

Unsupervised Cross-Domain Deep Image Generation

Yaniv Taigman, Adam Polyak, Lior Wolf

Facebook AI Research (FAIR)

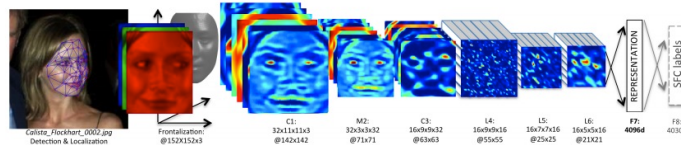
Tel Aviv



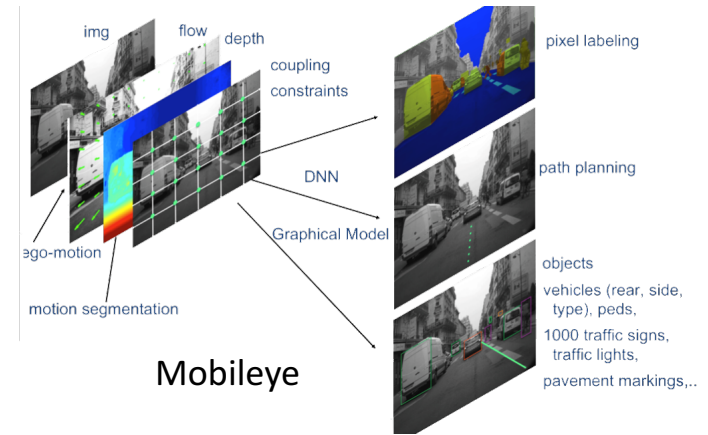
Supervised Learning; $\{X_i, y_i\} \rightarrow F$



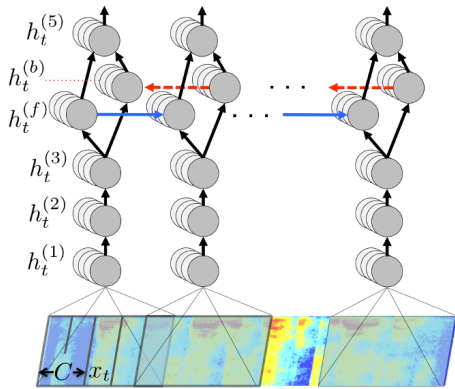
Kaiming et al. (MASK R-CNN / FAIR)



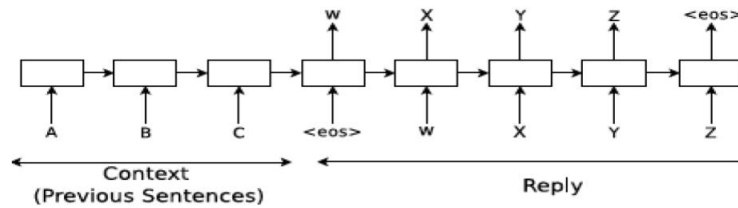
Face Recognition (DeepFace / FAIR)



Mobileye



Speech Recognition
(IBM, Google, MSFT & Baidu)



Seq2Seq for Machine Translation, Quick-Reply, etc.
(Sutskever et al.)

What's still impossible / limited ?

- Human-Machine communication, dialog systems ('Chatbots')
 - Answering free-form questions by memorizing the Web
 - Teaching robots to learn everything.
 - ...
-
- Instead of requiring millions of $\{X, y\}$ samples, maybe **learn to transfer knowledge** between domains, modalities.
 - This talk: Learning to transfer samples between visual domains

Latest Trends



1. Style Transfer (Gatys et al.)

- Replaces statistics/texture given an exemplar

Not semantic

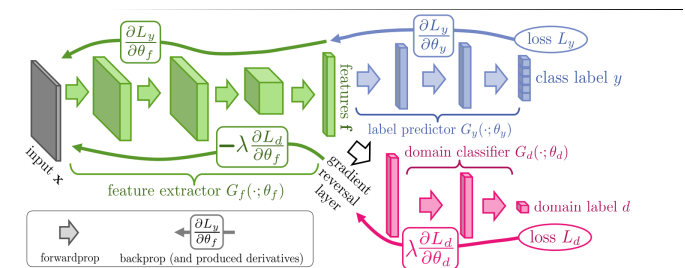


Latest Trends

2. Domain adaptation and Domain Confusion

- “Domain-adversarial training of neural networks” Ganin et al.

Not generative



1. Style Transfer (Gatys et al.)

- Replaces statistics/texture given an exemplar

Not semantic

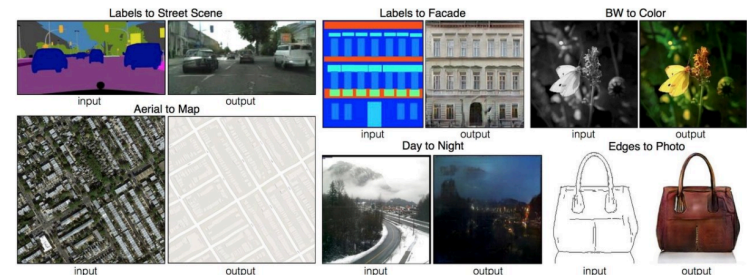
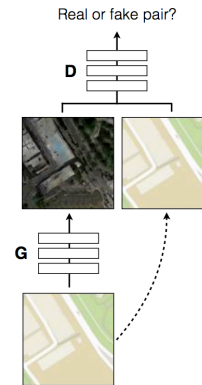


Latest Trends

3. Supervised GANs

- “Image-to-Image Translation with Conditional Adversarial Nets” Isola et al.

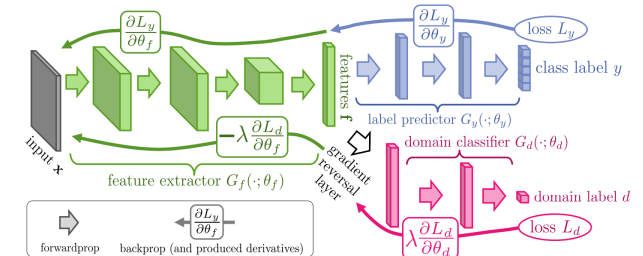
Fully supervised (pairs given)



2. Domain adaptation

- “Domain-adversarial training of neural networks” Ganin et al.

