FIT5230 Malicious AI

Deepfakes I

Overview

- Refocus on Security Properties

- AI attacks Security
  - CONF
  - INT / AUTH

- Deepfakes
  - 1st order Motion Model
  - Motion-supervised Co-part Segmentation

Deepfakes

AI attacks Security

Security Properties

- CONF
- INT
- AUTH

Security: CONF C

- CONFidentiality: secret not leaked
  - cannot prevent access/intercept/compromise/leakage

in storage in transit

  - prevent understanding/comprehension of secret:
    - transform secret m to incomprehensible form c
    - cryptography: encrypt/encipher

Security Properties

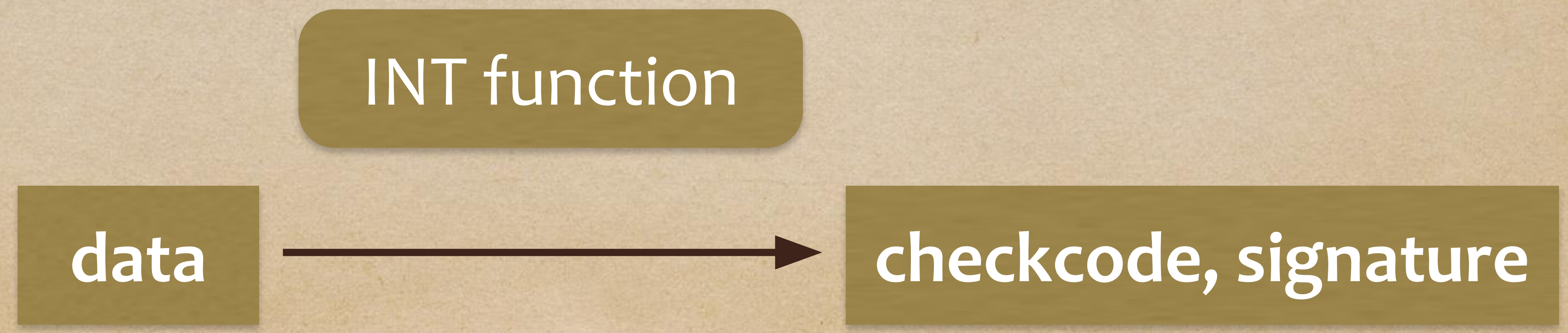
- CONF
  - encryption
- INT
- AUTH

Security: INT

I

- INTegrity: data not changed, originally from source
  - cannot prevent modification of data m into some z

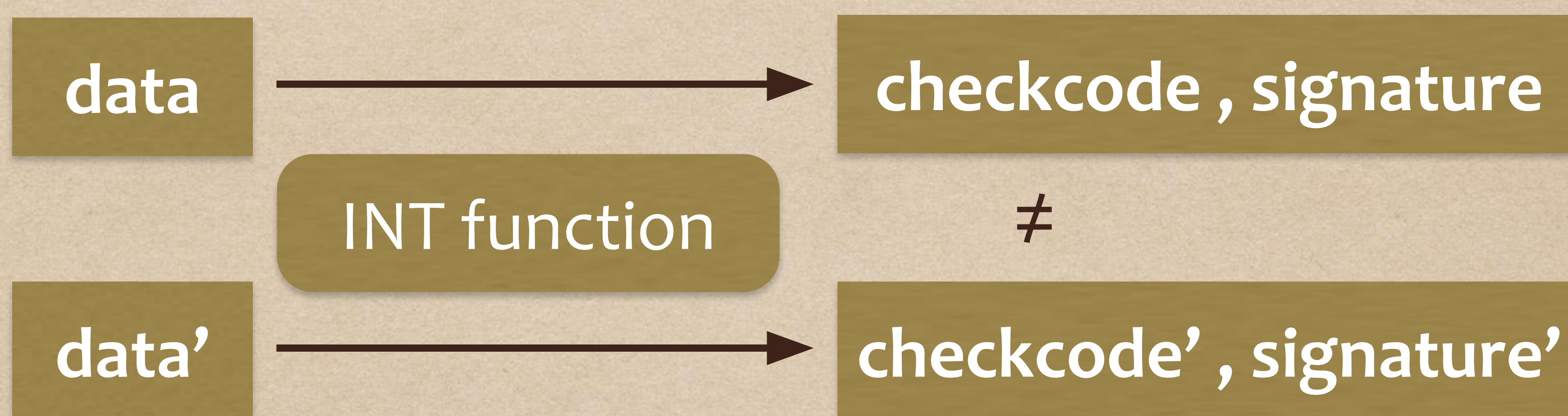
- prevent undetected modifications
  - check if same, using metadata (like checksum)
  - cryptography: message authentication code, signature



Security: INT

I

- INTegrity: data not changed, originally from source
  - check if same, using metadata (like checksum)



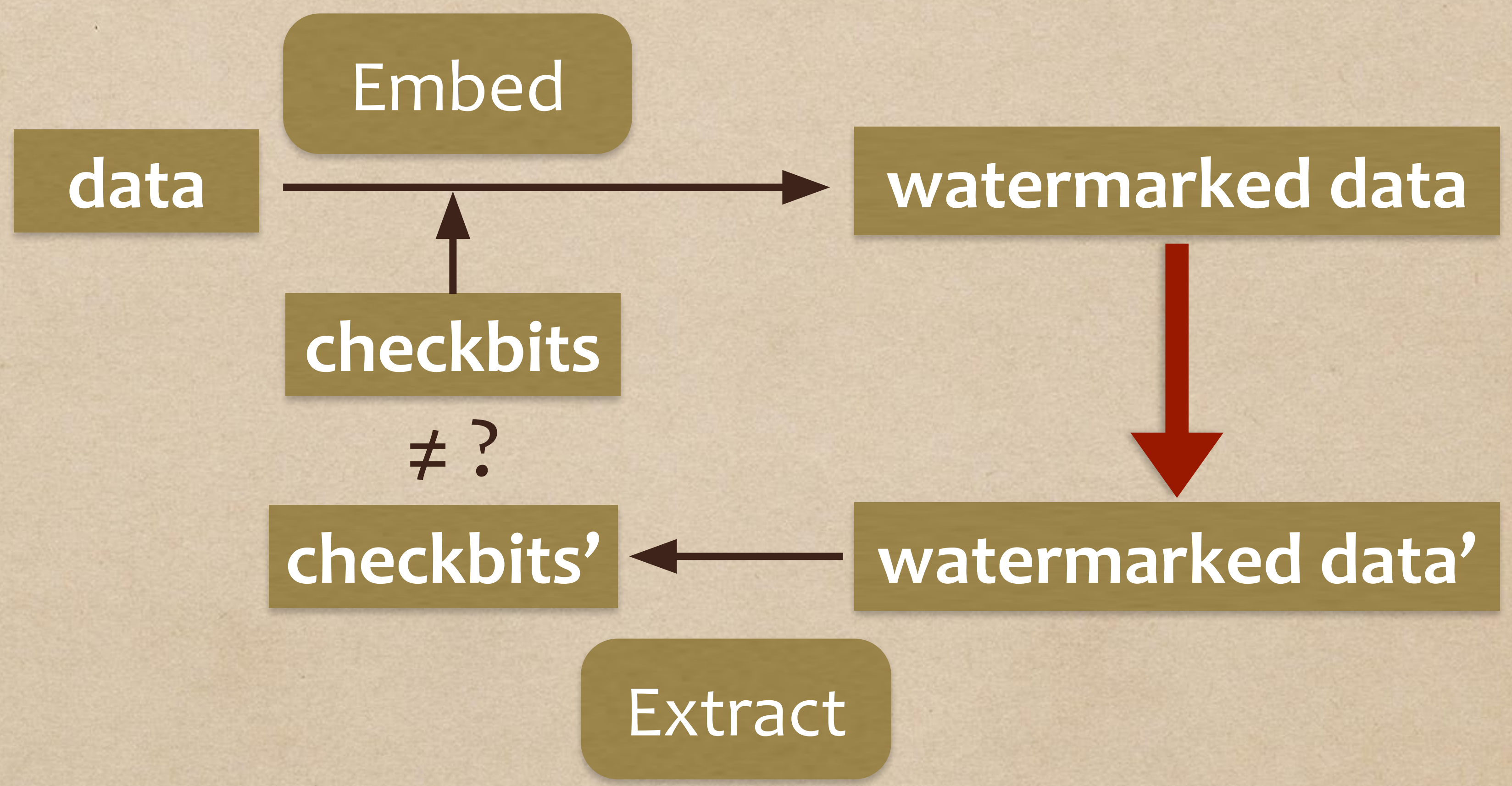
Security Properties

- CONF
- INT
  - crypto: MAC, digital signature
- AUTH

Security: INT

I

- INTegrity: data not changed, originally from source
  - pre-embed checkbits into data



  - if data changed, checkbits should be different
  - information hiding: fragile watermarking

Security Properties

- CONF

- INT
    - crypto: MAC, digital signature
    - signal/image/video processing: watermarking

- AUTH

Security: AUTH

A

- AUTHentication: source/origin is correct
  - check uniqueness of AUTH factor
    - what only you know: passwords, PINs, IC no., …
    - what only you have: passport, ATM card, …
    - what only you are: biometrics
      - static: e.g. facial, fingerprint, …
      - dynamic/soft: e.g. gait, gesture, keystrokes, …
    - who you know: mutual friends, …

Security Properties

- CONF
- INT
- AUTH
  - does not matter what technique you use, need to start with something unique

AI attacks Security

Adversarial AI: data, compute power, brains

AI vs Security

- AI
  - learn, based on past/current observations
  - to recognize/detect: discriminative model D
  - to generate/simulate: generative model G
- Security
  - problems caused by selfish / malicious humans
  - CONFidentiality / INTegrity / AUTHentication

C I A

AI attacks Security

- They have all the data
  - web / social media / video conf hosting
    - Zoom, Google Hangout, Microsoft Teams, Cisco Webex
  - IoE: internet of everything
    - smart speakers: Amazon Alexa, Google Assistant, Apple Siri, …
- They have the best architectures: computers, GPUs, … .
  - Amazon AWS, Google Cloud Platform/Colab, Microsoft Azure Cloud, …
- They have the latest research / top experts
  - Google AI / DeepMind, Microsoft AI / Research, Nvidia Research, Adobe Research, …

- Q: Do we want to go against such an AI adversary?

AI attacks Security

C

- CONFidentiality: secret not leaked
  - human adversary: brains, manual
  - computer adversary: automated, brute force
  - but AI: beyond human capability, & with brains

- AI: could do inference attacks vs privacy/CONF
  - classify: recognise patterns
  - regress: predict relationships
  - cluster: recognise similar patterns

AI attacks Security

C

bbc.com/news/technology-51309186

Facebook settles facial recognition dispute

30 January 2020

- Q: do we want to be recognized by strangers?
- Q: do we want to be seen by people we know although they were not present?

AI attacks Security

C

cnet.com/home/smart-home/google-knows-what-you-look-like-heres-what-it-means-and-how-to-opt-out/

Google knows what you look like. Here's what it means and how to opt out

Google's Face Match technology isn't everywhere yet, but it's always looking. Find out what's happening with your face data and what you can do to stop it.

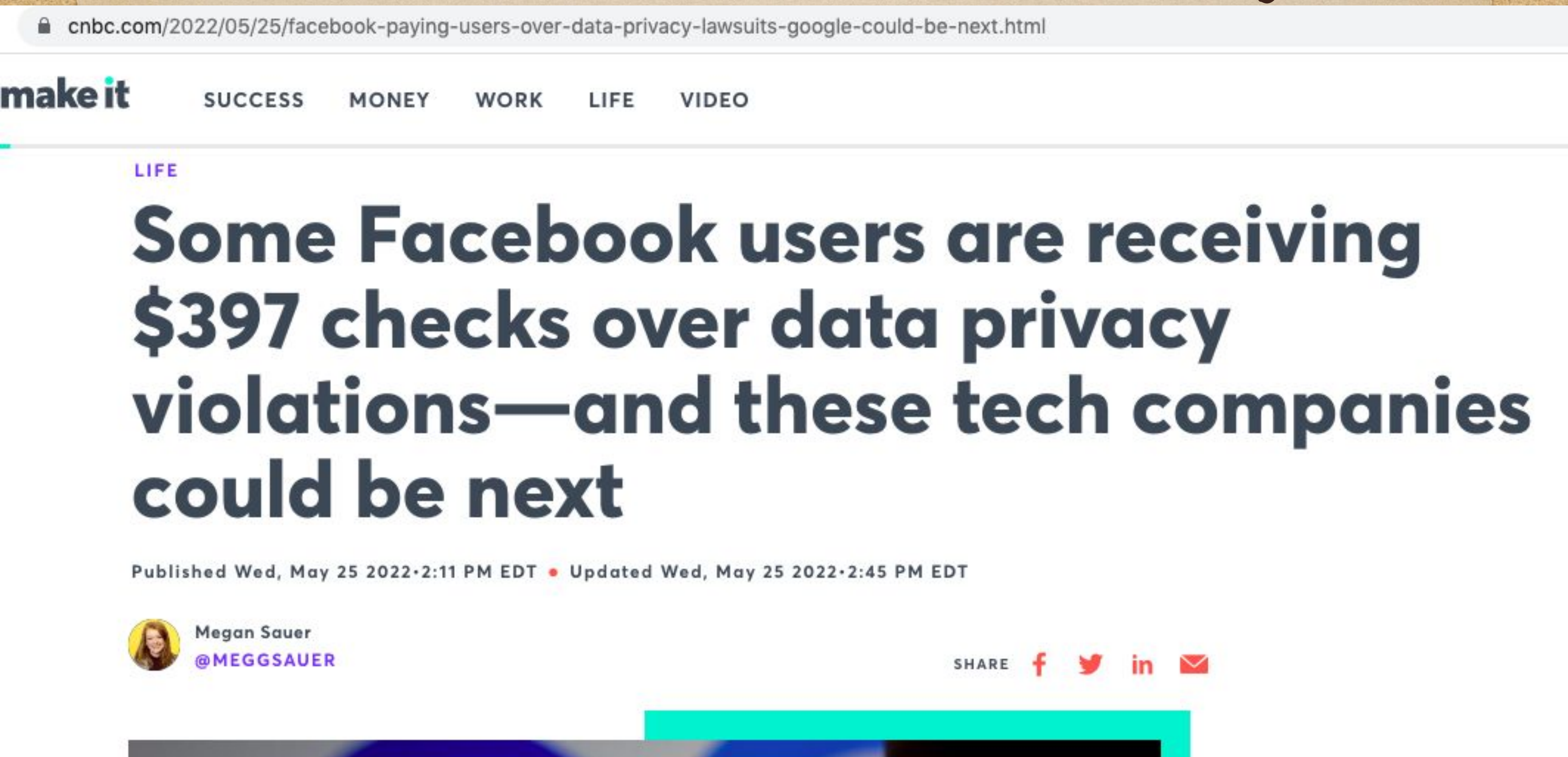
Dale Smith Feb. 4, 2020 5:00 a.m. PT

LISTEN - 07:09

28

- Note the dates, they are recent issues

AI attacks Security C


cnbc.com/2022/05/25/facebook-paying-users-over-data-privacy-lawsuits-google-could-be-next.html

make it SUCCESS MONEY WORK LIFE VIDEO

LIFE

Some Facebook users are receiving $397 checks over data privacy violations—and these tech companies could be next

