SEED

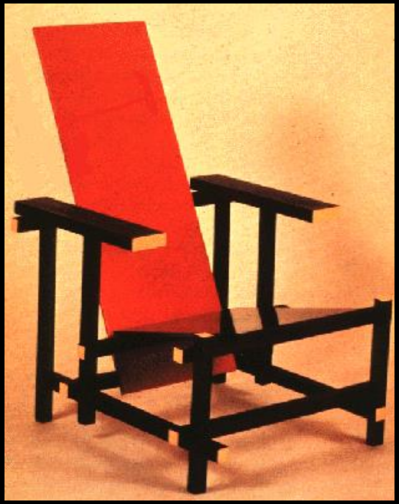
Learning to See in a Dynamic World

ERC Consolidator Grant

Cristian Sminchisescu
Lund University

Challenges: Intra-class variation



Challenges: viewpoint variation

Michelangelo 1475-1564

Challenges: illumination

Challenges: Articulation and Shape of Humans

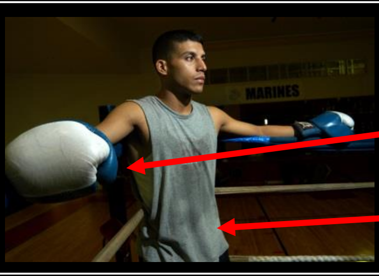
General poses with many d.o.f.