Not generative



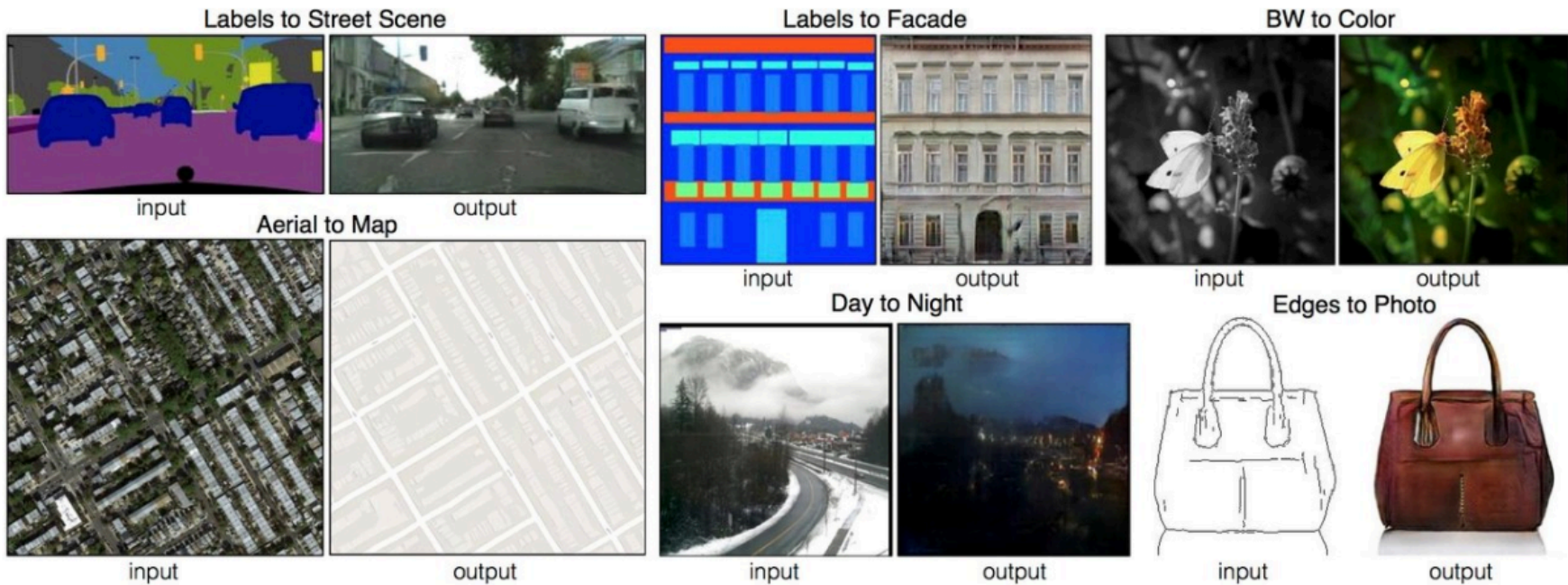
1. Style Transfer (Gatys et al.)

- Replaces statistics/texture given an exemplar

Not semantic



Pix2pix (Isola et al.)



Fully Supervised: Million of ($S_i \rightarrow T_i$) pairs

Vision applications that require **Unsupervised** Image Transfer Methods

Day → Night



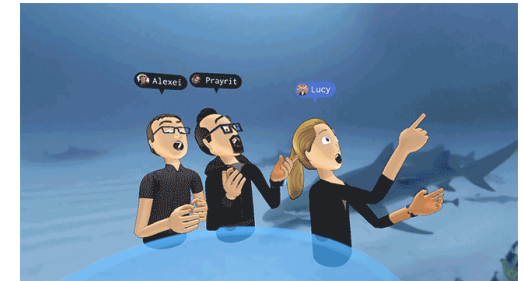
NVIDIA

Bag → Shoe



SKT Brain

World → VR



Oculus

Transfer Learning

AirSim
(Microsoft)



Playing for Data: Ground Truth from Computer Games
(Intel)



Yan Duan et al. OpenAI

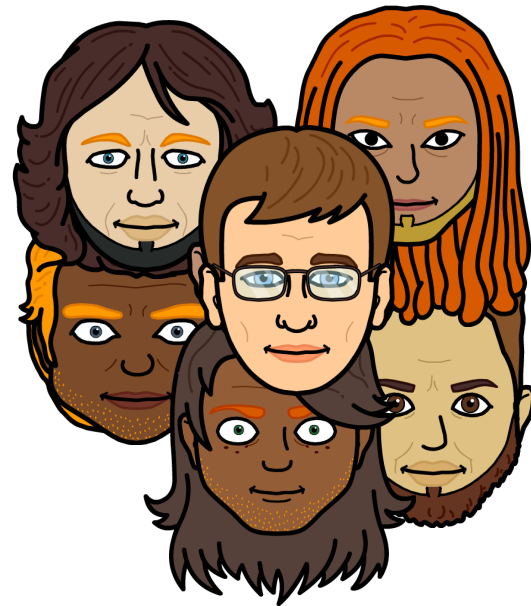


Tzeng et al. Berkley

True AI needs no explicit supervision



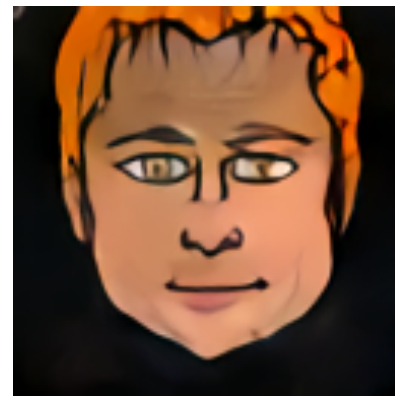
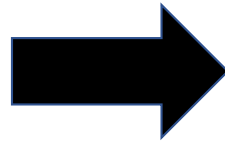
Bag of face images



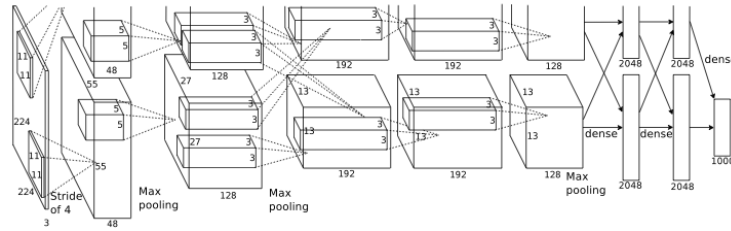
Bag of Emoji

Visual Analogies: Solving the Cartooning task

- Transfer a sample in domain S to its corresponding sample in domain T
- Without any correspondences
- Is this even possible?

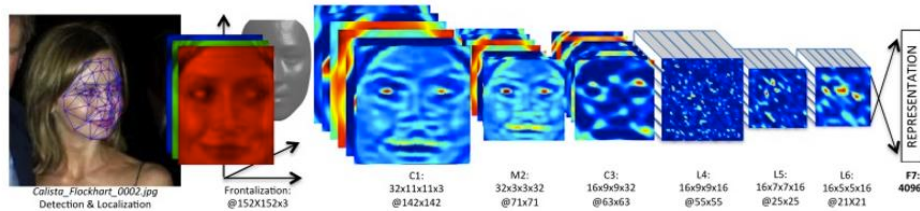


Experts

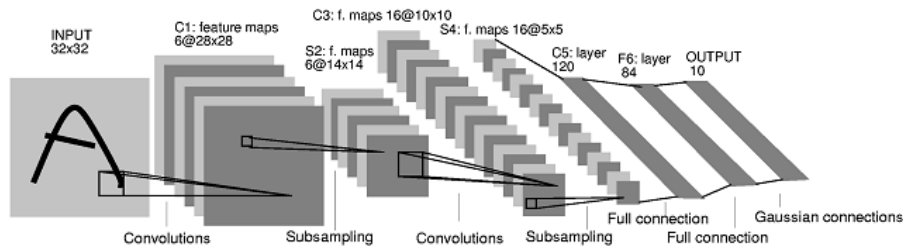


ImageNet
(Krizhevsky et al.)