Published Wed, May 25 2022·2:11 PM EDT • Updated Wed, May 25 2022·2:45 PM EDT

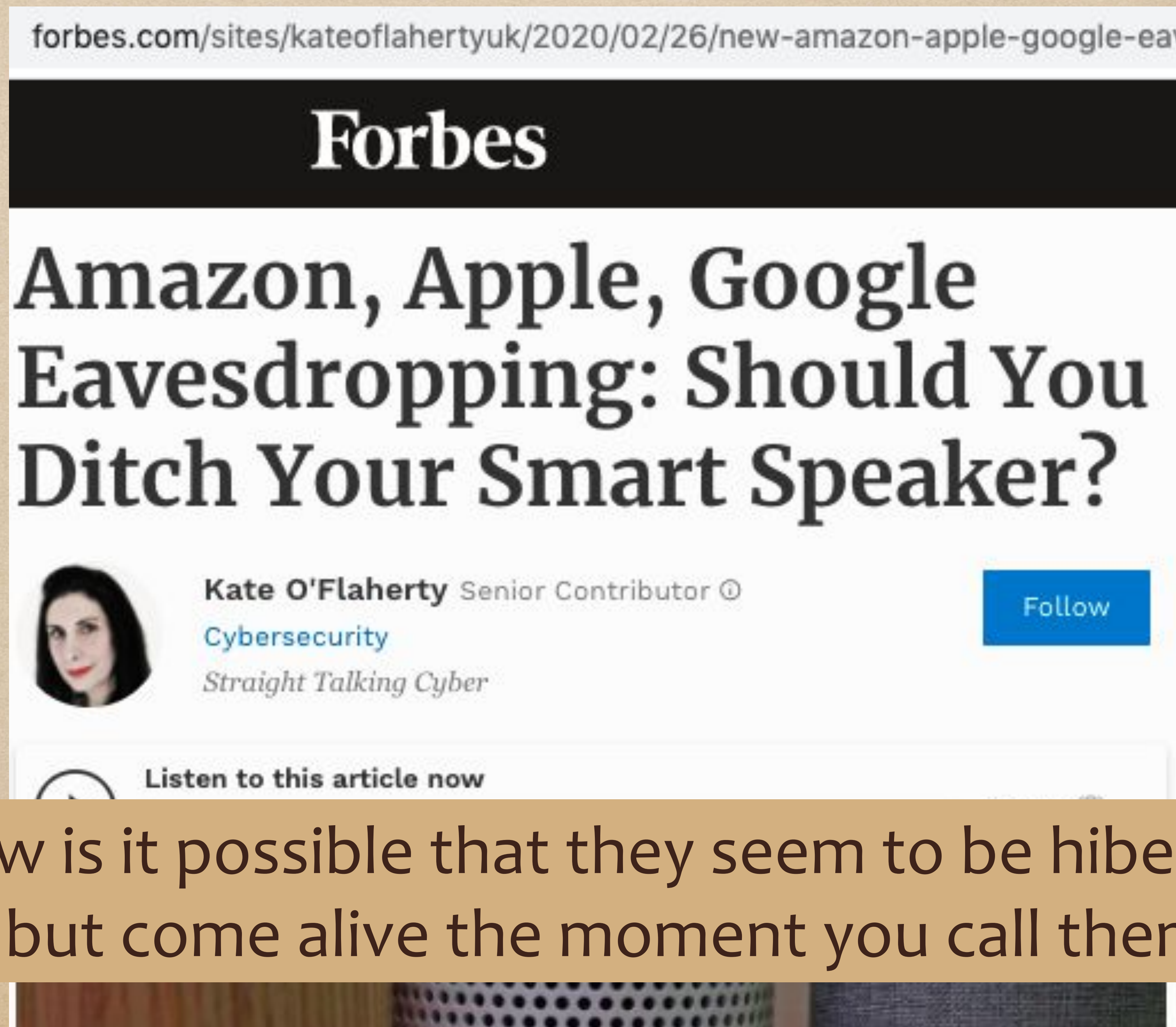
Megan Sauer
@MEGGSAUER

SHARE

- Note the dates, they are recent issues

AI attacks Security

C



forbes.com/sites/kateoflahertyuk/2020/02/26/new-amazon-apple-google-eav

Forbes

Amazon, Apple, Google Eavesdropping: Should You Ditch Your Smart Speaker?

Kate O'Flaherty Senior Contributor

Follow

Cybersecurity

Straight Talking Cyber

Listen to this article now

- Q: how is it possible that they seem to be hibernating, but come alive the moment you call them?

C

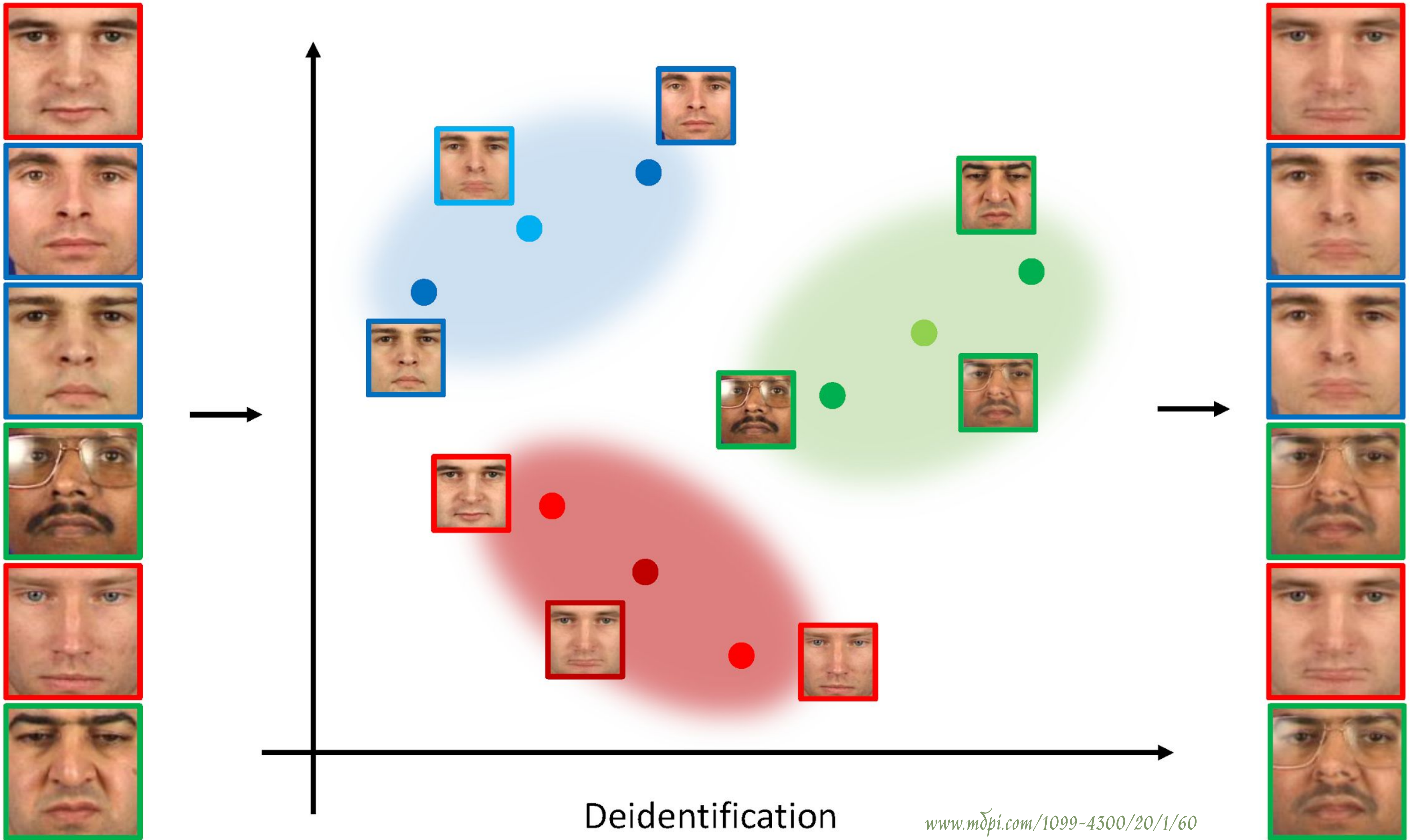
Privacy vs Inference Attacks

- Example: Database security
  - k-anonymity
  - l-diversity
  - t-closeness
  - differential privacy

Inference Attacks on CONF

C

- k-anonymity



AI attacks Security I

- AI attacks CONF: C
  - inference attacks on dBs / datasets
  - pattern recognition attacks on images/signals

- AI attacks INT & AUTH: I A
  - deepfakes

Deepfakes: AI attacks Security

I

- Integrity: data not changed, originally from source
  - AI to tamper without being detected
    - change existing, or
    - generate new fakes

- Deepfakes:
  - change existing images/videos/audio
    - face swap/transplantation A (via attacking INTegrity)
    - facial expression transfer incl lip syncing I
    - motion transfer I
  - generate new
    - puppet master I / A (if AUTH via soft biometrics)

AI attacks Security I

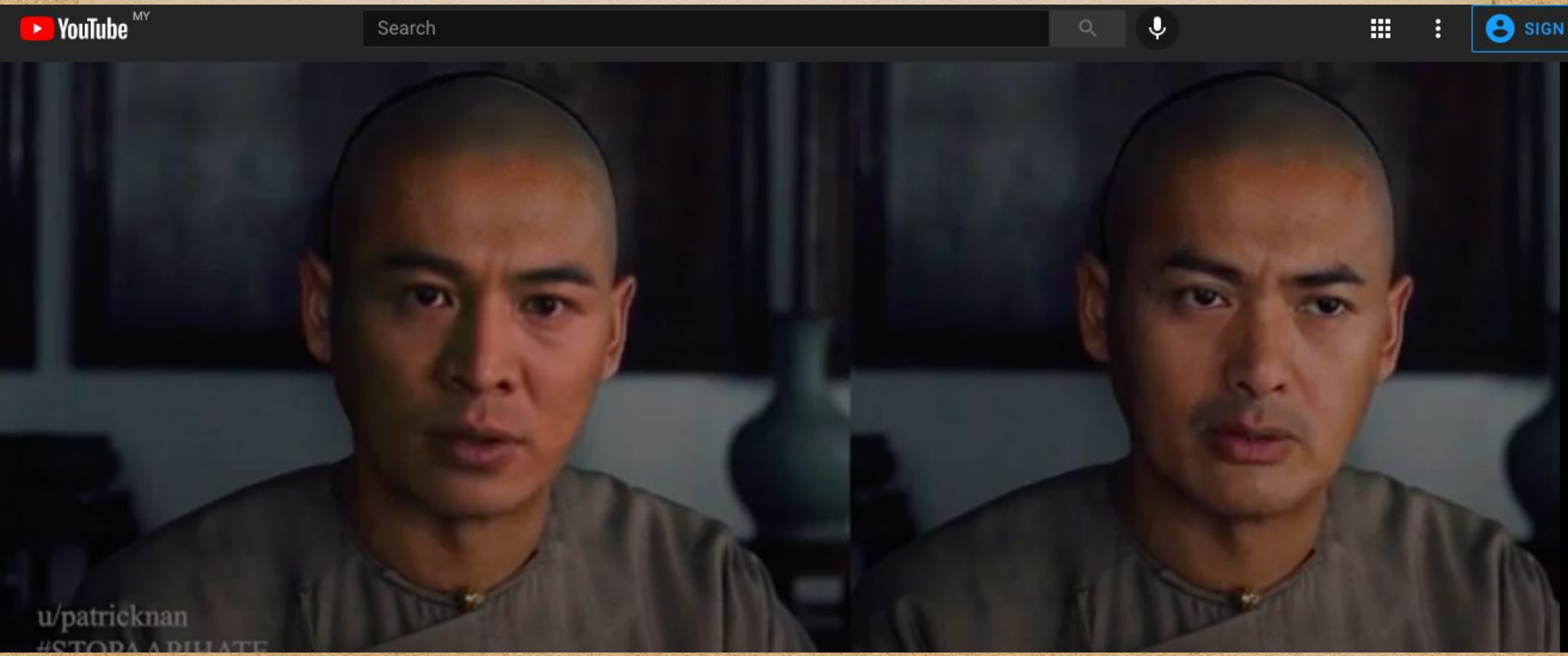
- AI attacks CONF: C
  - inference attacks on dBs / datasets

- AI attacks INT & AUTH: I A
  - deepfakes
    - face swap: attack on AUTH
      - e.g. Jet Li in Crouching Tiger Hidden Dragon https://www.youtube.com/watch?v=TWDA61-ht6Q