Self-occlusions

Difficult to segment the individual limbs

Different

body sizes



Loss of 3D information in the perspective projection

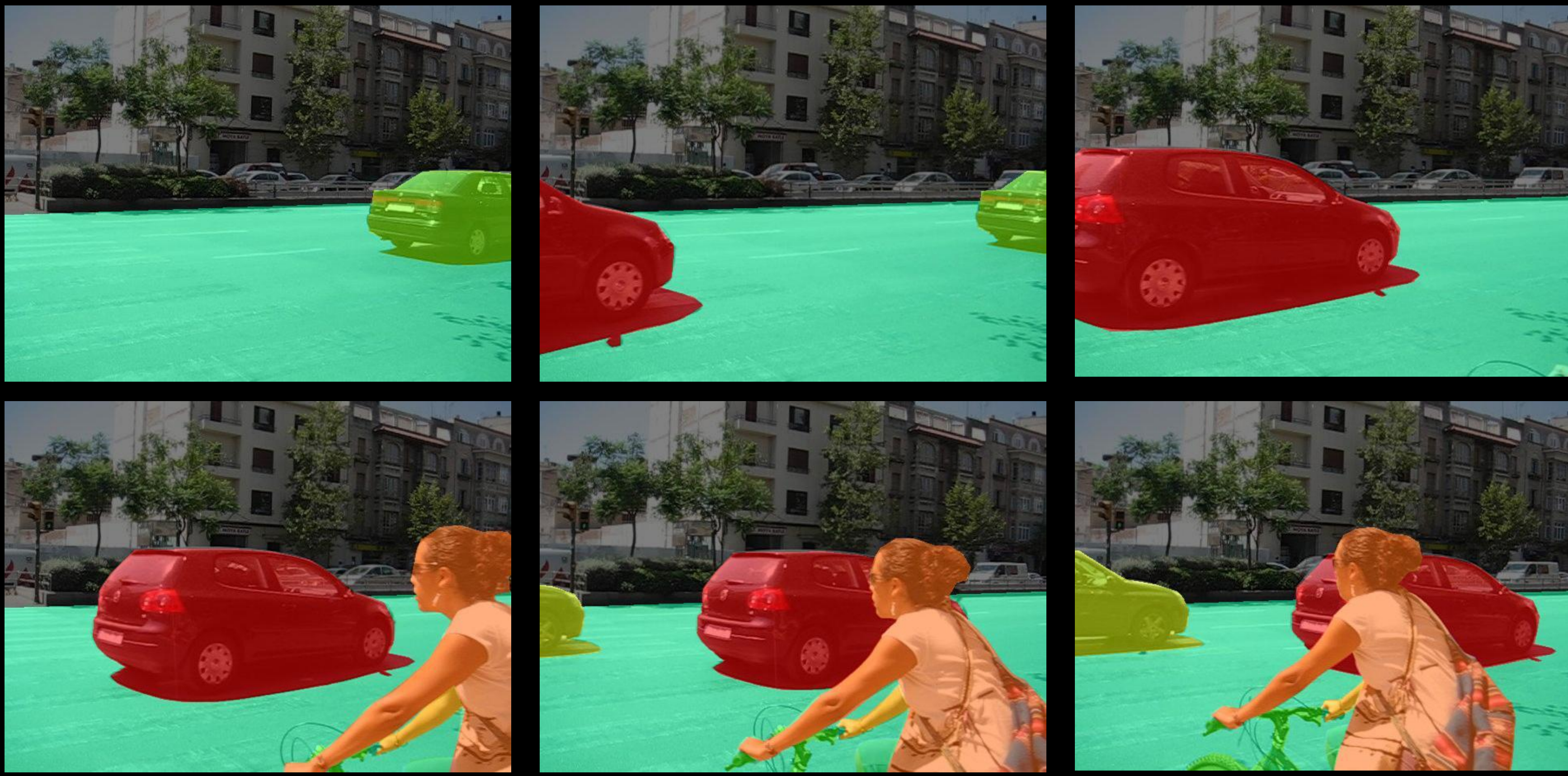
Partial views

Accidental allignments
Motion blur

Several people, occlusions

Reduced observability of body parts due to loose fitting clothing

Dynamic Scenes: *The complexity*



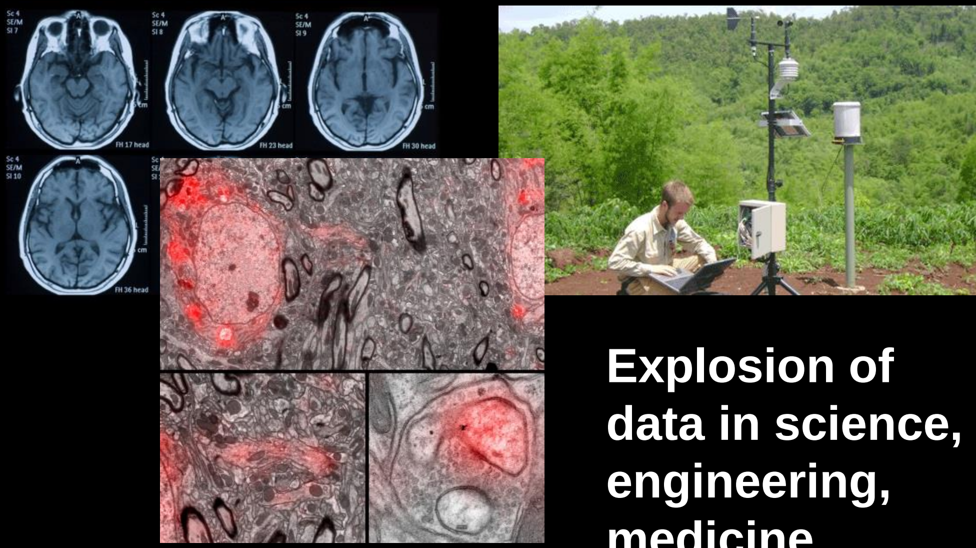
All previous ones shown for images

+

Large inter-frame displacements, occlusions

Why Machine Learning?

Defining how an object or a human looks like in images and video, under a representative set of variations, would be too difficult to specify by hand, but can be characterized by datasets with strong statistical regularity…