$f :=$



Faces
(Taigman et al.)



Digits
(LeCun et al.)

Domain Transfer Problem

Given two related domains S, T

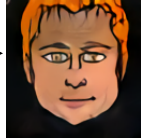
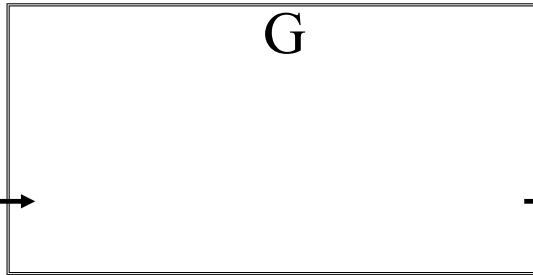
Learn a generative $G: S \rightarrow T$,

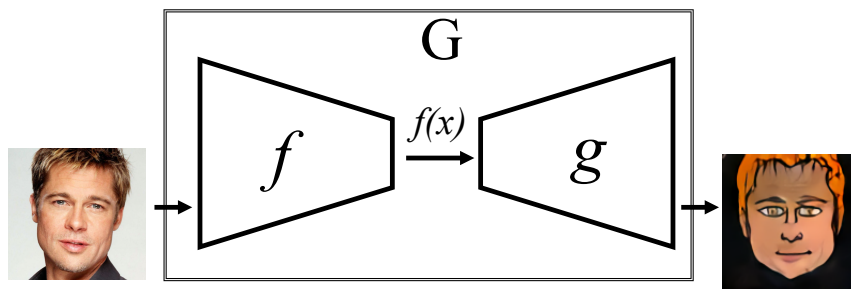
Such that for some $f: S \cup T \rightarrow \mathbb{R}^d$,

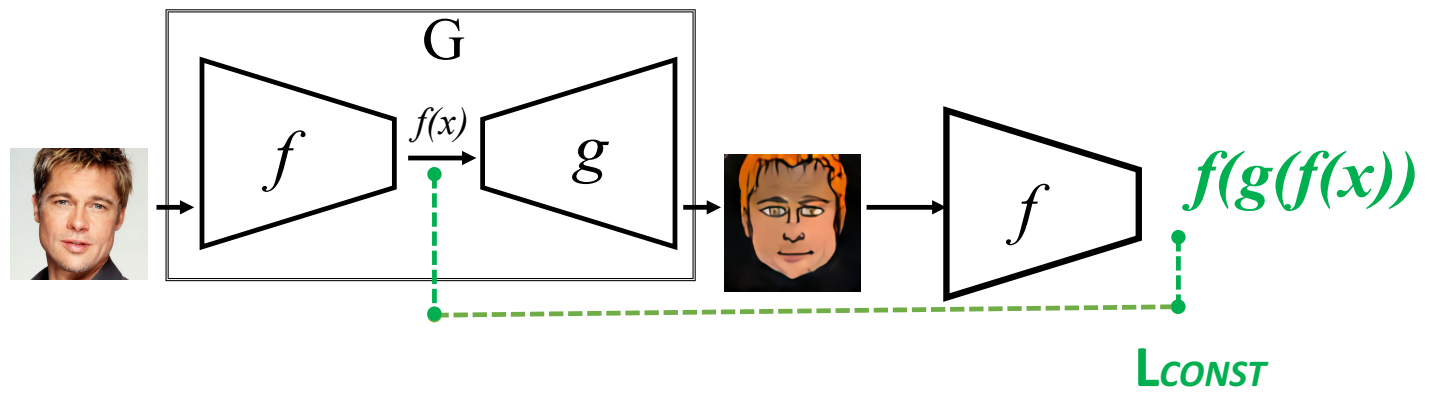
$$f(G(x)) \sim f(x)$$

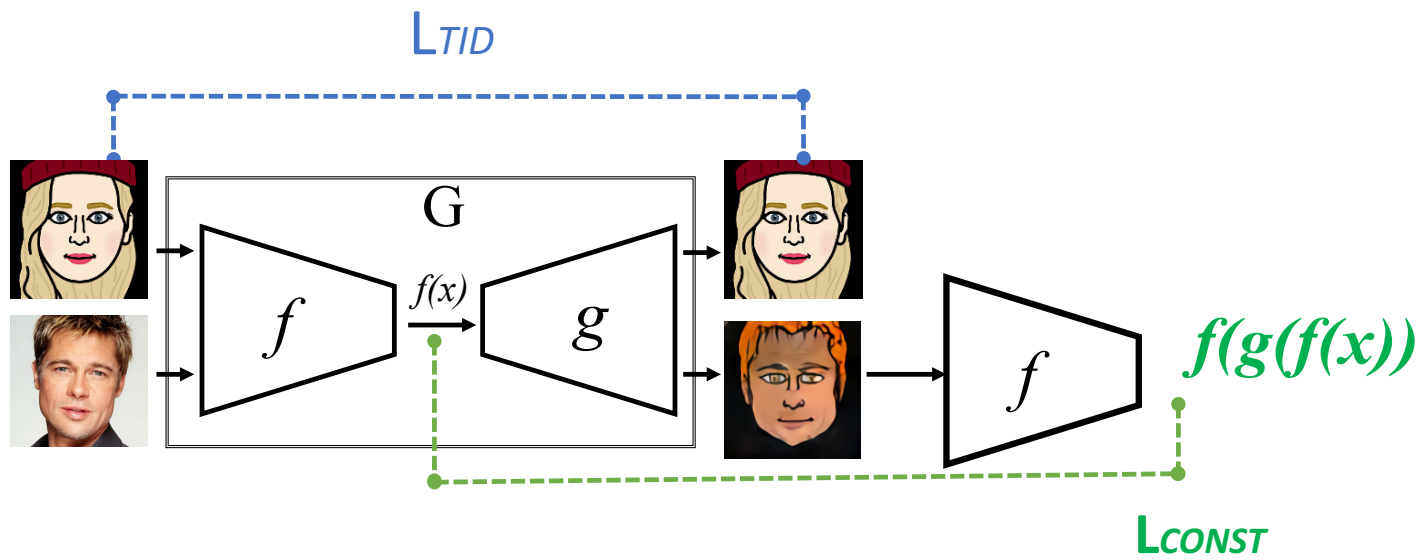
- Unsupervised

- Samples are unlabeled, either in S or T
- No pairs of samples are given, e.g. $\{(s_i, t_i) \mid s_i \rightarrow t_i\}$
- f is asymmetric/unadapted, i.e. was not trained in T