AI attacks Security

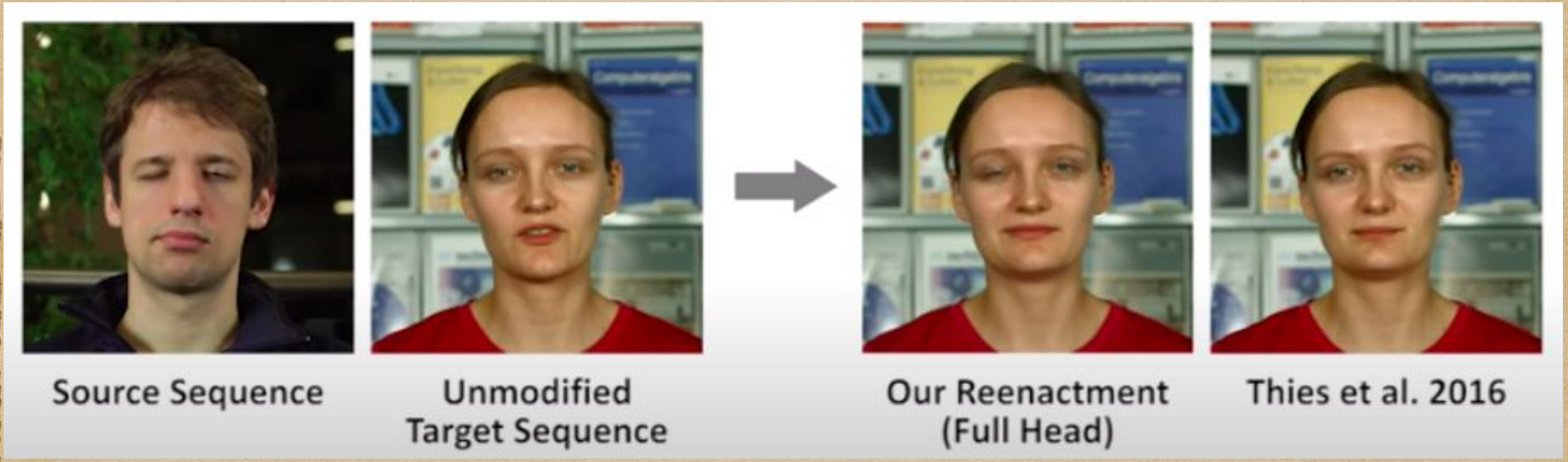
I

- face swap: attack on AUTH
  - e.g. Jet Li in Crouching Tiger Hidden Dragon https://www.youtube.com/watch?v=TWDA61-ht6Q



AI attacks Security I

- AI attacks INT & AUTH:
  - deepfakes
    - facial expression transfer: attack on INT
    - e.g. https://www.youtube.com/watch?v=qc5P2bvfl44



AI attacks Security

- AI attacks INT & AUTH:
  - deepfakes
    - face swap: attack on AUTH
    - facial expression transfer: attack on INT
    - puppet master: attack on INT / AUTH
      - https://www.youtube.com/watch?v=pAoTmlqMqjg
      - https://www.youtube.com/watch?v=UXGodiDAqiE
      - https://www.youtube.com/watch?v=qc5P2bvfl44

AI attacks Security

I

- puppet master: attack on INT / AUTH
  - https://www.youtube.com/watch?v=qc5P2bvfl44

AI attacks Security I

- puppet master: attack on INT / AUTH
  - https://www.youtube.com/watch?v=pAoTmIqMqj

Deepfakes: AI attacks Security

I

- Paper: Siarohin et al.: First Order Motion Model for Image Animation @NeurIPS 2019

- Code: https://colab.research.google.com/github/AliaksandrSiarohin/first-order-model/blob/master/demo.ipynb

- Demo: https://www.youtube.com/watch?v=lE-4w8q_5GU

First Order Motion Model for Image Animation

Aliaksandr Siarohin
DISI, University of Trento
aliaksandr.siarohin@unitn.it

Stéphane Lathuilière
DISI, University of Trento
LTCI, Télécom Paris, Institut polytechnique de Paris
stephane.lathuilire@telecom-paris.fr

Sergey Tulyakov
Snap Inc.
stulyakov@snap.com

Elisa Ricci
DISI, University of Trento
Fondazione Bruno Kessler
e.ricci@unitn.it

Nicu Sebe
DISI, University of Trento
Huawei Technologies Ireland
niculae.sebe@unitn.it

@NIPS 2019

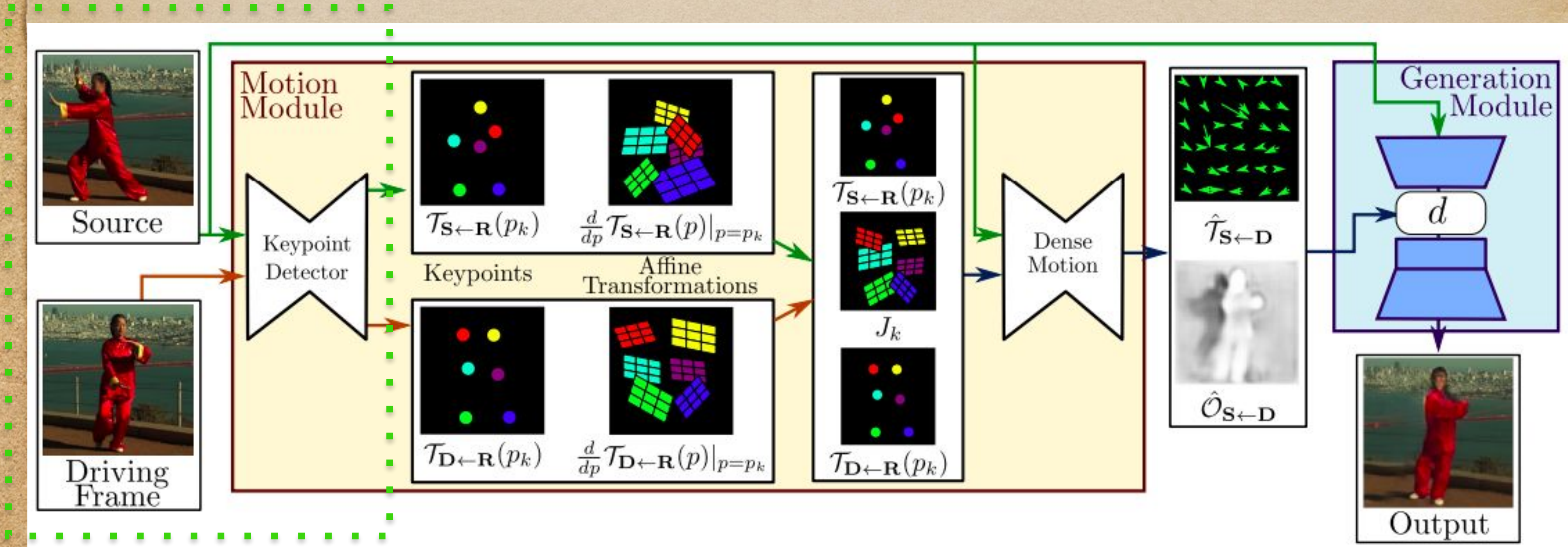
Abstract

Image animation consists of generating a video sequence so that an object in a source image is animated according to the motion of a driving video. Our framework addresses this problem without using any annotation or prior information about the specific object to animate. Once trained on a set of videos depicting objects of the same category (*e.g.* faces, human bodies), our method can be applied to any object of this class. To achieve this, we decouple appearance and motion information using a self-supervised formulation. To support complex motions,

1st Order Motion Model

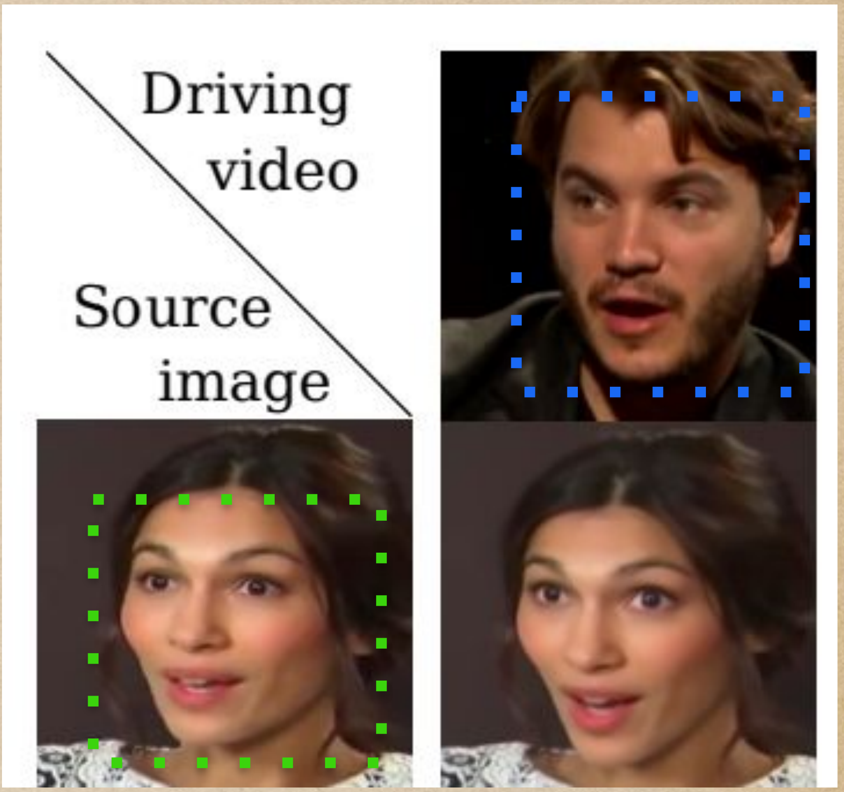
1st Order Motion Model

- Inputs
  - source image S: extract appearance & retain
  - driving video D: extract motion to animate S
- Output
  - source image S will move like video D

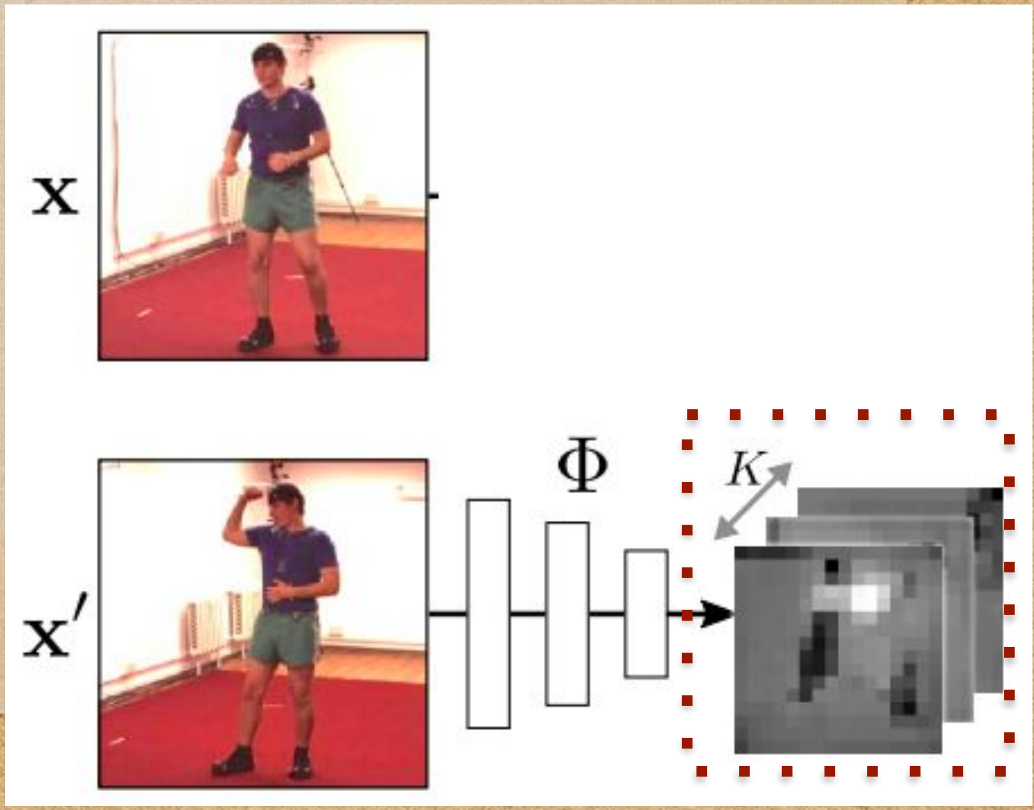


1st Order Motion Model

- Inputs:
  - source image S, driving video D
  - appearance, motion

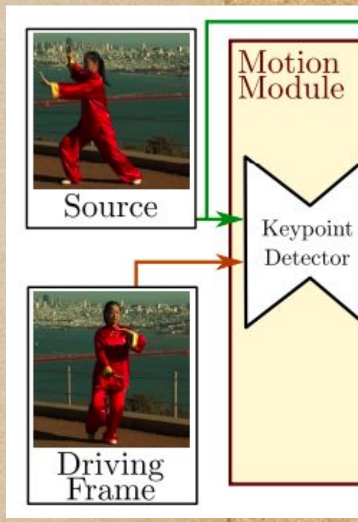


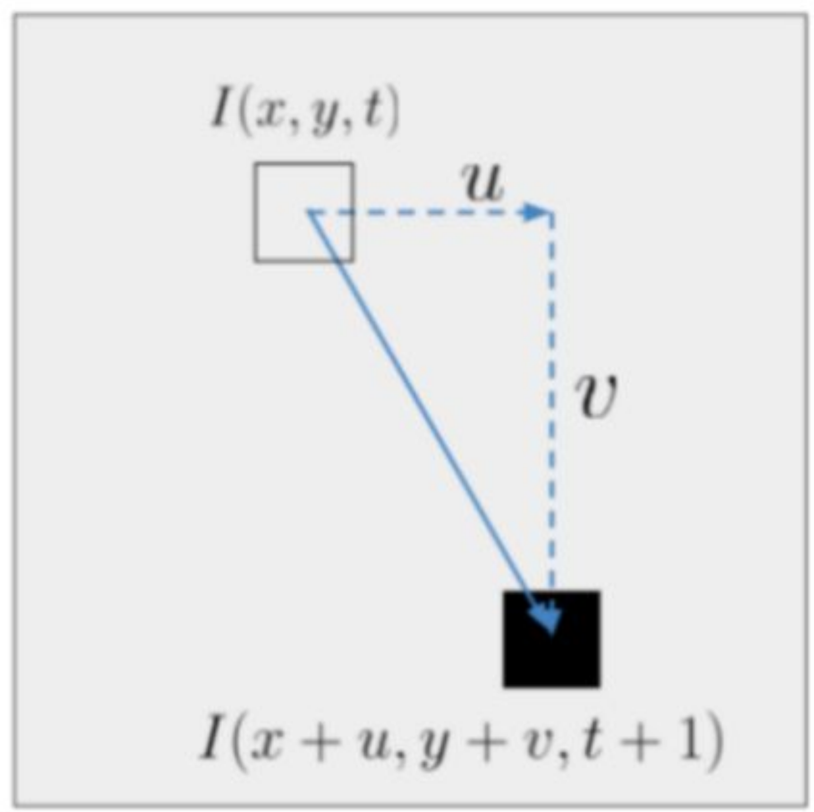
- Keypoint Detector
  - unsupervised detector
  - predicts key points K from S & D
  - key points / landmarks / heat map of important points