Explosion of data in science, engineering, medicine

40% of internet traffic is video

Big Data ≈ Image Data

Real-time sensors

24hrs/day

8 billion

flickr

100,000 hrs/day

YouTube

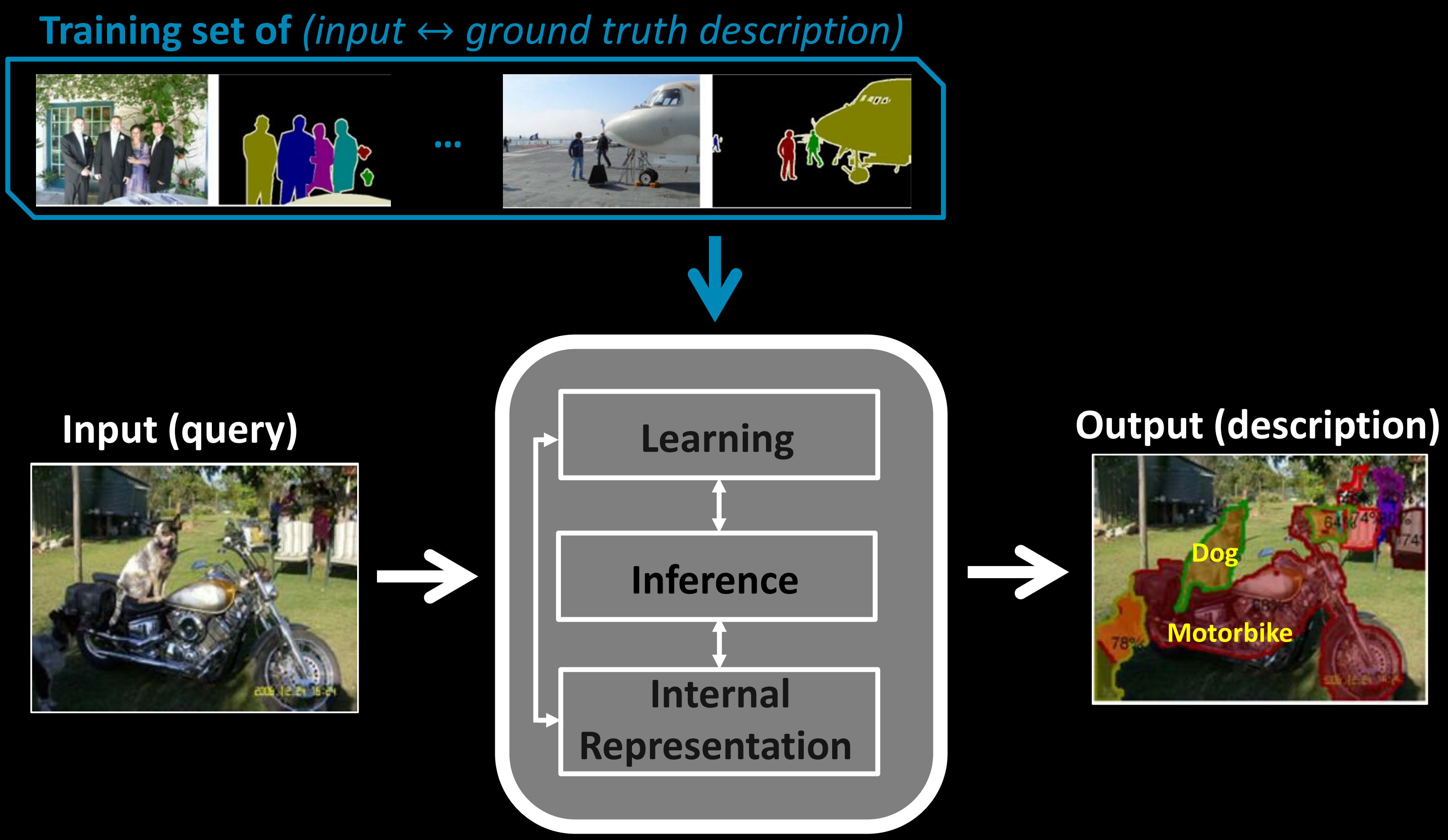
250 billion

facebook

Human (infant) visual learning requires billions of images.
It is (also) a big data problem

