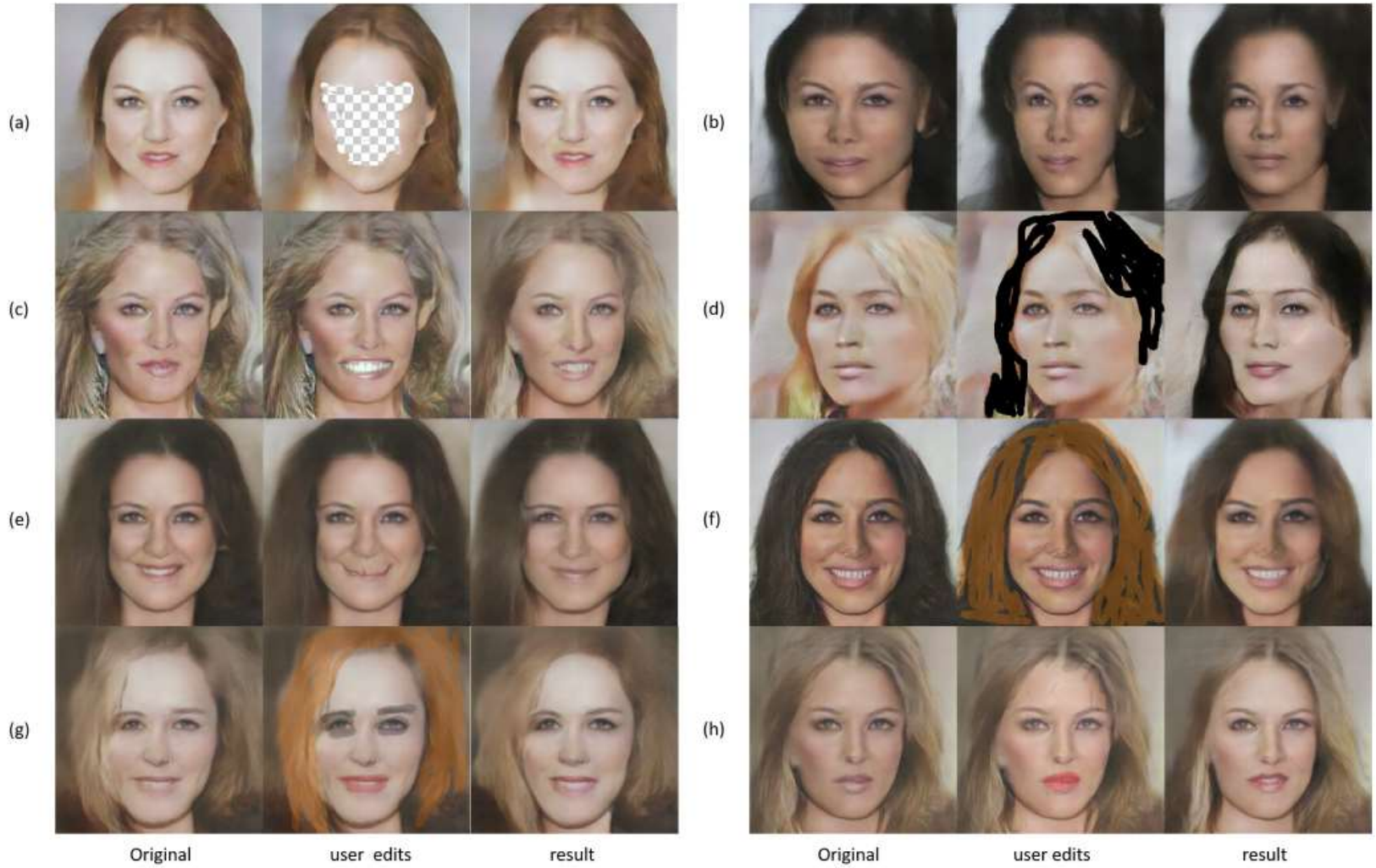

# Proposed Model


$(a^0, a^1, a^2, a^3, a^4)$
$(b^0, b^1, b^2, b^3, b^4)$
Original Image
$(a^0, a^1, b^2, b^3, b^4)$
Less smiling
$(b^0, b^1, a^2, a^3, a^4)$
smiling
Fused Image
$(a^0, a^1, a^2, a^3, a^4)$
$(b^0, b^1, b^2, b^3, b^4)$
Original Image
$(a^0, b^1, b^2, b^3, b^4)$
Face swapped
$(b^0, a^1, a^2, a^3, a^4)$
Fused Image

# Results

- Train a GAN for high-resolution image generation
- Train an encoder that maps an image to codes
- Adopt user inputs from a painting panel
- Optimize the image codes to fit user inputs
  - Gradient descent

## Manipulation of Image Generation


(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
Original
user edits
result
Original
user edits
result

# Results


# Questions?

# Abstract


- Many image processing and computer vision tasks, such as image segmentation, stylization, and abstraction, can be posed as image-to-image translation problems. This talk presents two different image-to-image translation approaches, one is rule-base and the other is learning-based.
- The rule-based algorithm is capable of stylizing an input face photo using a single exemplar image. Since the numbers and varieties of patch samples are highly limited, special cares are put into sample selection to best preserve the identity and content of the input face. A two-phase procedure is also designed, where colors are transferred first in a semantic-aware manner, followed by edge-preserving texture transfer.
- The learning-based algorithm employs Conditional Generative Adversarial Networks (GANs) to perform general cross-domain image-to-image translation. It requires a large set of training images, but unlike existing approaches, the images do not need to be labeled. To train in an unsupervised manner, two GANs are constructed to translate images in opposite directions, forming a closed loop. As a result, images from either domain can be translated to the other and then reconstructed, enabling a reconstruction error term for training.