1st Order Motion Model

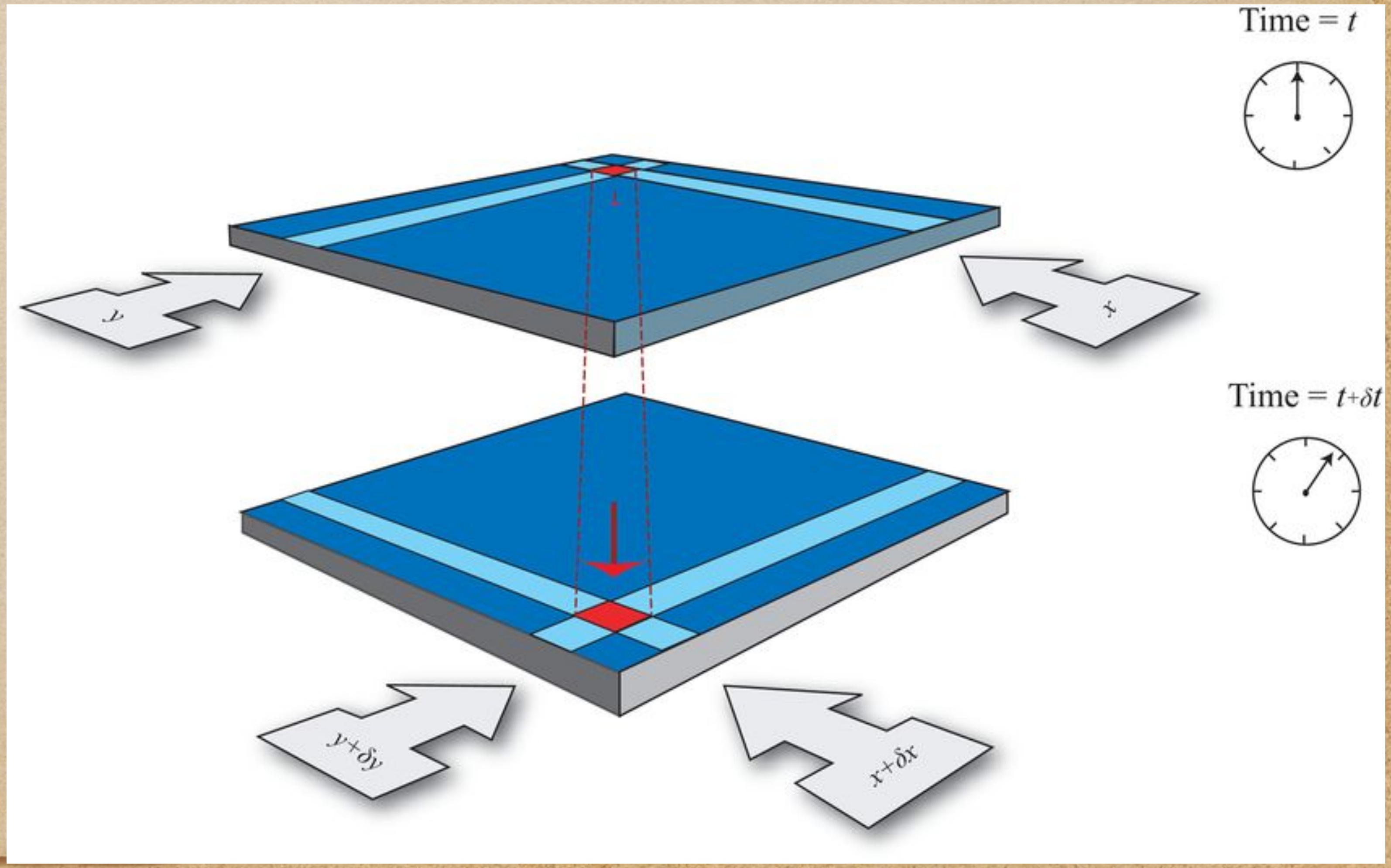
- Keypoints detected based on
  - comparing two frames (S & D with reference frame)
  - so includes motion info
- motion
  - displacement of pixel location by (u,v): x⟶x+u , y⟶y+v
    from current frame t to
    next frame t+1
- I(x,y,t) represents pixel intensity
  at location (x,y) in frame t
- Q: I(x+u,y+v,t+1) represents?





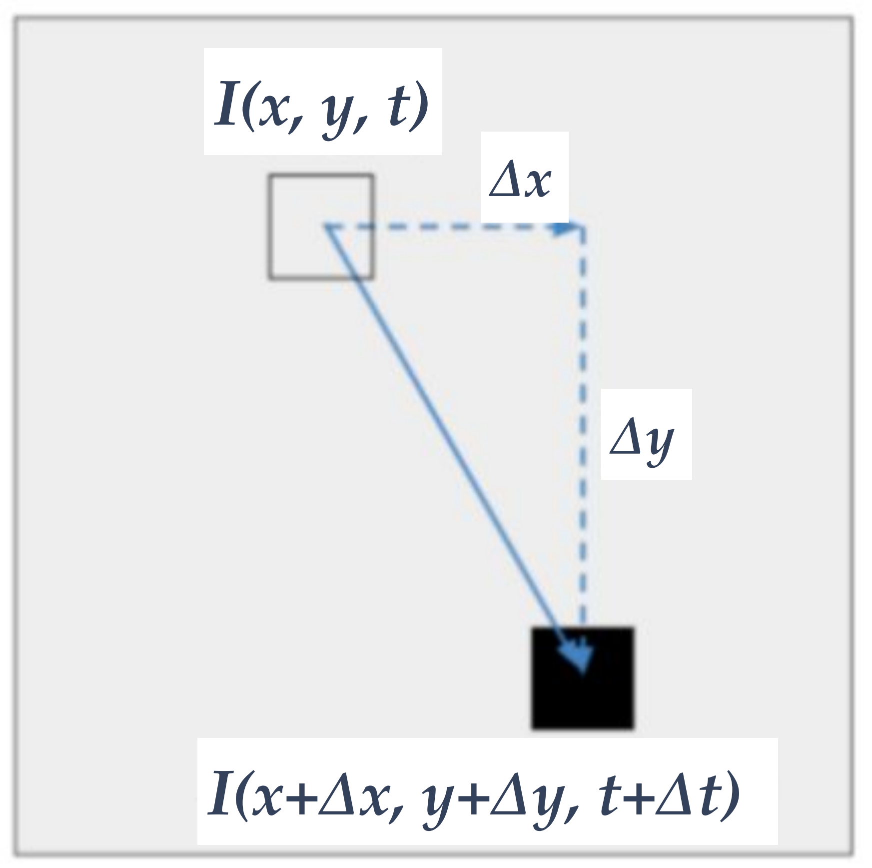
Motion Representation

- motion = displacement of pixel at (x,y) by (δx,δy) from current frame t to next frame t+δt , e.g. δt=1
  - optical flow (robotics, machine vision)
  - motion vector (video coding)



Motion Representation

- motion = displacement of pixel I(x,y) by (Δx,Δy) from current frame t to next frame t+Δt



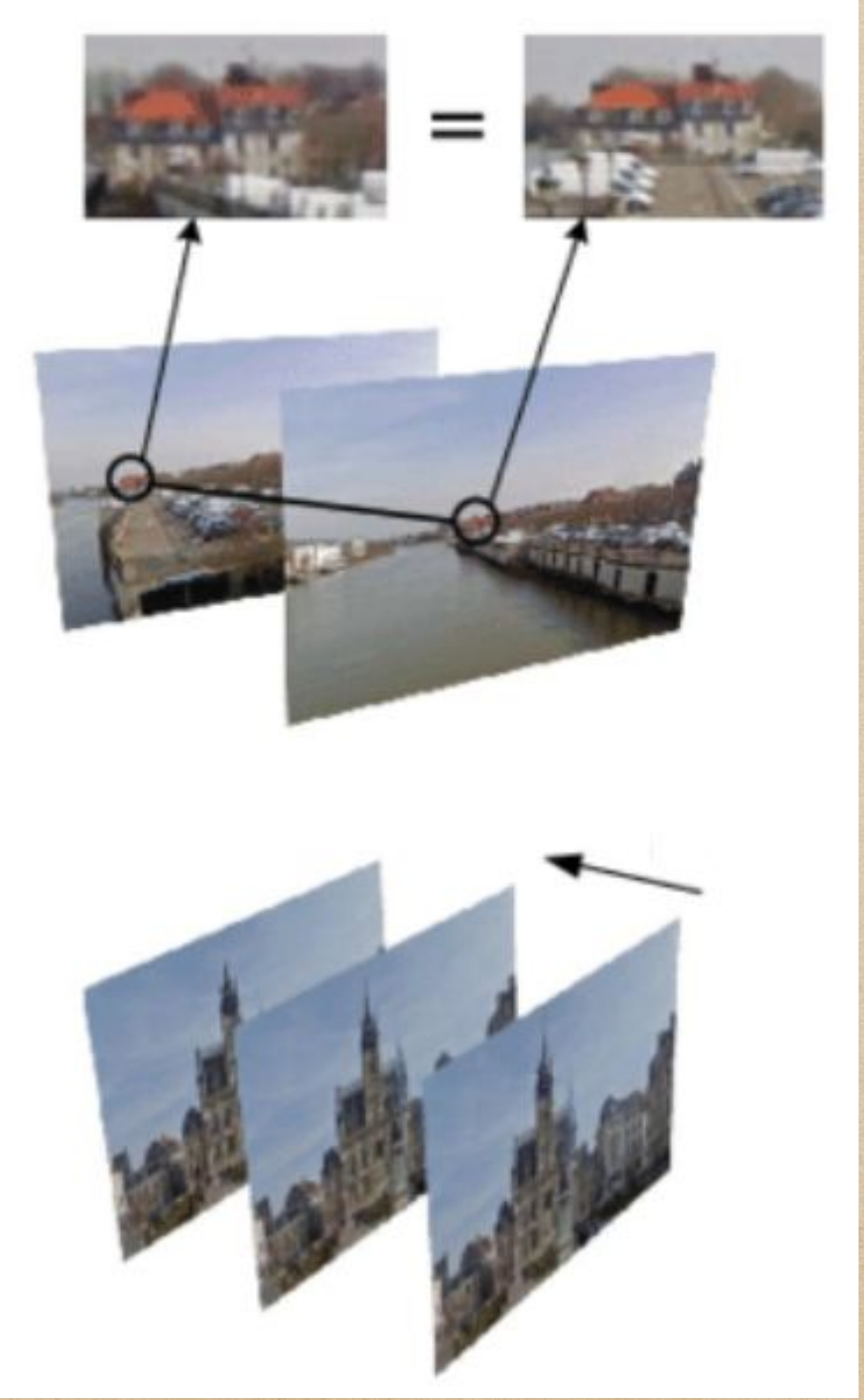
- optical flow (robotics,machine vision)
  - brightness constancy assumption:

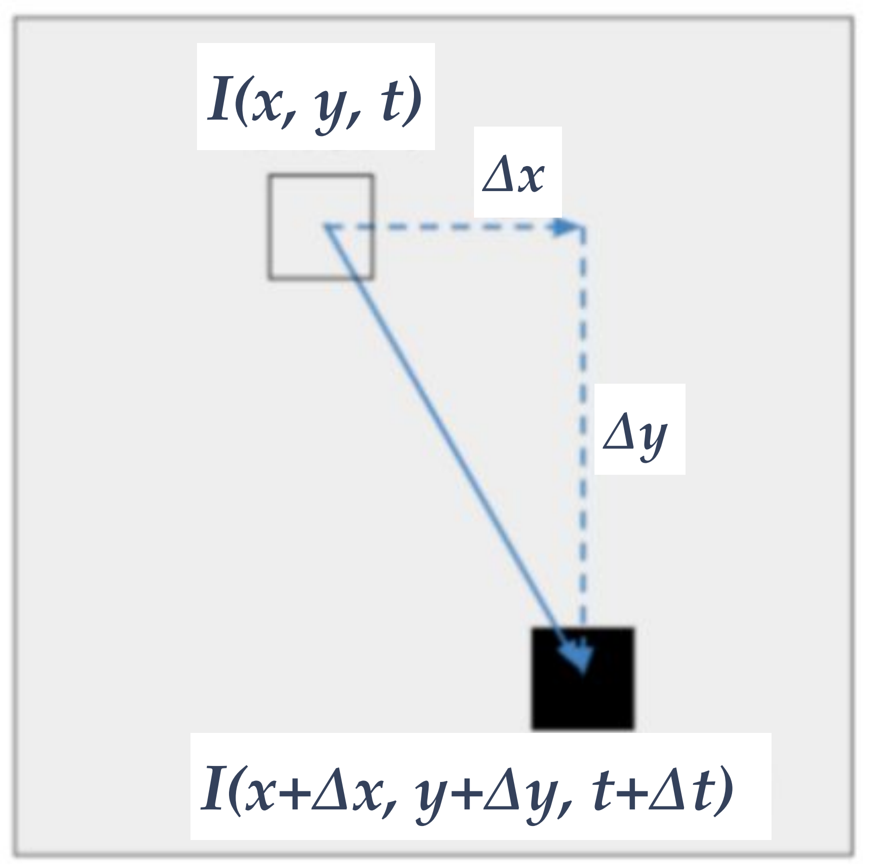
brightness (intensity) of small patch remains constant as it moves across time

$$\boxed{I(x,y,t) = I(x+\Delta x, y+\Delta y, t+\Delta t)}$$

i.e. though pixel at (x,y) of frame t has moved to (x+Δx, y+Δy) in frame t+Δt, the intensity remains the same

Brightness Constancy Assumption

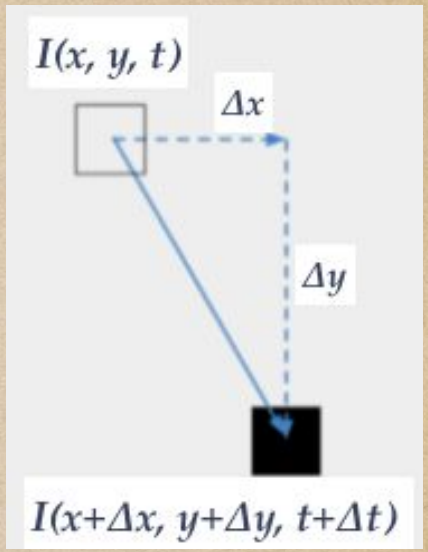




https://miro.medium.com/max/1000/1*6q7TXhpKGkLcvJjk6UZYFA.png

https://www.researchgate.net/figure/Feature-Tracking-Assumptions-The-brightness-constancy-which-assumes-that-the_fig19_265126161

Optical Flow*



- optical flow (robotics,machine vision)

$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t)$$

- Recap: Taylor series expansion of a function:

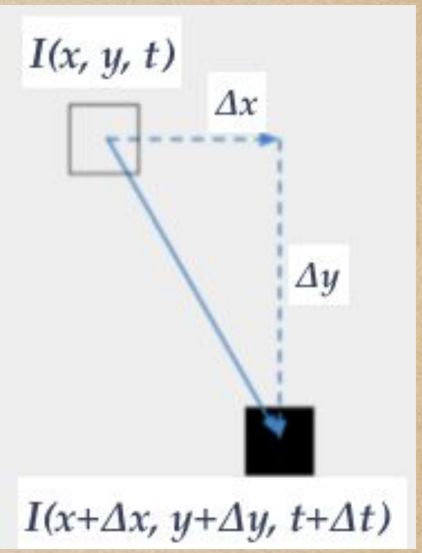
$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!}(x-a)^k = \boxed{f(a) + f'(a)(x-a)} + \frac{f''(a)}{2!}(x-a)^2 + \cdots$$

- Linear approximation of a function:

$$f(x) \approx \boxed{f(a) + f'(a)(x-a)}$$

let x = a+δ, then f(x) ≈ $\boxed{\text{f(a) + f'(a)(δ)}}$

Optical Flow*



- optical flow (robotics,machine vision)

$$I(x,y,t) = I(x+\Delta x, y+\Delta y, t+\Delta t)$$

- Taylor series expansion of a function:

$$I(x+\Delta x, y+\Delta y, t+\Delta t) = I(x,y,t) + \frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial u}\Delta y + \frac{\partial I}{\partial t}\Delta t + \ldots$$

- Linear approximation of a function:

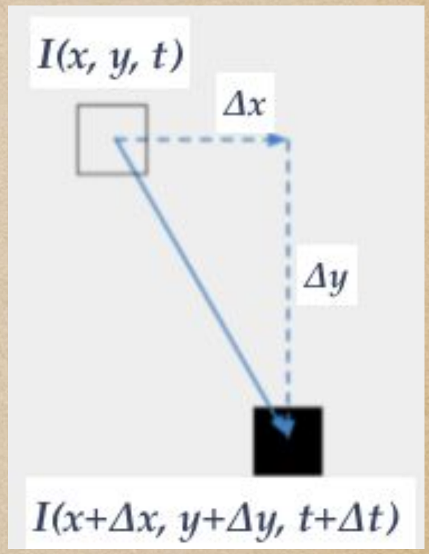
$$I(x+\Delta x, y+\Delta y, t+\Delta t) = I(x,y,t) + \frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t$$

- Due to brightness constancy assumption:

$$\frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t = 0$$

Optical Flow*

- optical flow (robotics,machine vision)



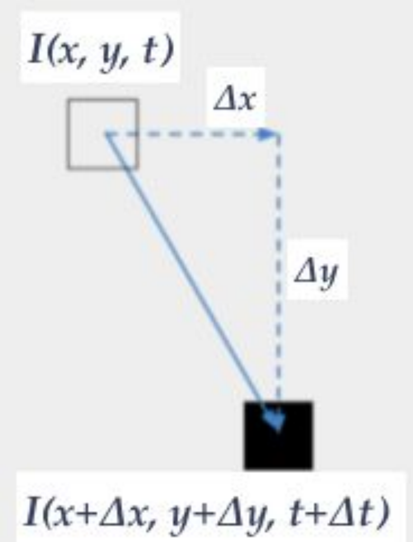
$$\frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t = 0$$

$$\frac{\partial I}{\partial x}\frac{\Delta x}{\Delta t} + \frac{\partial I}{\partial y}\frac{\Delta y}{\Delta t} + \frac{\partial I}{\partial t}\frac{\Delta t}{\Delta t} = 0$$

$$\frac{\partial I}{\partial x}V_x + \frac{\partial I}{\partial y}V_y + \frac{\partial I}{\partial t} = 0$$

- V_x and V_y are the x and y components forming the optical flow of I(x,y) from frame t to frame t+Δt

Motion Representation



- optical flow (robotics,machine vision)

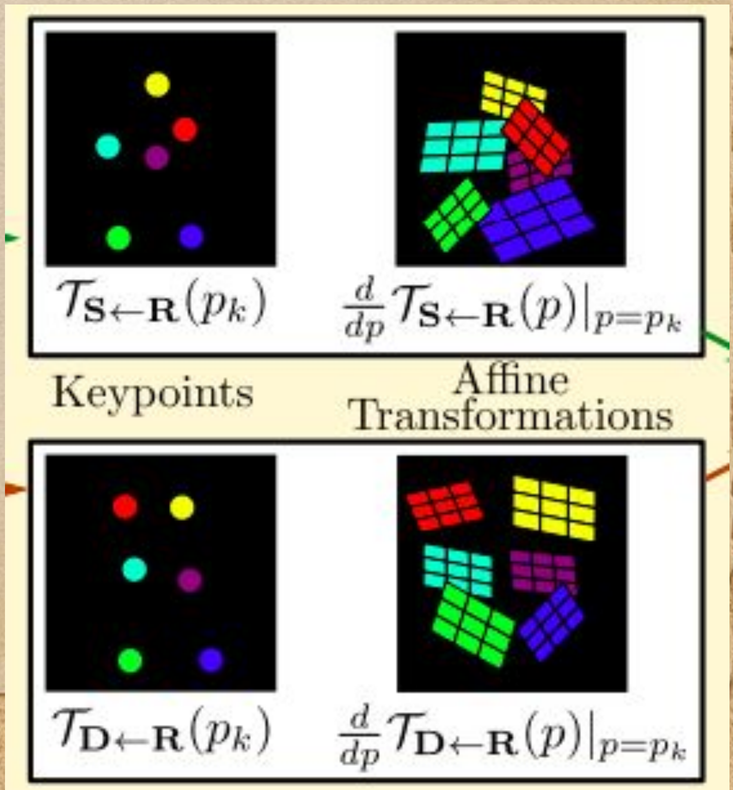
$$V_x = \frac{\Delta x}{\Delta t} \qquad V_y = \frac{\Delta y}{\Delta t}$$

⇒change in x and y as move from frame t to t+1

- Jacobian = matrix of all 1st order partial derivatives

⇒change in I(x,y,t) as x,y,t changes

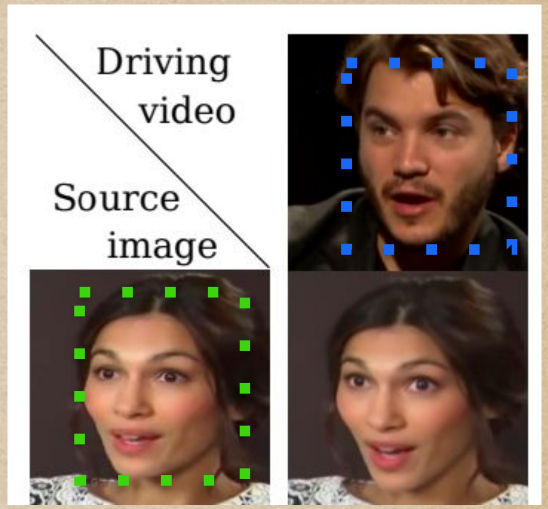
$$\begin{pmatrix} \frac{\partial I}{\partial x} \\ \frac{\partial I}{\partial y} \\ \frac{\partial I}{\partial t} \end{pmatrix}$$

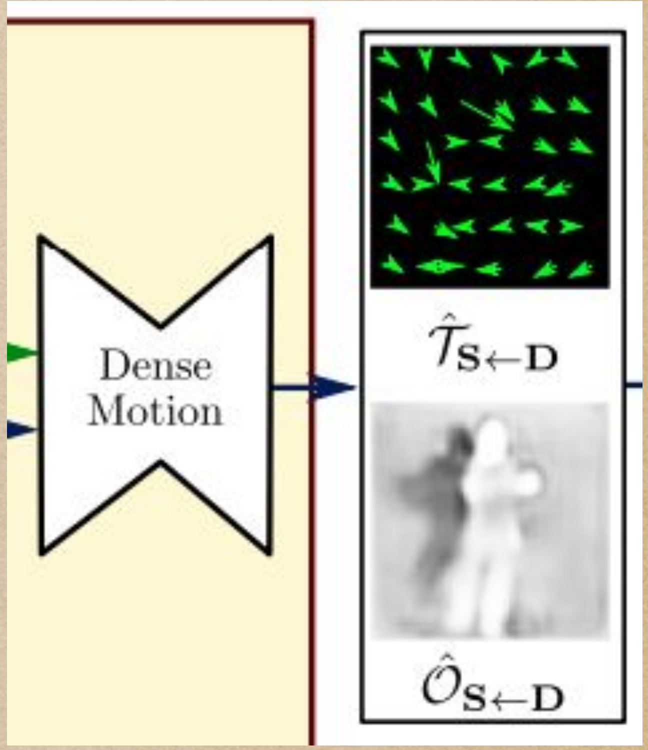


1st Order Motion Model

- Recall, aim: S to move like D

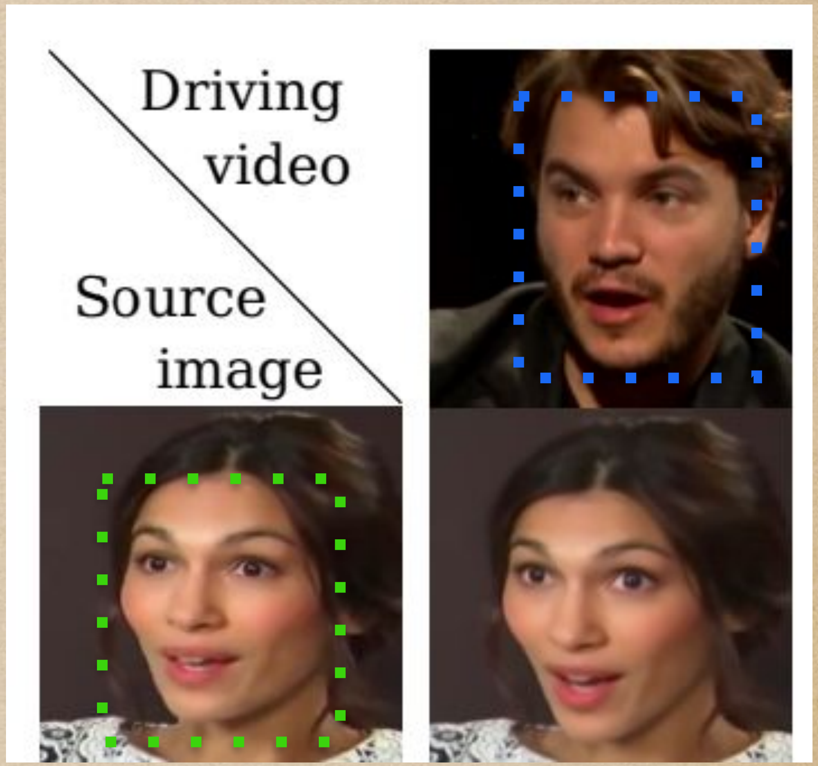
- Motion Mapping from D to S
  - incl warping
  - else source image S movements will be distorted e.g. mouth location in S & D different

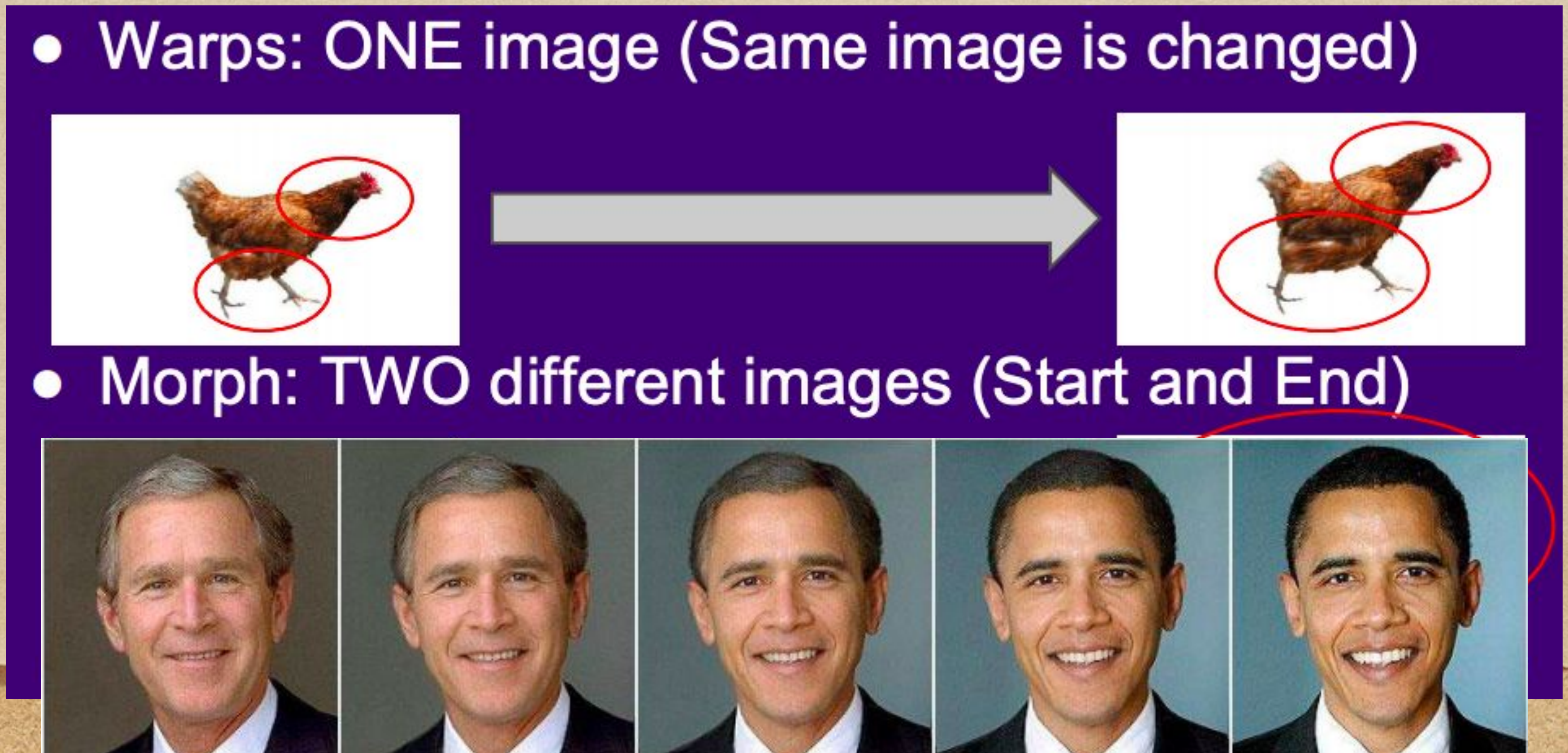




Warping

- Warping
  - change/distort the form/shape
- Morphing
  - gradual transformation of one image to another