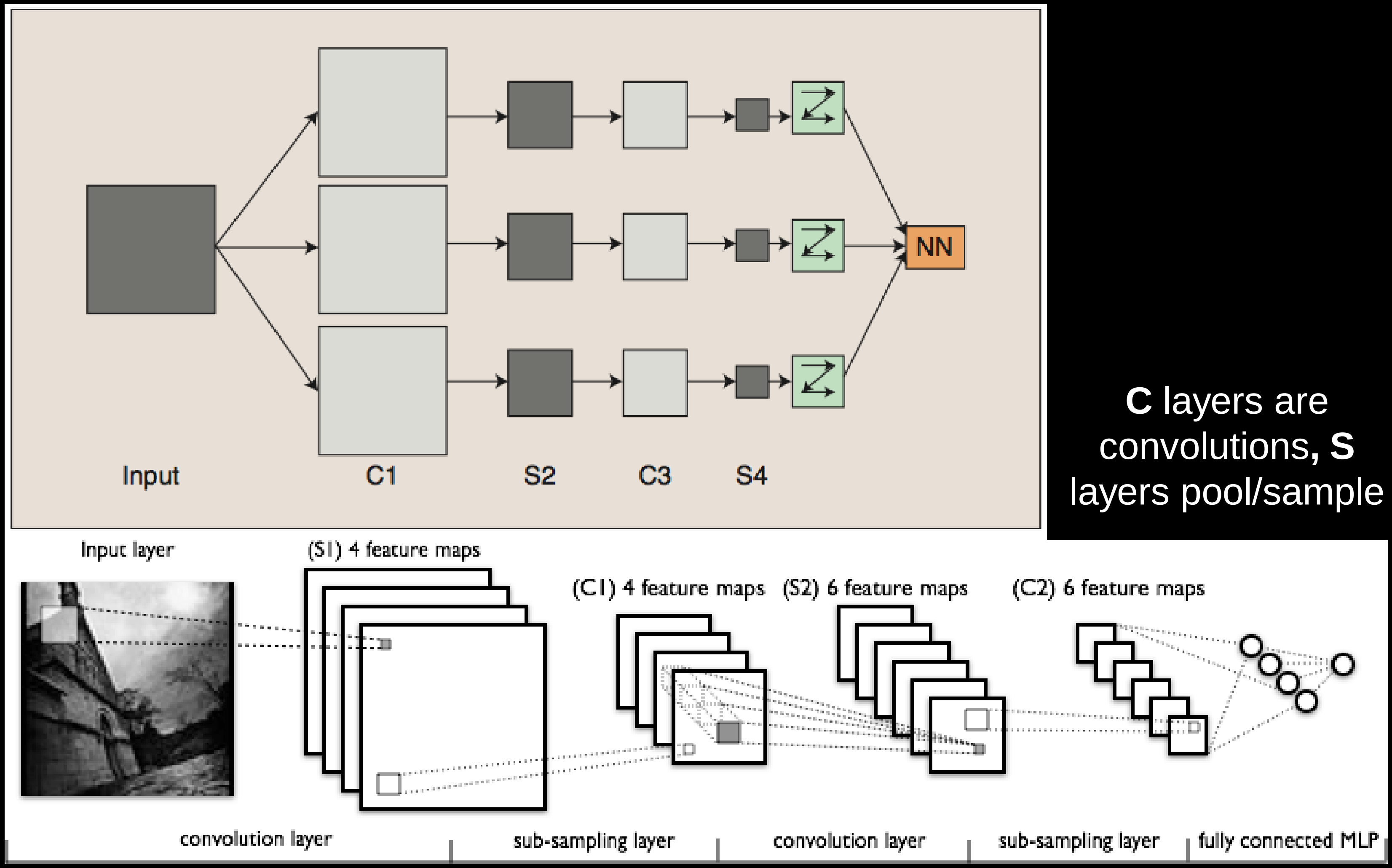
A robot may require just as much data in order to competently learn to see and interact with the visual world