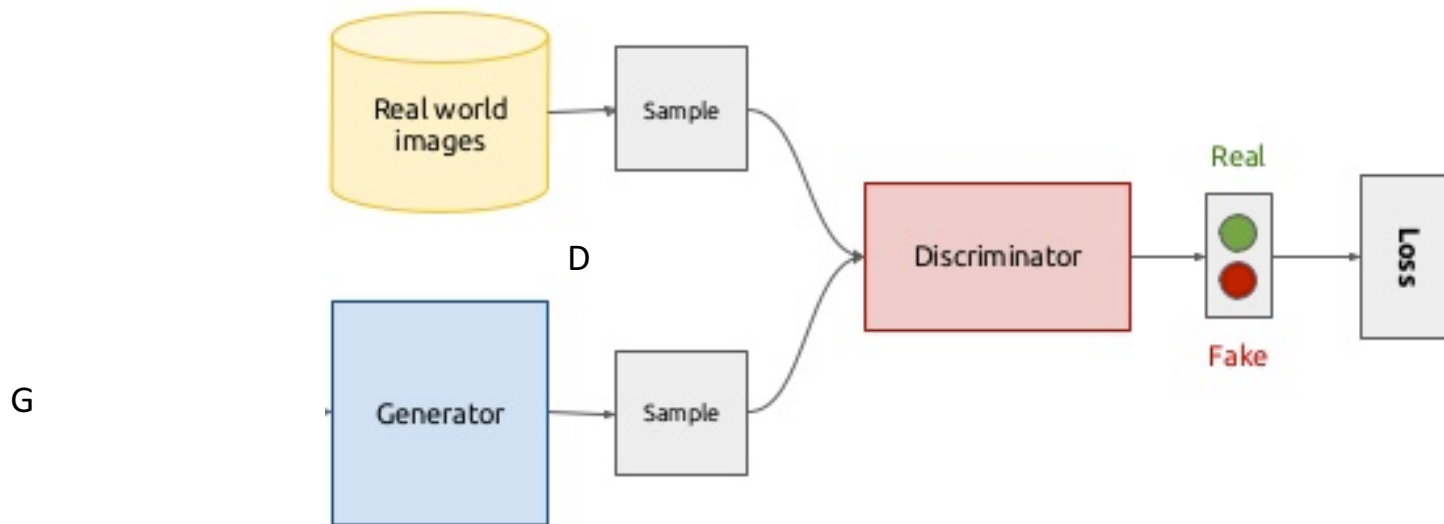


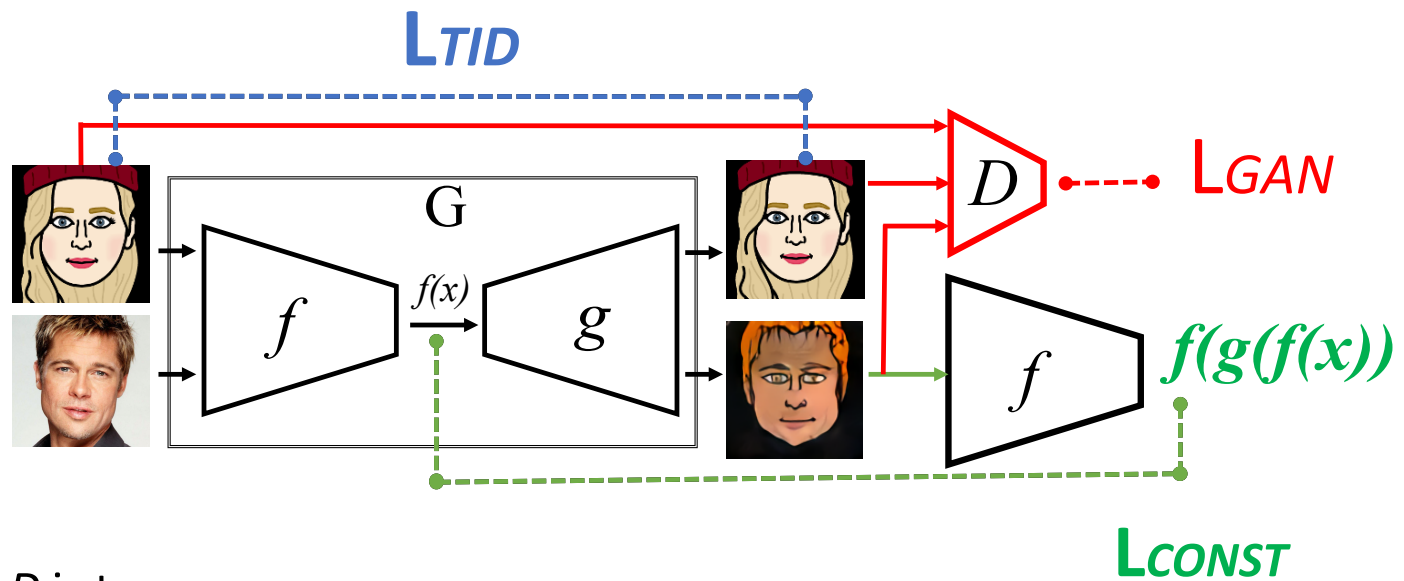
Background: Generative Adversarial Networks (Goodfellow et al.)

Learn networks D, G together.

D identifies which images are real and which created by G

G tries to fool D





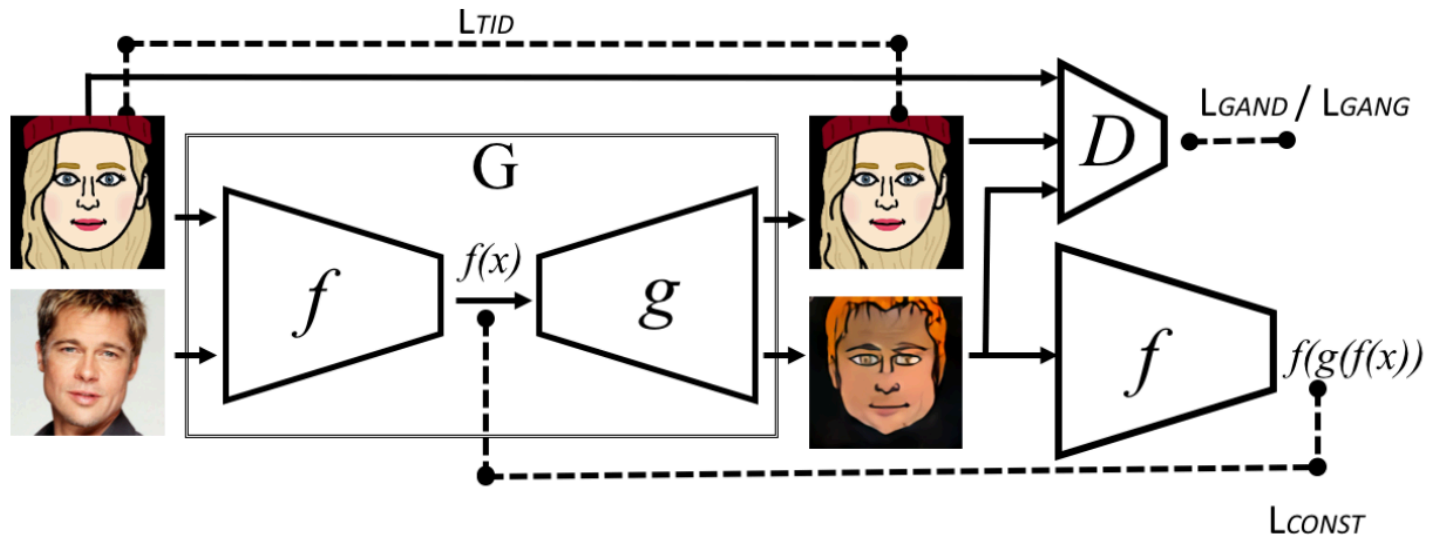
D is ternary:

class 1: authentic emoji

class 2: emoji created by G for an emoji input

class 3: emoji created by G for a photograph

Domain Transfer Network



$$L_{CONST} = \sum_{x \in \mathcal{S}} d(f(x), f(g(f(x))))$$

$$L_{TID} = \sum_{x \in \mathcal{T}} d_2(x, G(x))$$

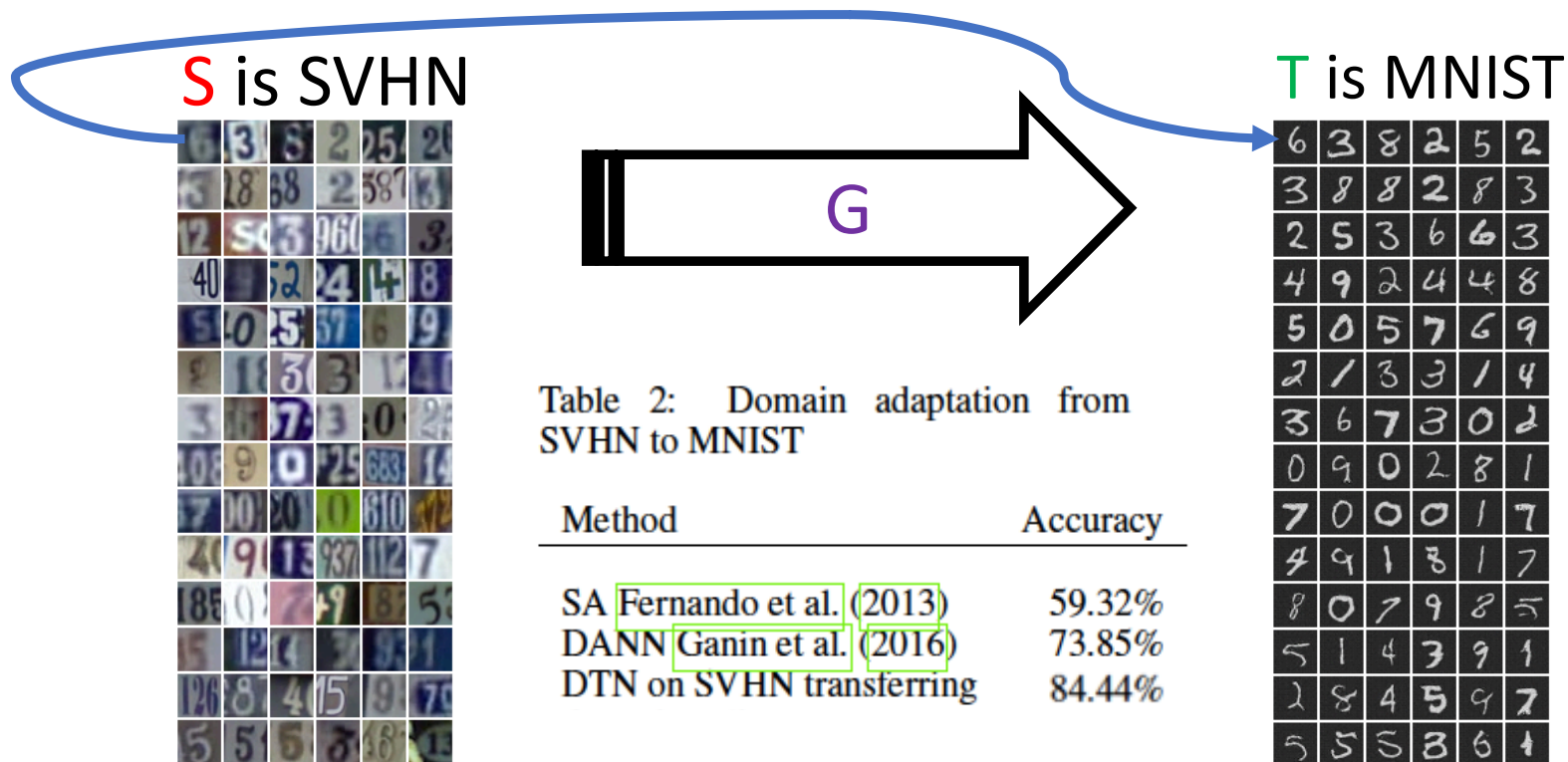
$$L_D = -\mathbb{E}_{x \in \mathcal{S}} \log D_1(g(f(x))) - \mathbb{E}_{x \in \mathcal{T}} \log D_2(g(f(x))) - \mathbb{E}_{x \in \mathcal{T}} \log D_3(x)$$

$$L_{GANG} = -\mathbb{E}_{x \in \mathcal{S}} \log D_3(g(f(x))) - \mathbb{E}_{x \in \mathcal{T}} \log D_3(g(f(x)))$$