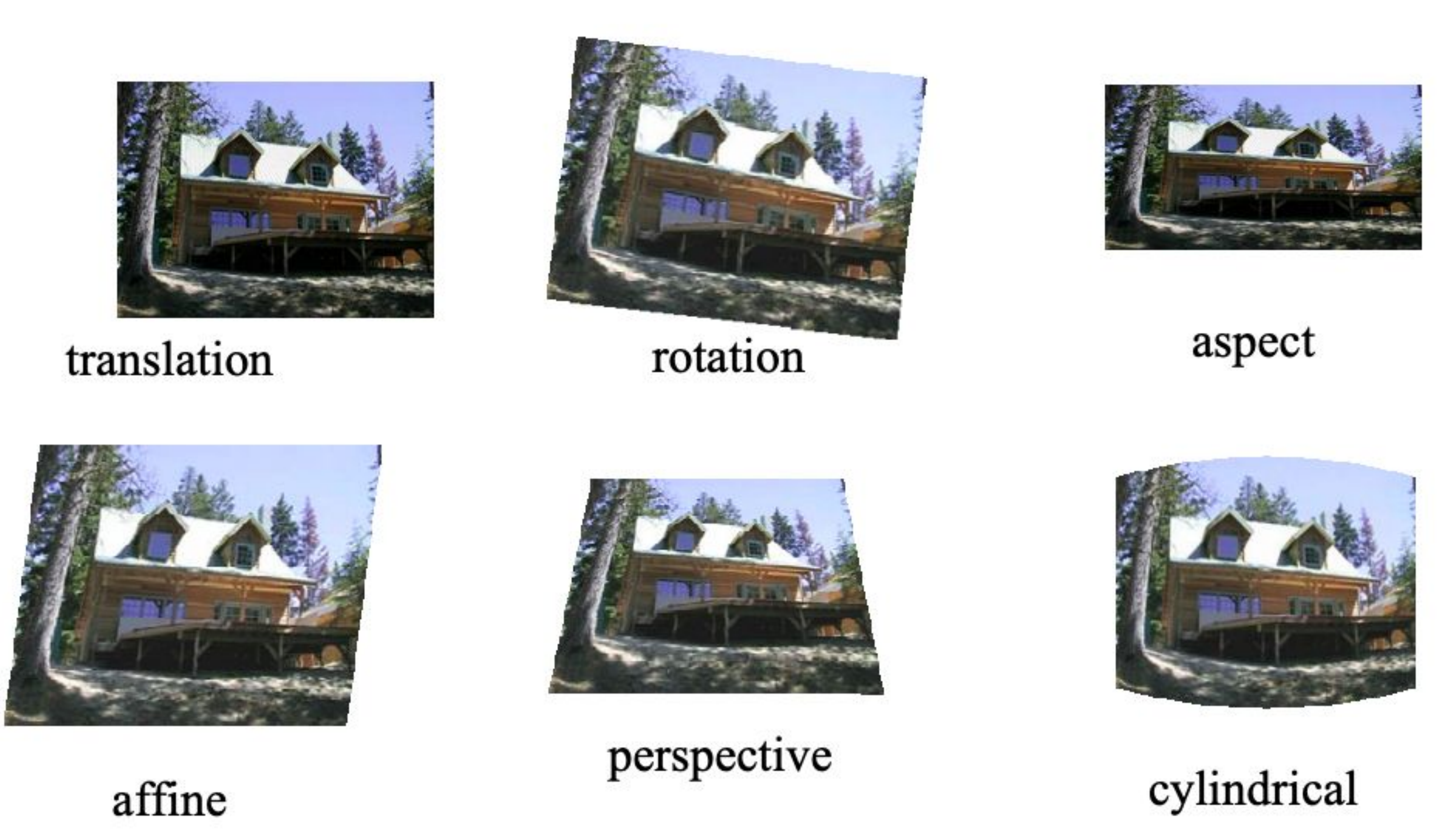




https://www.giss.nasa.gov/edu/nycri/research/files/2015/2015-Hostos-Prince_presentation.pdf

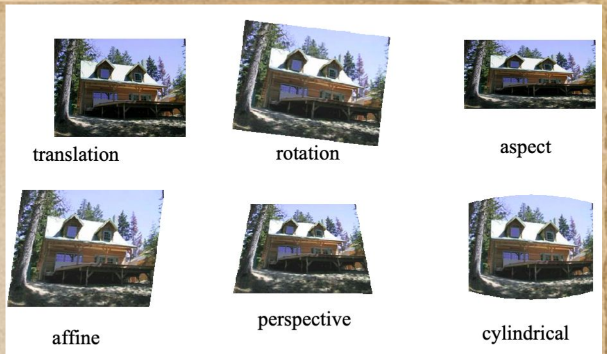
Warping

- change/distort the form/shape

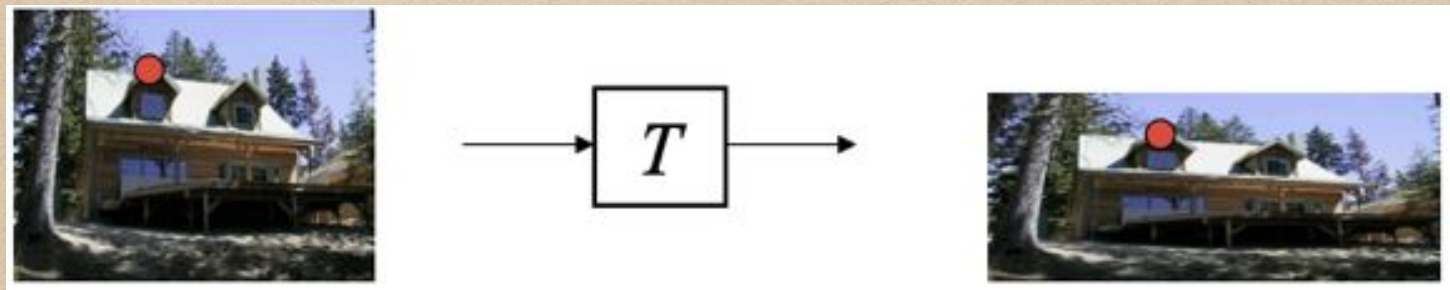


http://graphics.cs.cmu.edu/courses/15-463/2006_fall/www/Lectures/warping.pdf

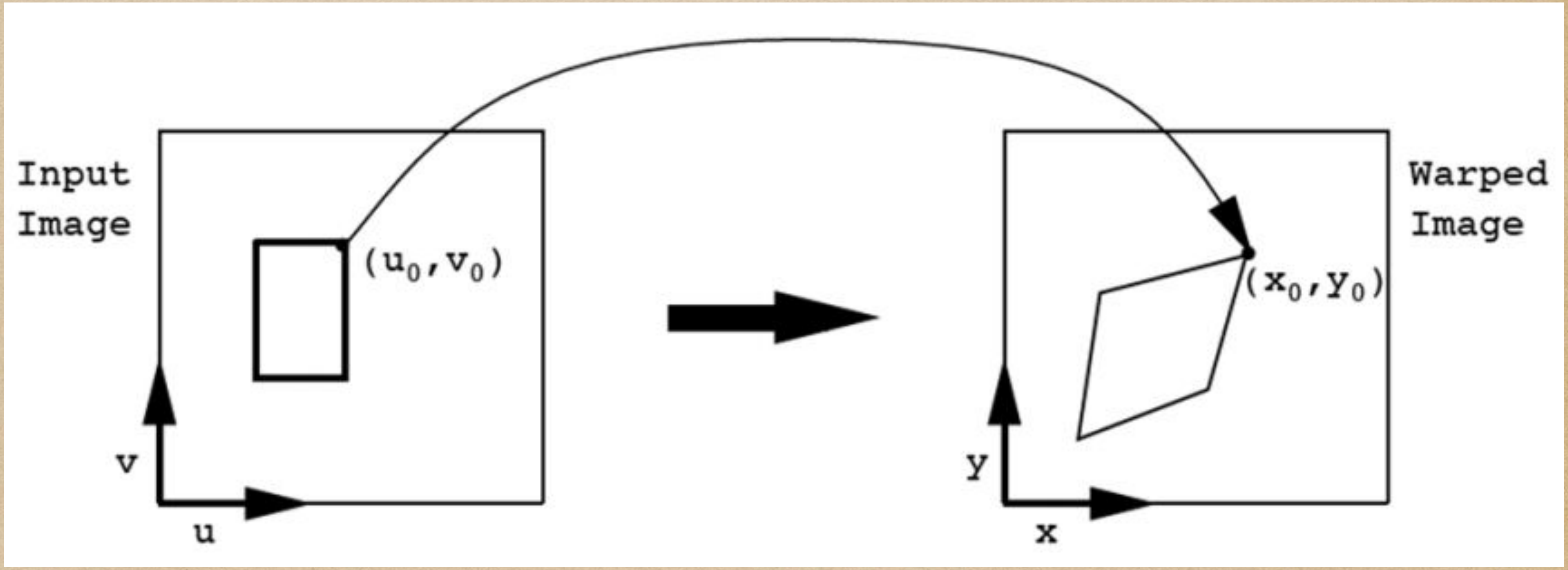
Warping



- change/distort the form/shape



for a point at (u,v), warped position at (x,y)



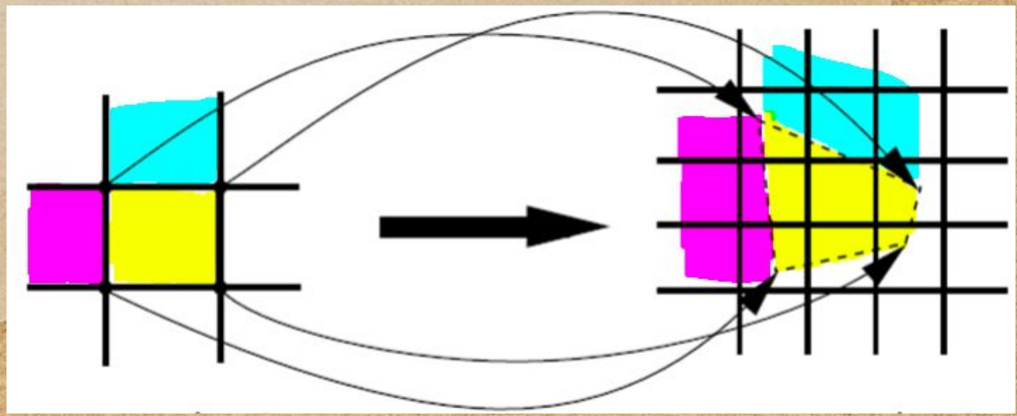
http://docdingle.com/teaching/cs545/presents/p12a_cs545_Warps.pdf

Recap: Geometric Transformations*

- express original (u,v) & warped (x,y) points as column vectors in homogeneous form (extra 1 row)
- then geometric transformations just matrix operations
- e.g. Affine transformation

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}$$

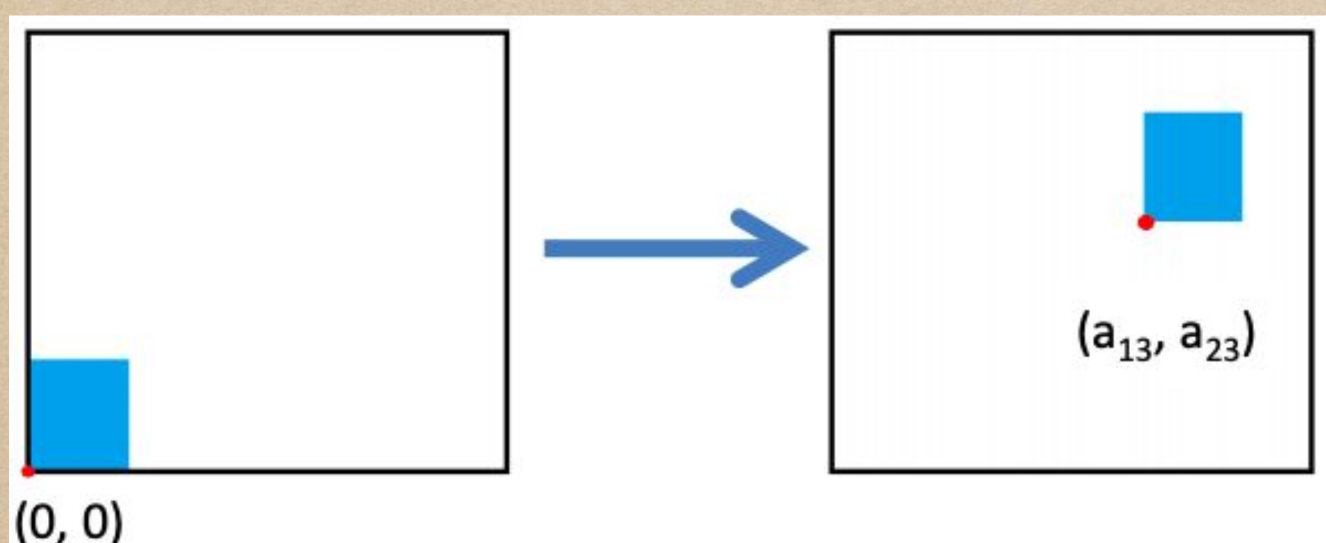
$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} a_{11}u + a_{12}v + a_{13} \\ a_{21}u + a_{22}v + a_{23} \end{bmatrix}$$

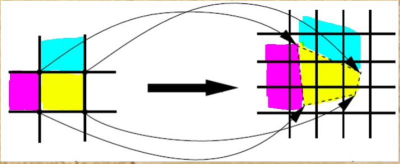


Recap: Geometric Transformations*

- Affine incl Scale, Translate, Rotate, Shear …
- Translate

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} u + a_{13} \\ v + a_{23} \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & a_{13} \\ 0 & 1 & a_{23} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}$$