Computer Vision Modeling



Learning Visual Representations

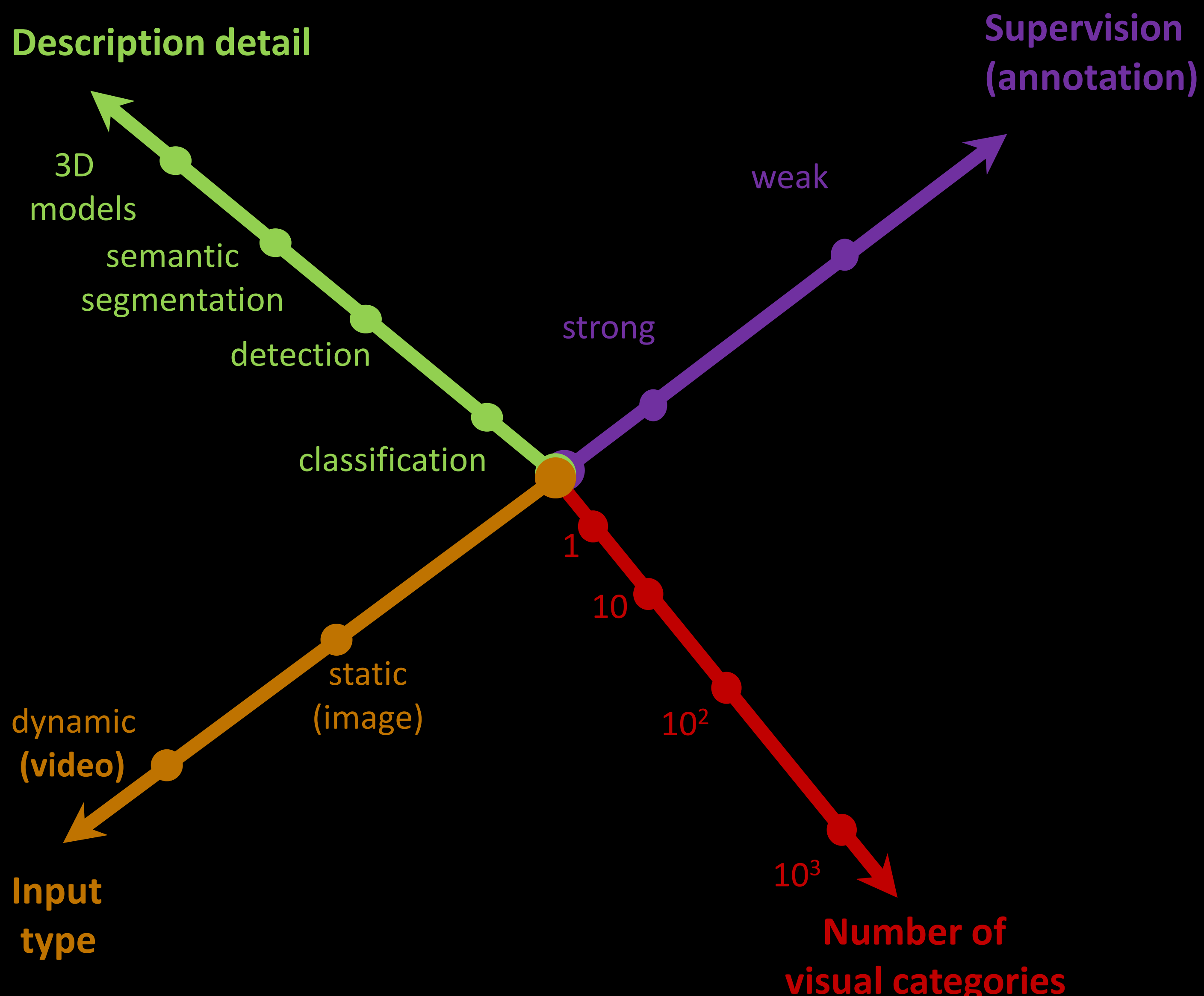
Convolutional Neural Networks



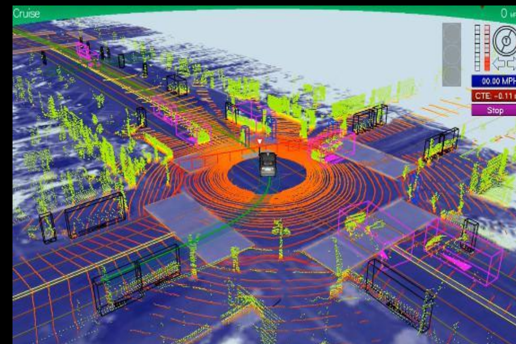
C layers are convolutions**, S** layers pool/sample