Evaluation of the DTN

1. Digits : SVHN → MNIST
2. Faces: Faces → Emoji

$f := \text{ConvNet}(S) \rightarrow \{0, \dots, 9\}$



Zero Shot

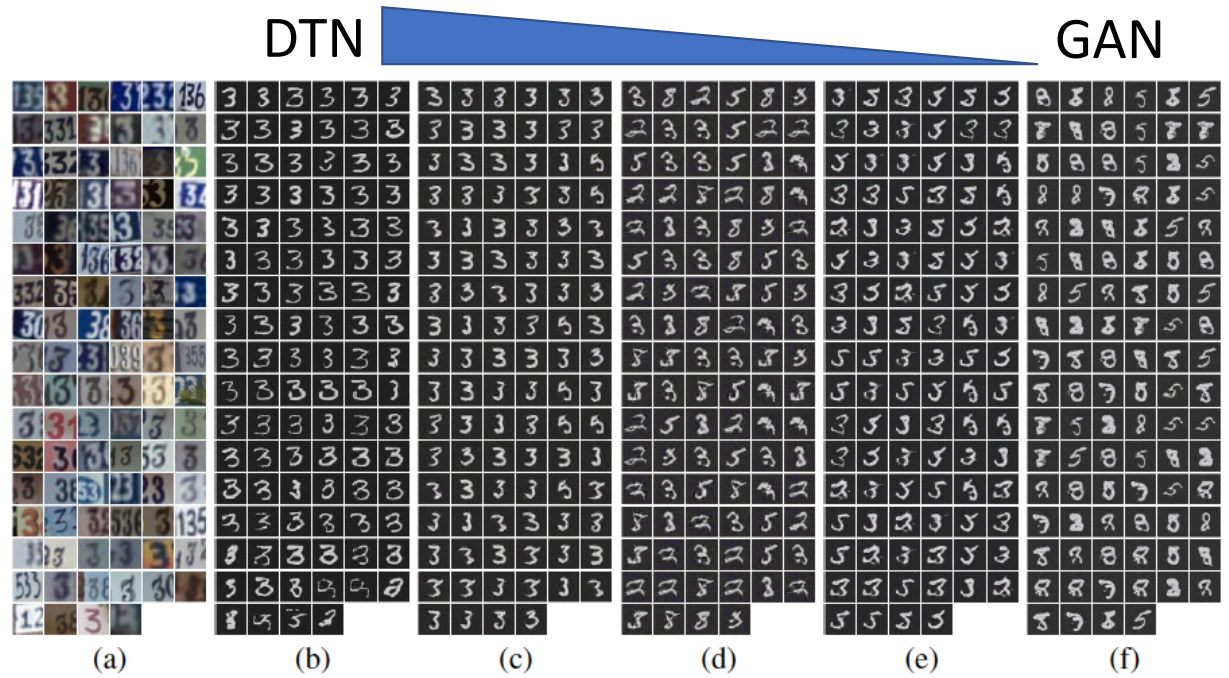
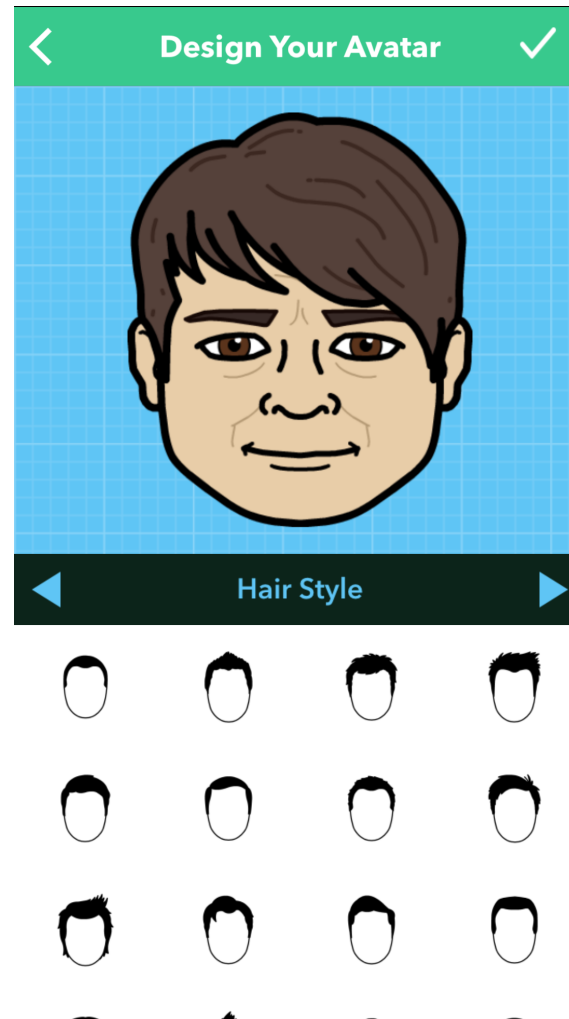
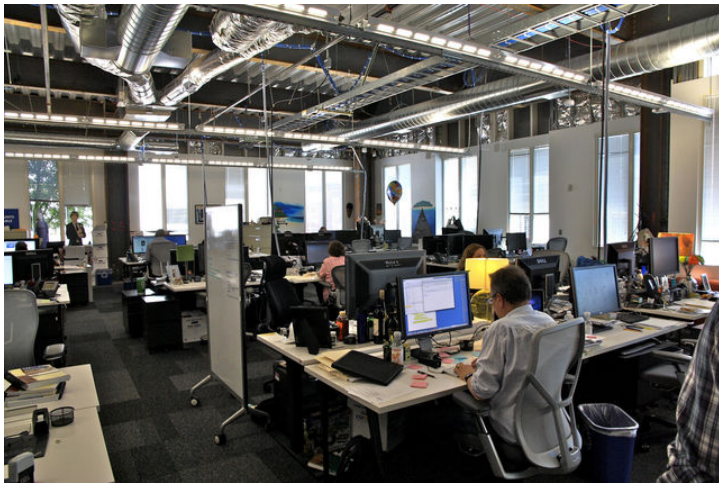


Table 3: Comparison of recognition accuracy of the digit 3 as generated in MNIST

Method	Accuracy of '3'
DTN	94.67%
'3' was not shown in s	93.33%
'3' was not shown in t	40.13%
'3' was not shown in both s or t	60.02%
'3' was not shown in s, t, and during the training of f	4.52%

Faces → Emoji

- ~ 5-10 minutes for an expert
- Limited selection of 'templates'
- Cartooning an open problem



Input **Manual** DTN



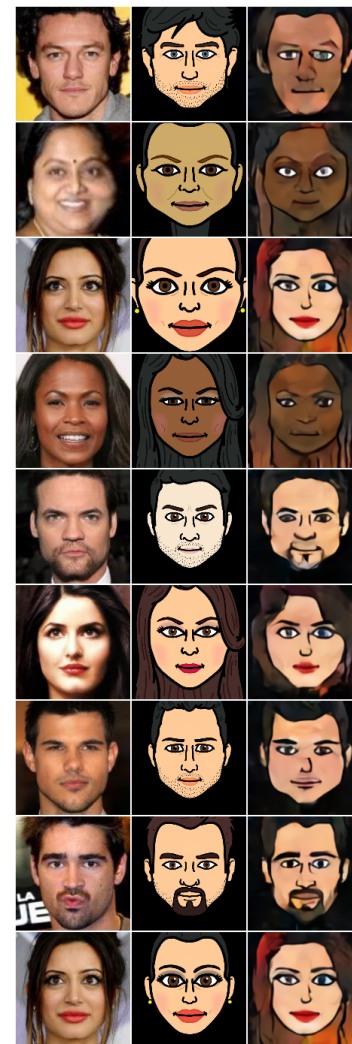
Input **Manual** DTN



Input **Manual** DTN




Input **Manual** DTN



How identifiable are those generations?