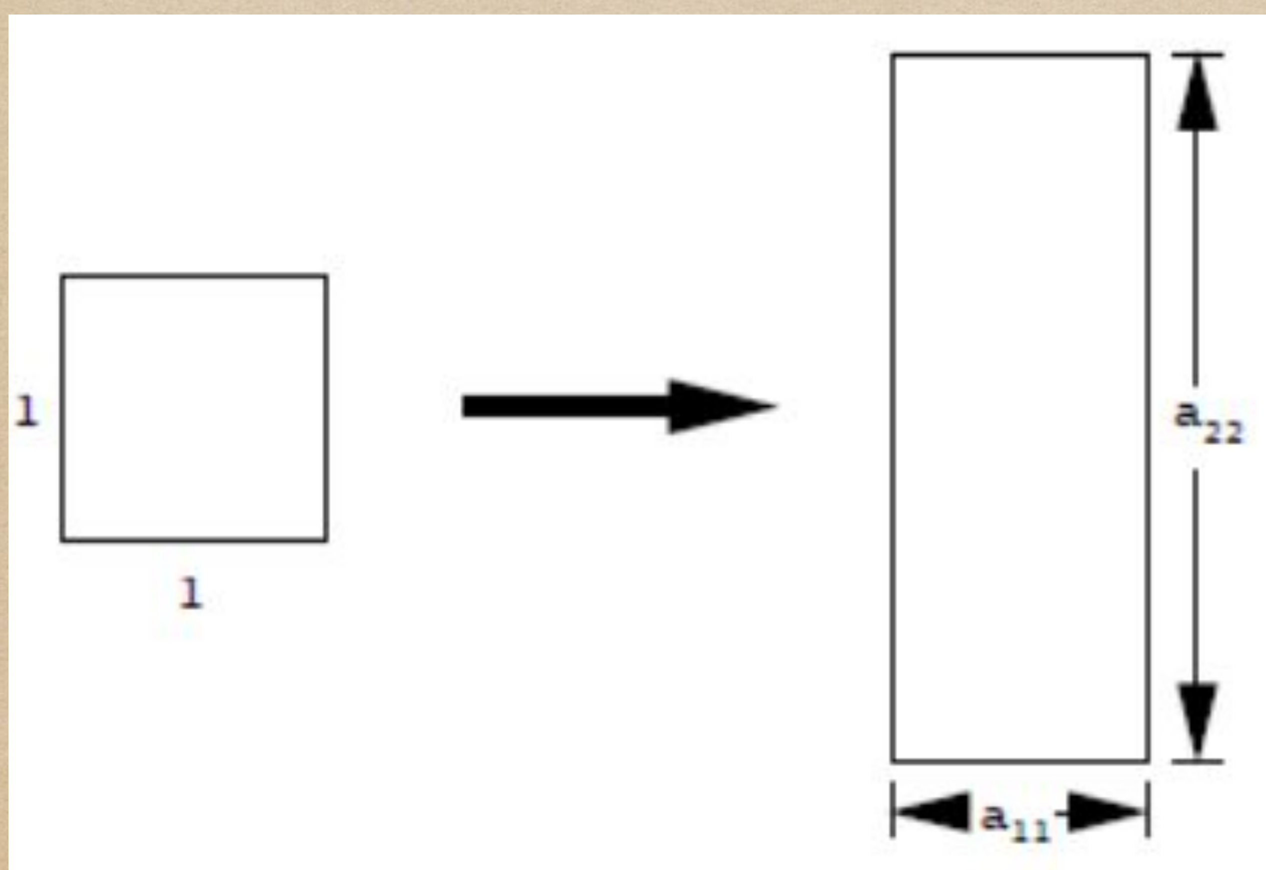


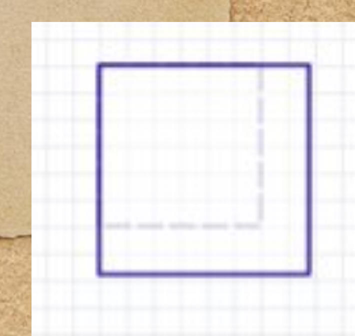


Recap: Geometric Transformations*

- Affine incl Scale, Translate, Rotate, Shear …
- Scale

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} a_{11}u \\ a_{22}v \\ 1 \end{bmatrix} = \begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}$$

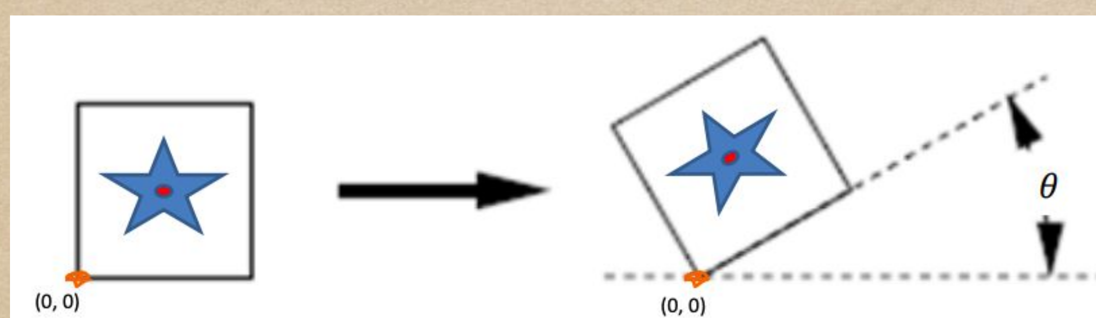


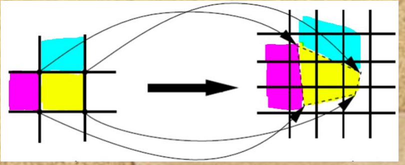


Recap: Geometric Transformations*

- Affine incl Scale, Translate, Rotate, Shear …
- Rotate

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}$$

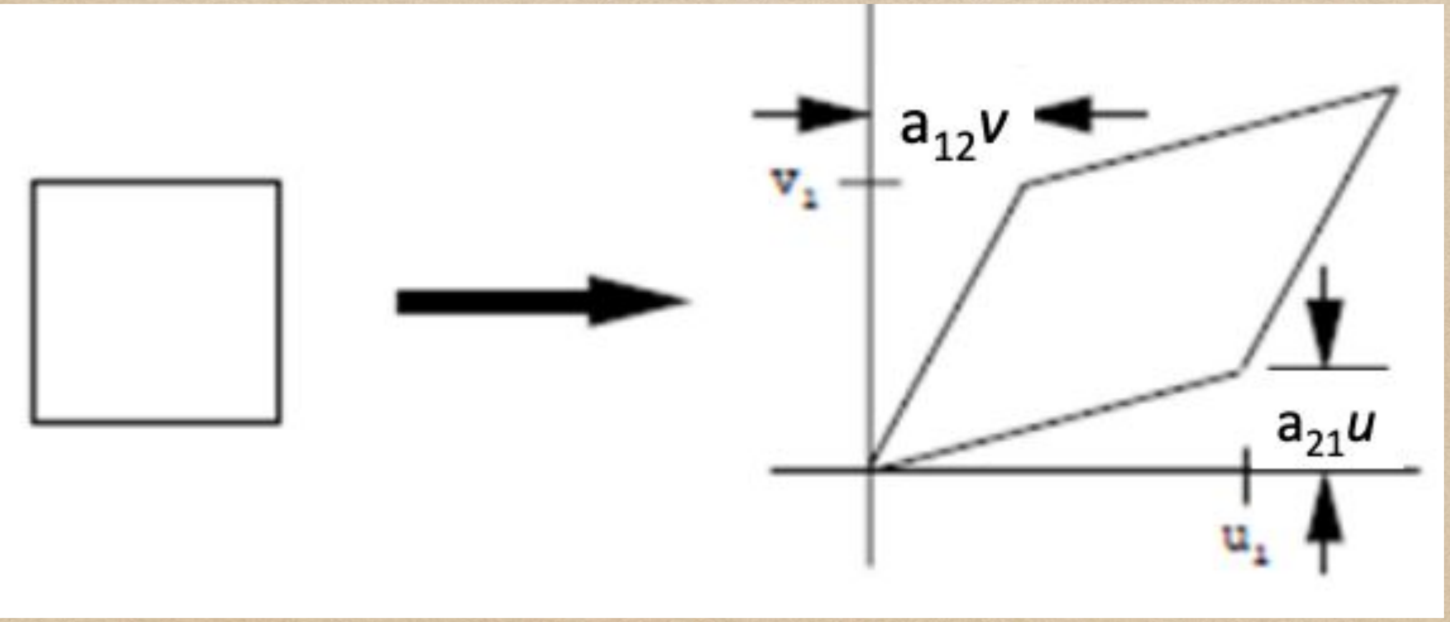


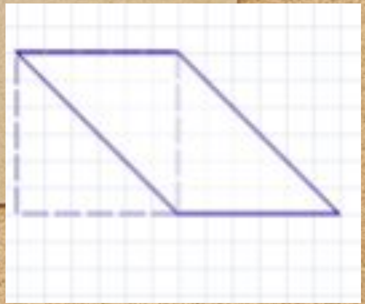


Recap: Geometric Transformations*

- Affine incl Scale, Translate, Rotate, Shear …
- Shear

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} u + a_{12}v \\ a_{21}u + v \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & a_{12} & 0 \\ a_{21} & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}$$





Projective Warping*

- Non-Affine transformation
- Homogeneous coordinate's 3rd element is no longer 1

e.g.

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ a_{31} & a_{32} & 1 \end{bmatrix} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} u \\ v \\ \boxed{a_{31}u + a_{32}v + 1} \end{bmatrix} = \mathrm{w}$$

- Perspective warping is special case where 1st and 2nd components (u,v) unchanged, only 3rd component is changed, as like above example

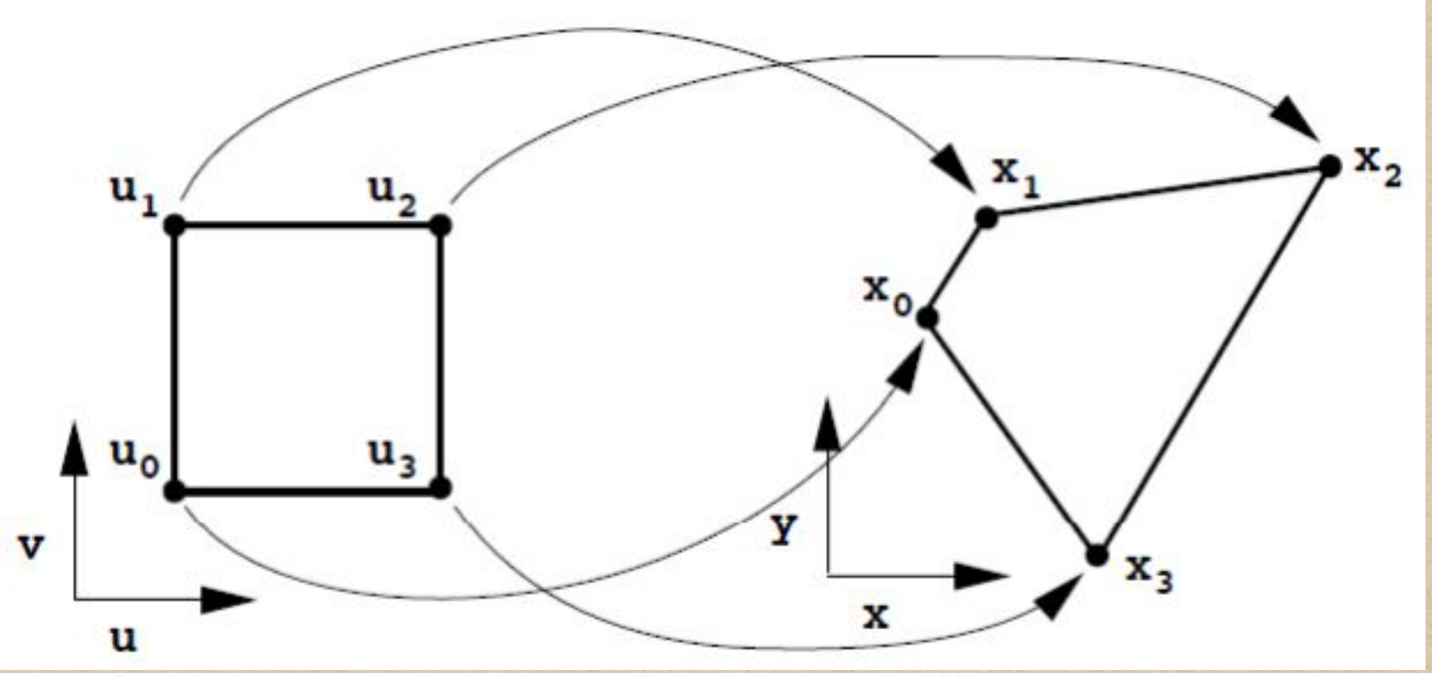
Projective Warping*

- Affine warping $\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}$
    - last row of transformation matrix only had [0 0 1]
- Non-Affine warping
    - last row of transformation matrix has non-zero/one

$$\begin{pmatrix} xw \\ yw \\ w \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \begin{pmatrix} a_{11}u + a_{12}v + a_{13} \\ a_{21}u + a_{22}v + a_{23} \\ a_{31}u + a_{32}v + a_{33} \end{pmatrix}$$

- to get back the 2D coordinates (x,y) of the warped point, just divide by w $\begin{pmatrix} xw \\ yw \\ w \end{pmatrix} \rightarrow \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}$

Projective Warping*



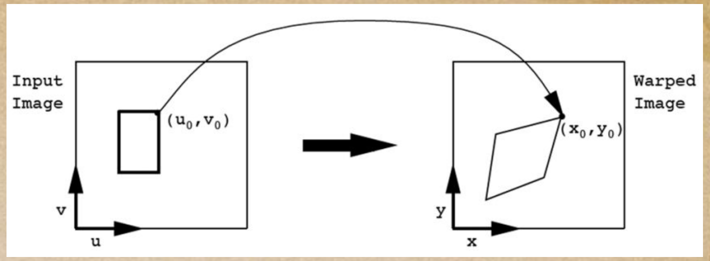
$$\begin{pmatrix} xw \\ yw \\ w \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \begin{pmatrix} a_{11}u + a_{12}v + a_{13} \\ a_{21}u + a_{22}v + a_{23} \\ a_{31}u + a_{32}v + a_{33} \end{pmatrix}$$

$$\begin{pmatrix} xw \\ yw \\ w \end{pmatrix} \rightarrow \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}$$