Lecun et al. 1998, Hinton et al., 2015

Scientific Challenges



Research Emphasis


Computers that `learn to see' in a dynamic world

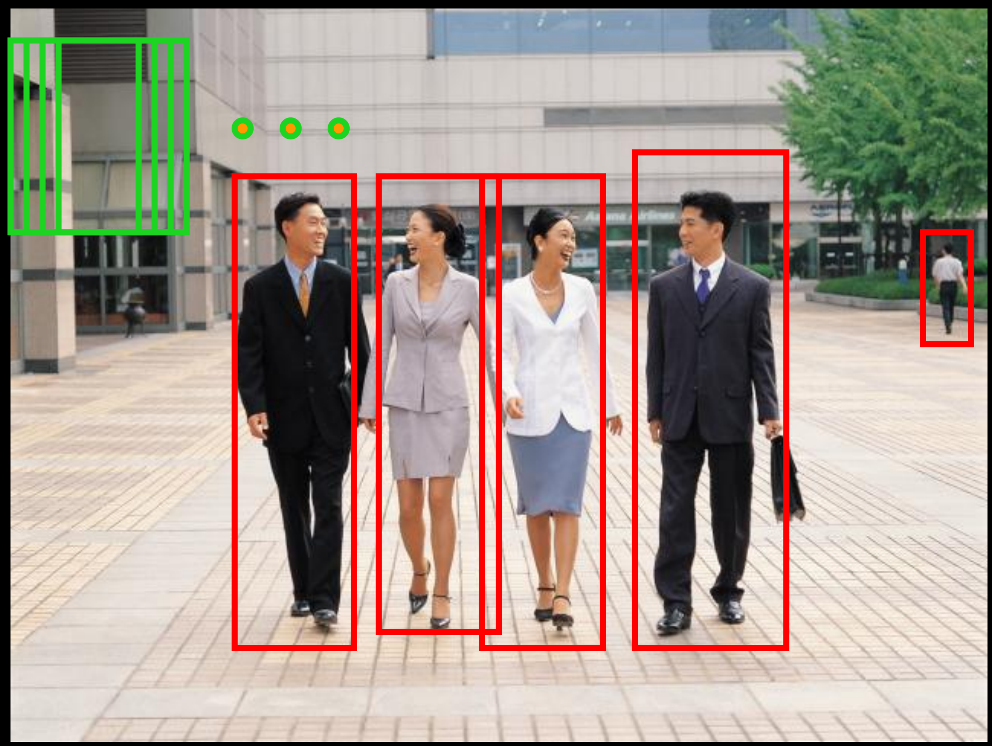
Static (image) → **Dynamic** *(video)*

Coarse analysis (holistic) → **Precise** description *(shape, 3D)*

Rigid (preset supervision) → **Flexible** *(continuous learning)*

Recognition by Detection

Is this an X?