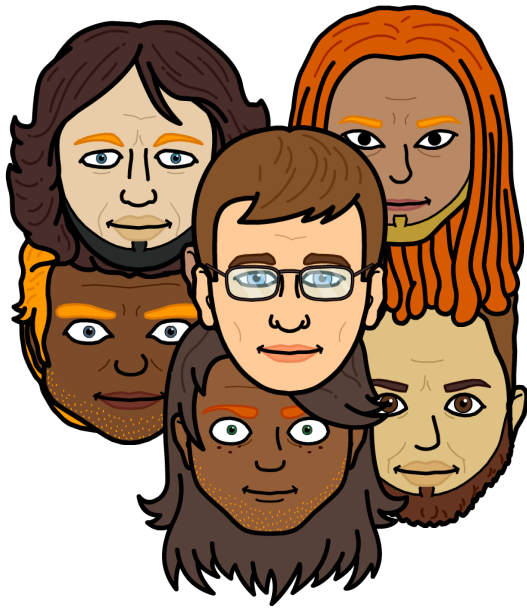
Table 4: Comparison of retrieval accuracy out of a set of 100,001 face images for either manually created emoji or the one created by the DTS network.

Measure	Manual	Emoji by DTN
Median rank	16311	16 
Mean rank	27,992.34	535.47
Rank-1 accuracy	0%	22.88%
Rank-5 accuracy	0%	34.75%

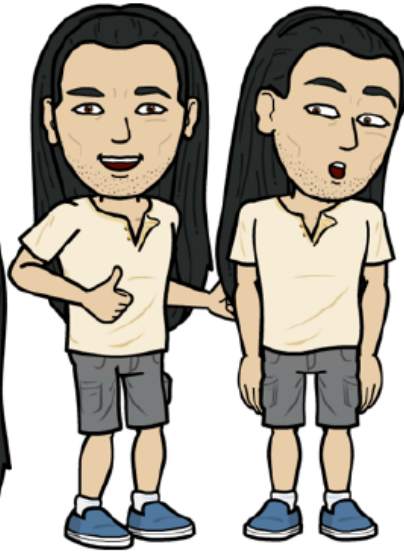


All 80 identities in Facescrub
(No Cherry-Picking)

Are we being totally faithful to the T?



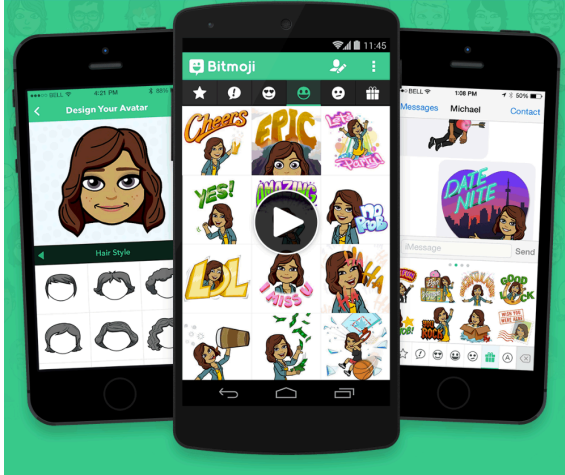
Emoji



DTN

The target domain is generated by an engine

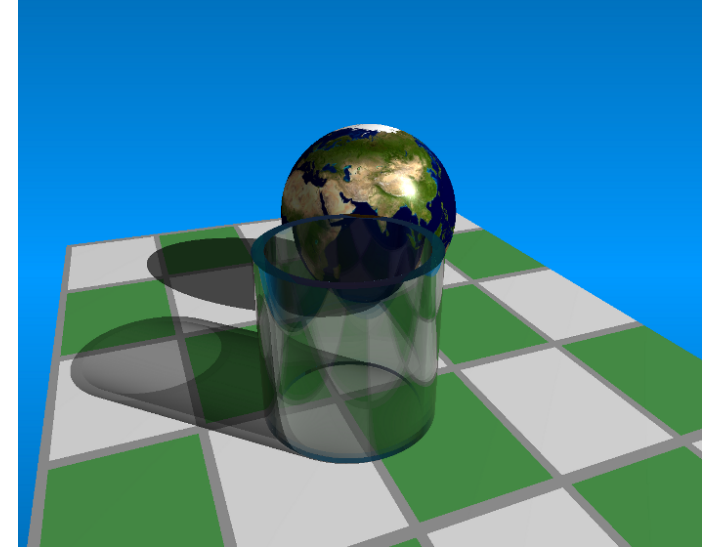
Emoji



3D Avatars



3D Computer Graphics



Engine: configuration \rightarrow image

Domain Transfer Revisited

Given two related domains S, T

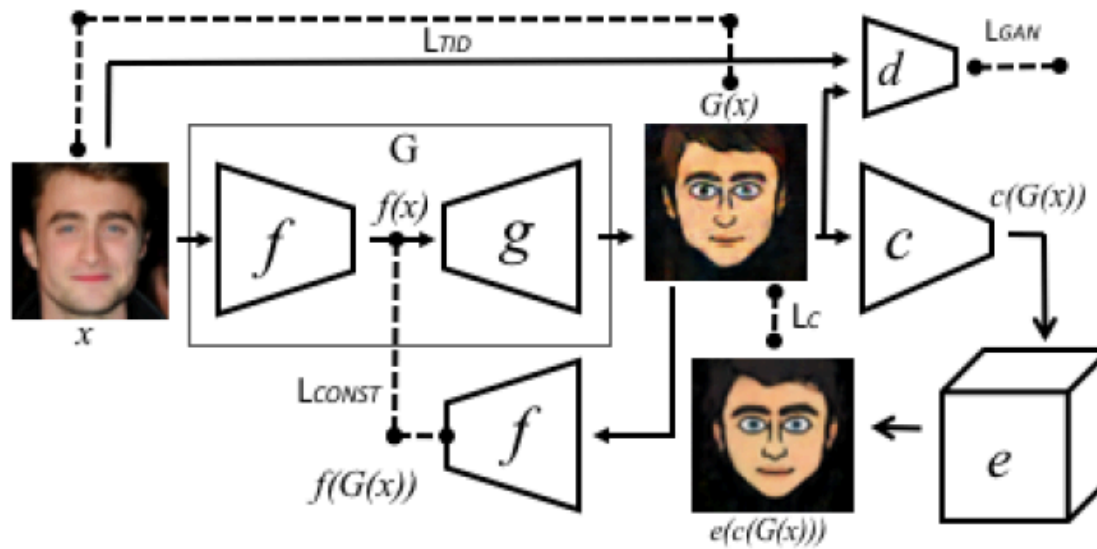
Learn a generative $G: S \rightarrow T$,