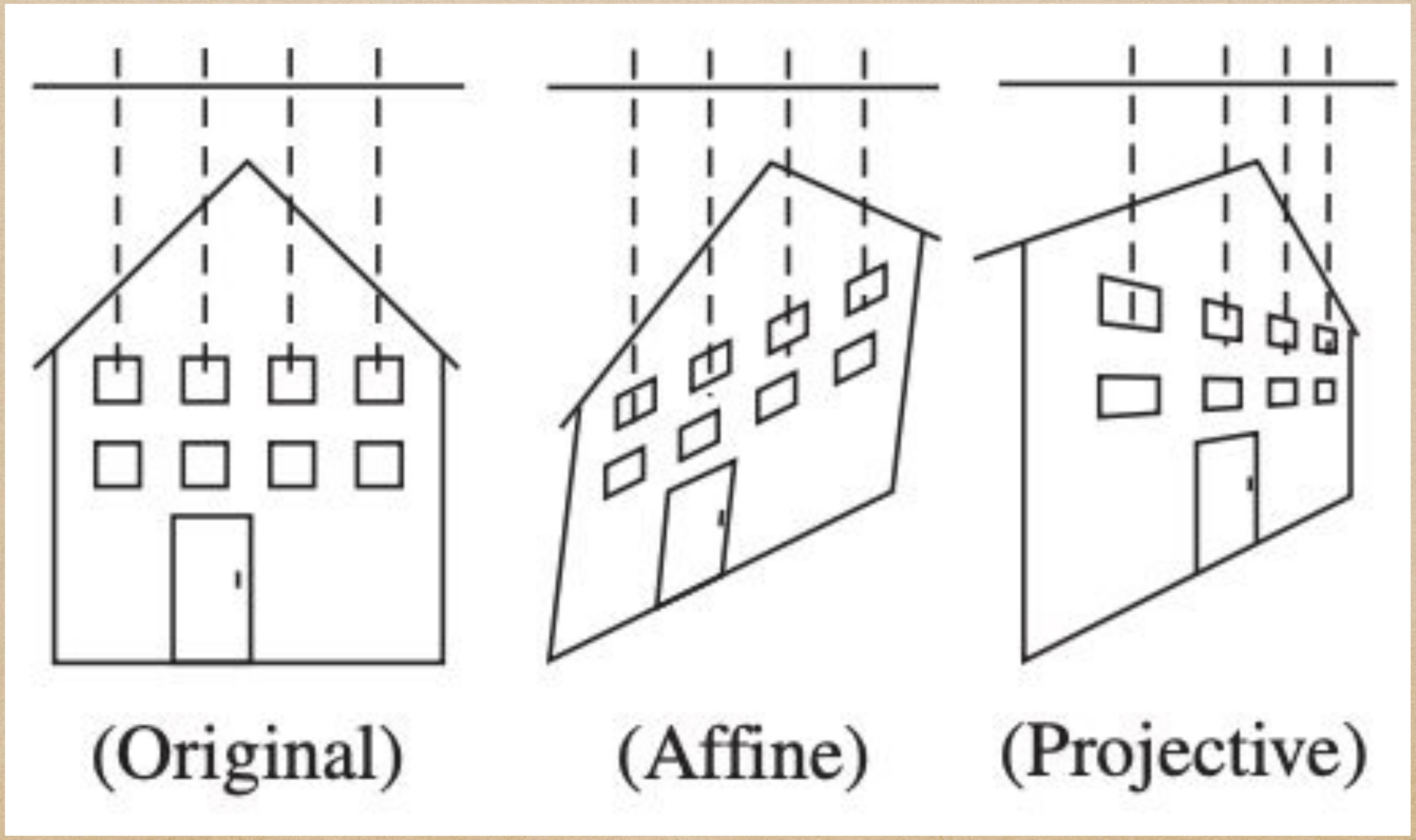
$$x_i = \frac{a_{11}u_i + a_{12}v_i + a_{13}}{a_{31}u_i + a_{32}v_i + a_{33}}$$

$$y_i = \frac{a_{21}u_i + a_{22}v_i + a_{23}}{a_{31}u_i + a_{32}v_i + a_{33}}$$

Warping



- change/distort the form/shape
- for a point at (u,v), warped position at (x,y)
- Affine vs Non-Affine (Projective)



Deepfakes: AI attacks Security

- Paper: Siarohin et al.: Motion-supervised Co-Part Segmentation @ICPR 2021

- Code: https://github.com/AliaksandrSiarohin/motion-cosegmentation

- Demo: https://www.youtube.com/watch?v=RJ4Nj1wV5iA

Motion-supervised Co-Part Segmentation

Aliaksandr Siarohin[1*], Subhankar Roy[1,4*], Stéphane Lathuilière[2], Sergey Tulyakov[3], Elisa Ricci[1,4], and Nicu Sebe[1,5]

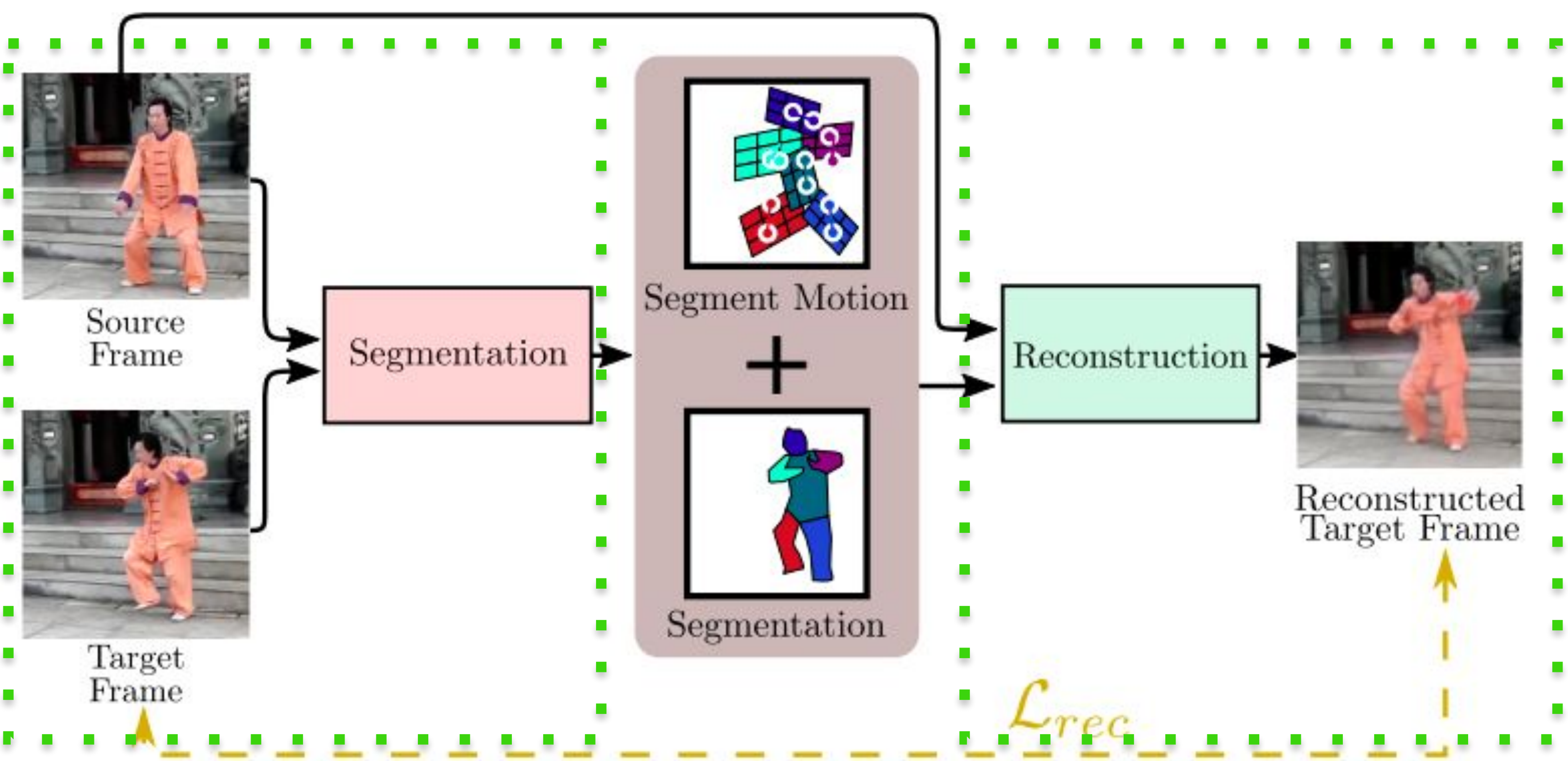
[1]DISI, University of Trento; [2]LTCI, Tlcom Paris, Institut polytechnique de Paris; [3]Snap Inc.; [4]Fondazione Bruno Kessler; [5]Huawei Technologies Ireland

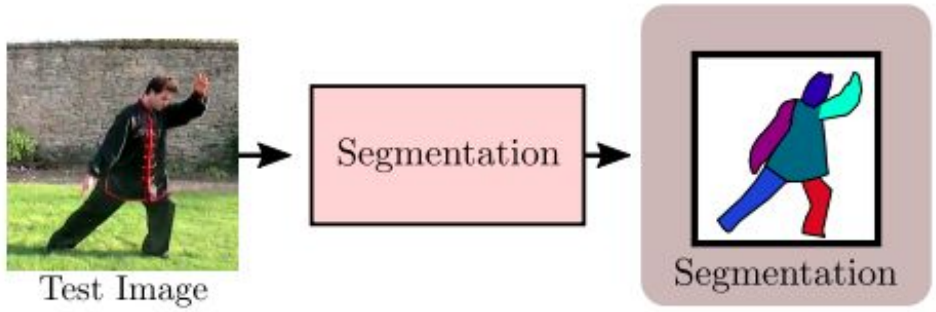
@ICPR 2021

Abstract. Recent co-part segmentation methods mostly operate in a supervised learning setting, which requires a large amount of annotated data for training. To overcome this limitation, we propose a self-supervised deep learning method for co-part segmentation. Differently from previous works, our approach develops the idea that motion information inferred from videos can be leveraged to discover meaningful object parts. To this end, our method relies on pairs of frames sampled from the same video. The network learns to predict part segments together with a representation of the motion between two frames, which permits reconstruction of the target image. Through extensive experi-

Motion-based CoSegmentation

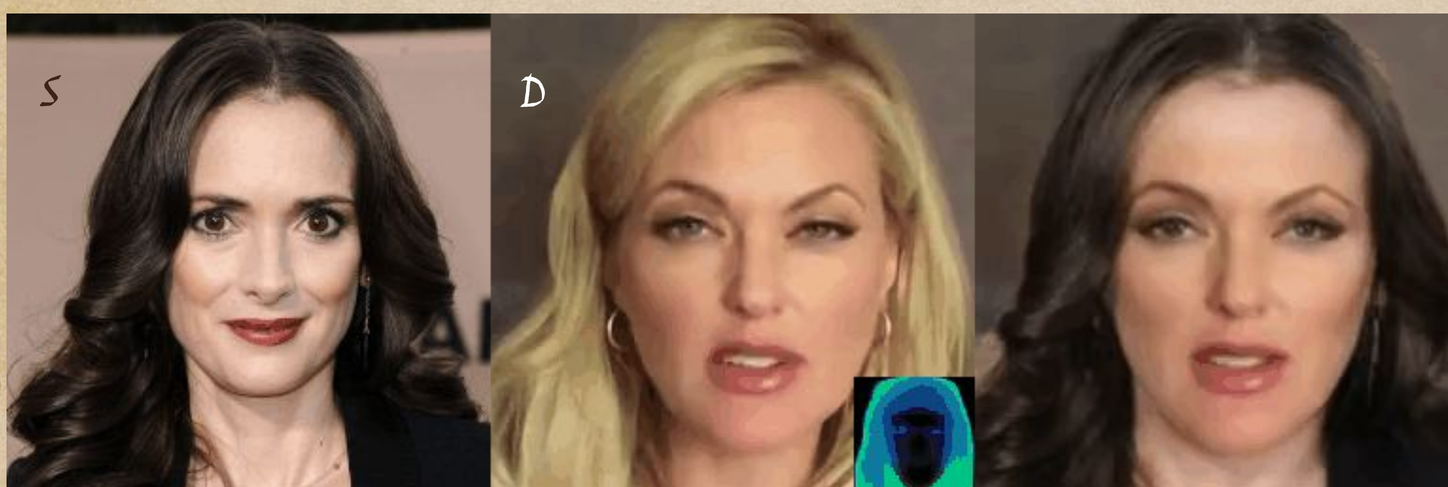
- Inputs
  - videos D having objects of same class w diff appearance
  - source image S
- Output
  - segment from image S transferred to video D

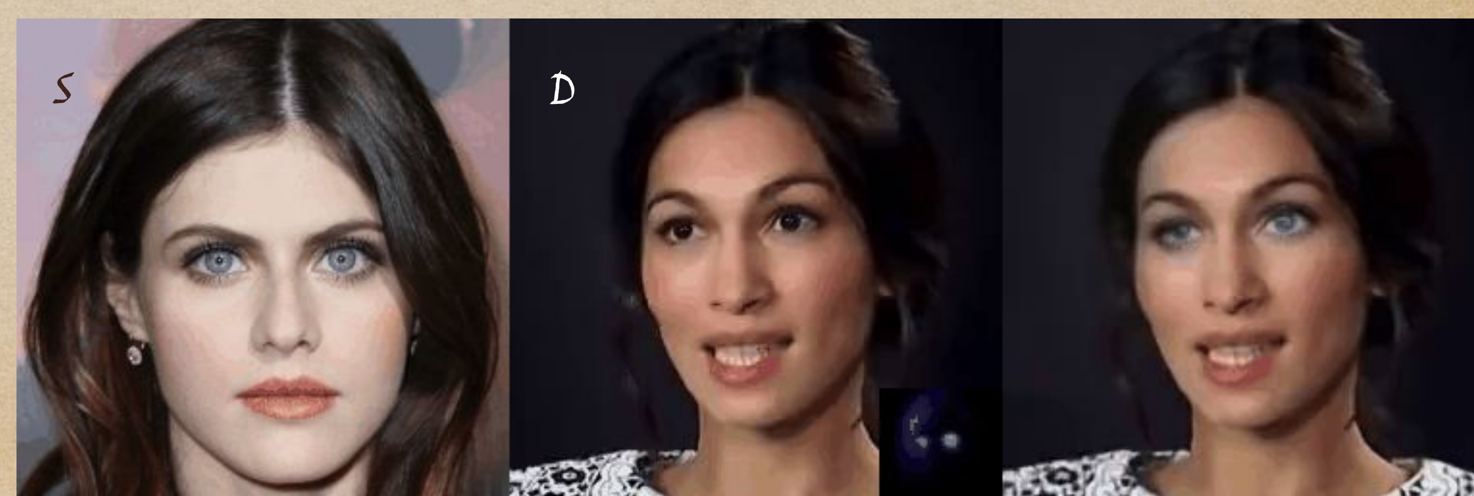




Motion-based CoSegmentation

- Inputs: video D , source image S
- Output:
  - segment from image S transferred to video D





Motion-based CoSegmentation

- Training: 2 random frames from input video D
- Segmentation
  - segment into K+1 parts (K for foreground, 1 background)
  - each segment groups pixels that move together based on a segment-wise optical flow
- Reconstruction
  - reconstruct target frames of D:
    - warp frame based on optical flow from S to D
- Q: how is motion used here?