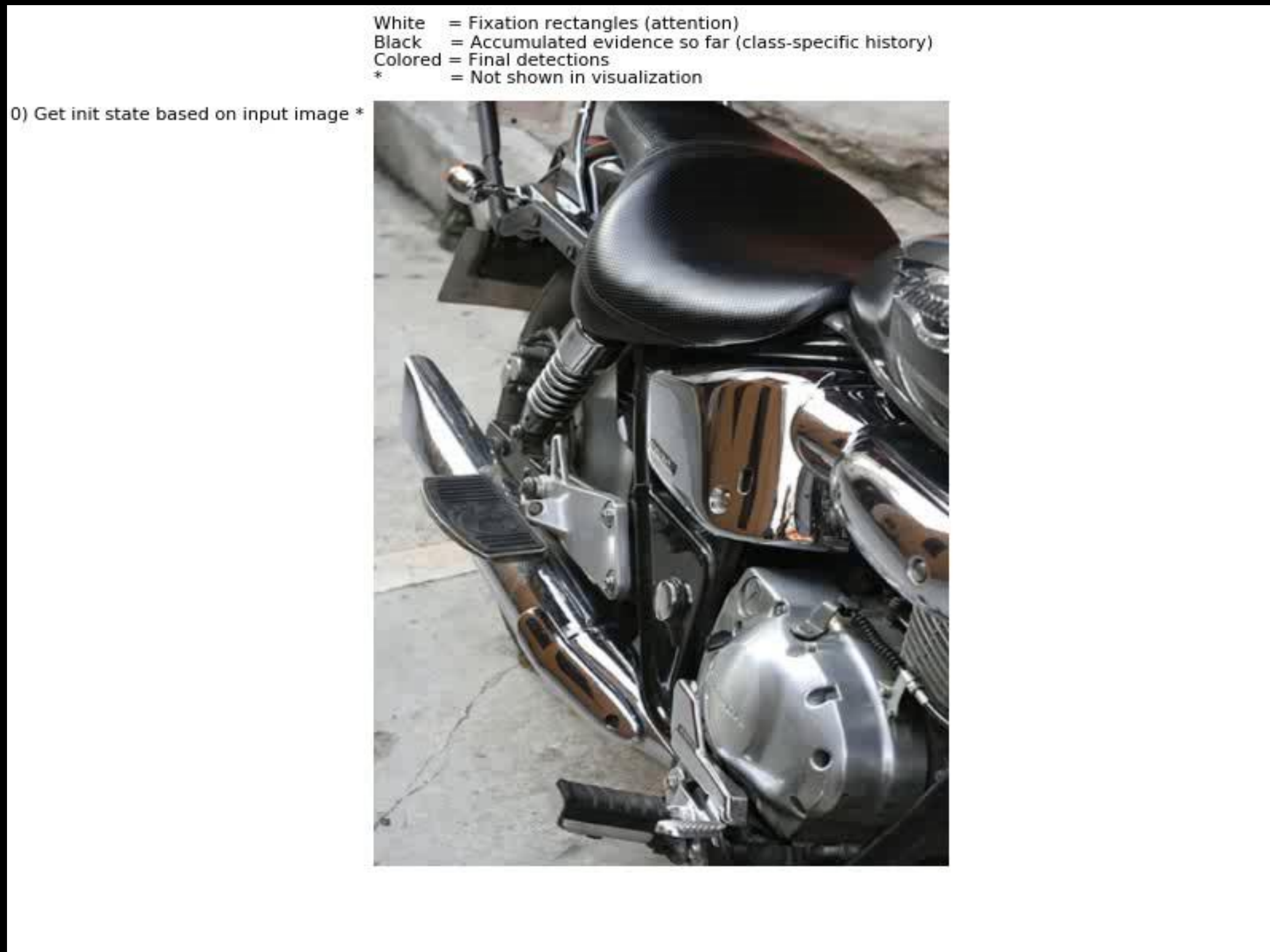


Search at multiple locations, scales and for all object categories of interest
Indiscriminately describe (mix) both foreground and background

Rowley, Baluja & Kanade 1996 (face detection)

Recognition in the Human Eye

Active Visual Recognition



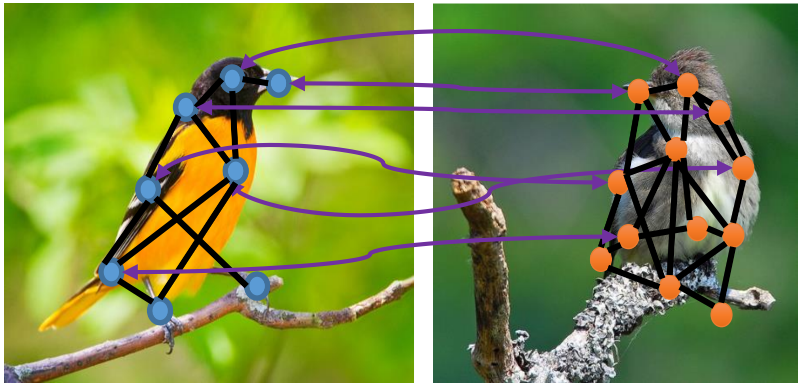
Deep reinforcement learning for thw
search strategy + detector + stopping criteria