Such that for some $f: S \cup T \rightarrow \mathbb{R}^d$,

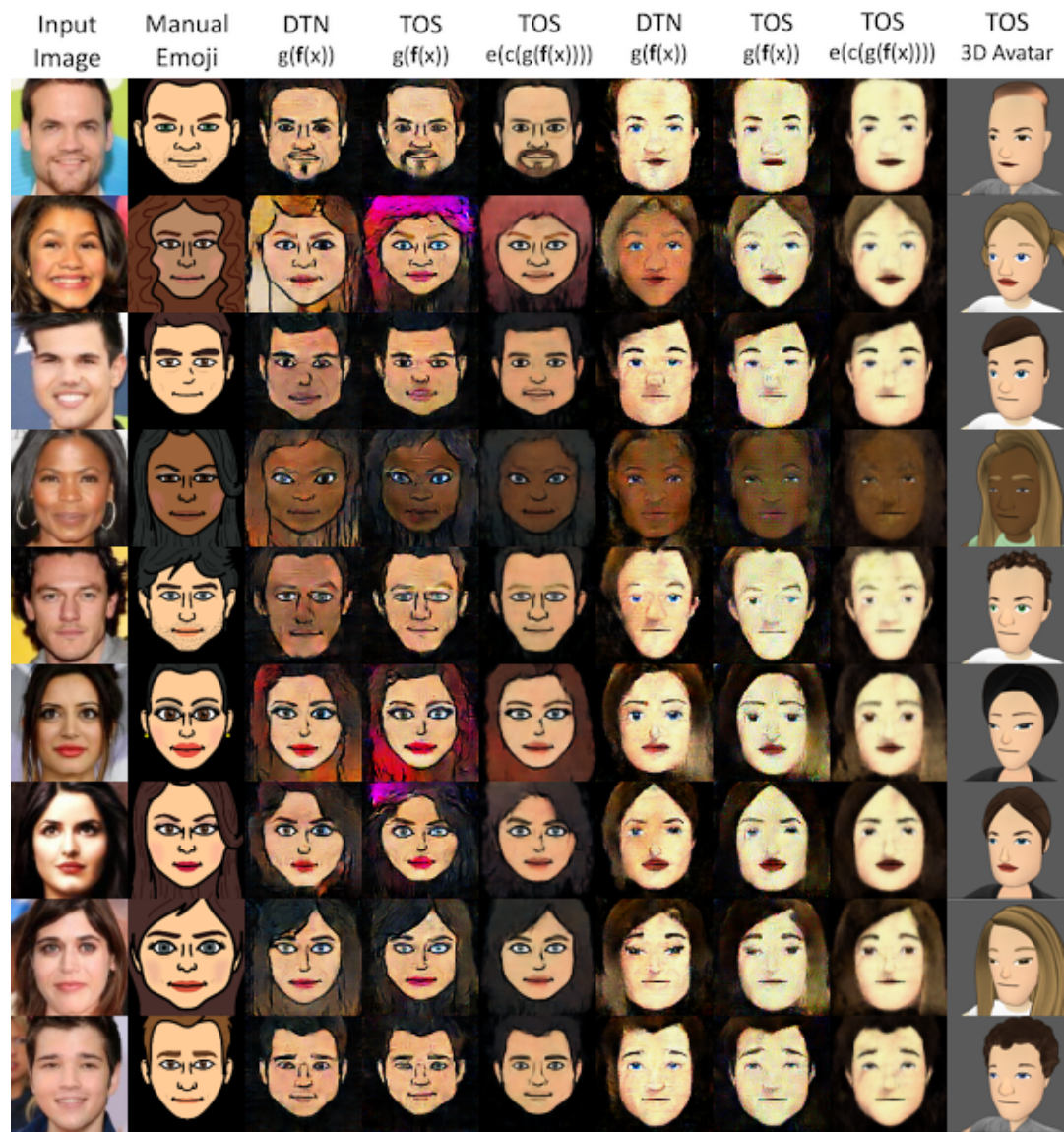
$f(G(x))$ is close to $f(x)$

And for a given engine E , there exist a configuration u
such that $G(x) = E(u)$

Tied Output Synthesis (TOS)

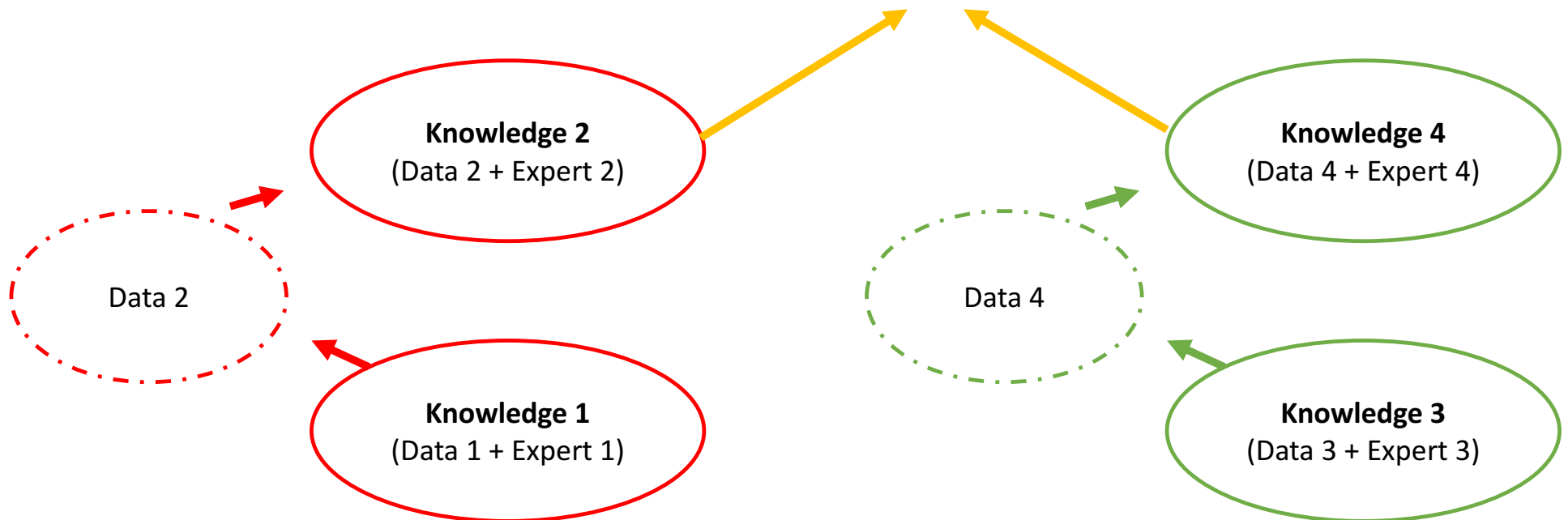


Unsupervised Creation of Parameterized Avatars; In submission



What's ahead?

- Unsupervised (Transfer) Learning in new domains



Thank you

- {yaniv, adampolyak, wolf}@fb.com



References:

Unsupervised Cross-Domain Image Generation; Taigman, Polyak, Wolf [ICLR 2017]

Unsupervised Creation of Parameterized Avatars; Wolf, Taigman, Polyak [In submission]