Pirinen and Sminchisescu, CVPR 2018

ESTABLISHING CORRESPONDENCES GRAPH MATCHING

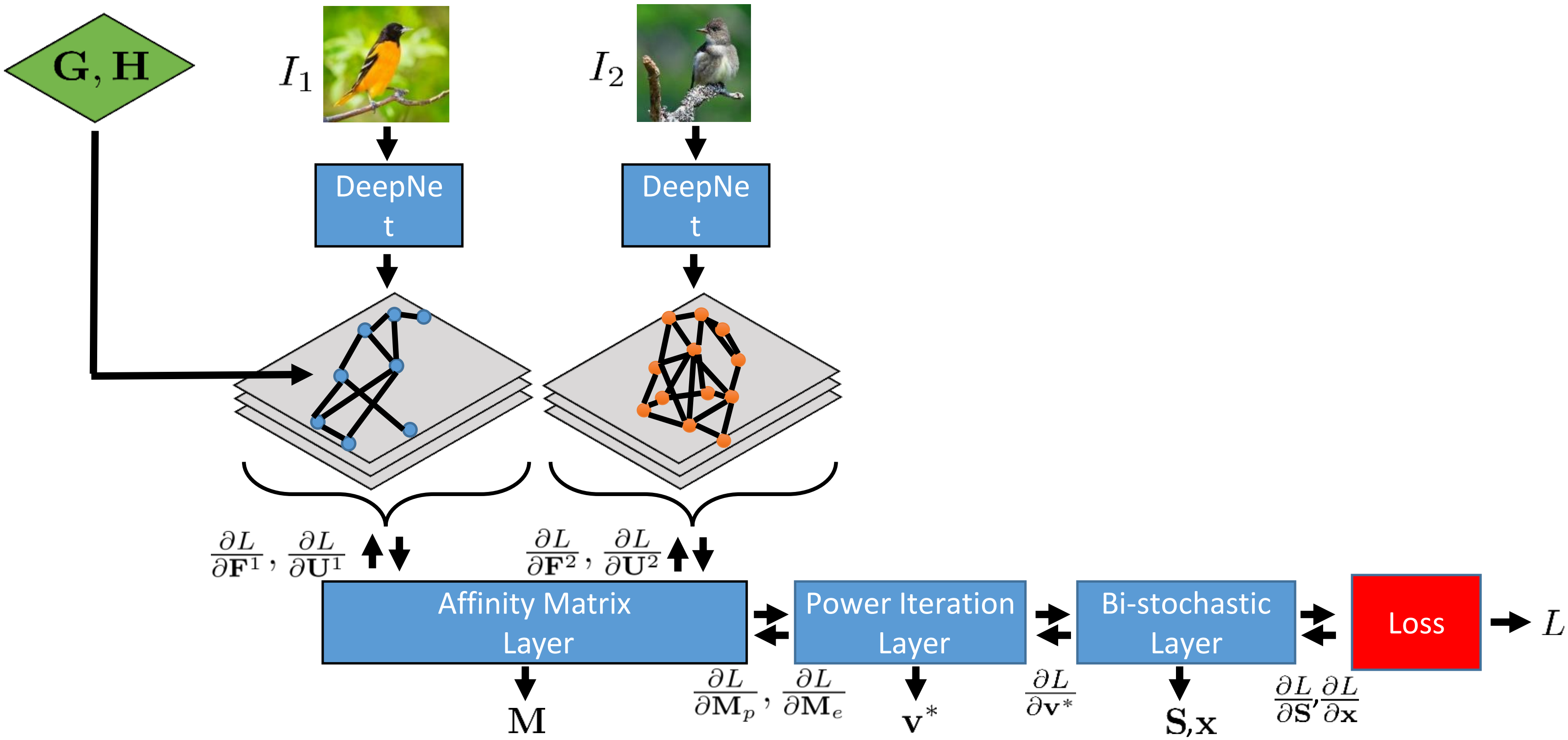
Input: two graphs $G_1 = (V_1, E_1)$ **and** $G_2 = (V_2, E_2)$

with $|V_1| = n, |V_2| = m, |E_1| = p, |E_2| = q$



Task: find a one-to-one mapping between the two graphs that accounts for structure, i.e. reflects both node and edge similarities

TRAINABLE GRAPH MATCHING NETWORK



Matrix backpropagation: generalization of backprop to matrix functions and global calculations like SVD, EIG, projectors (Ionescu, Vanzos, Sminchisescu, ICCV 2015)

Zanfir and Sminchisescu, best paper award honorable mention, CVPR 2018

Trainable either on **different images of the same video** *or* **same visual category**

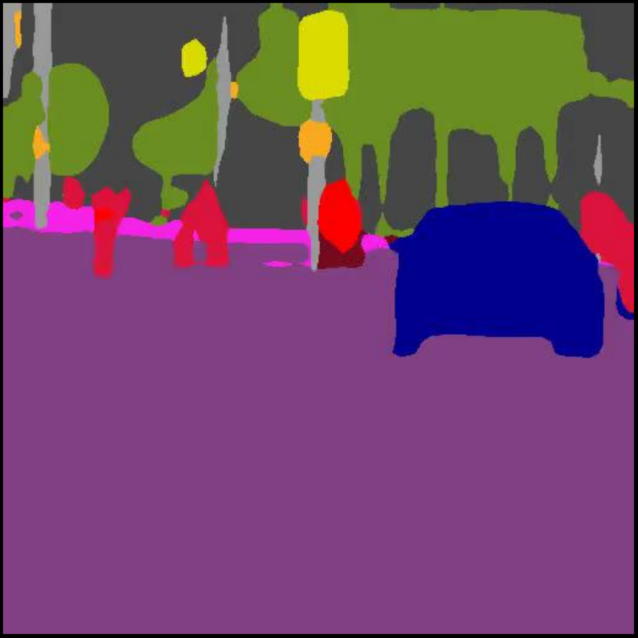
MPI-Sintel test partition exhibits large motions and occlusion areas

From top to bottom: source images with the initial grid of points overlaid and target images with corresponding matches. Colors are unique and encode correspondences. Even for fast moving objects, points tend to track the surface correctly, without sliding – see the dragon's wing, claw, and the flying monster

CUB: Green frames indicate ground truth correspondences, red frames estimates

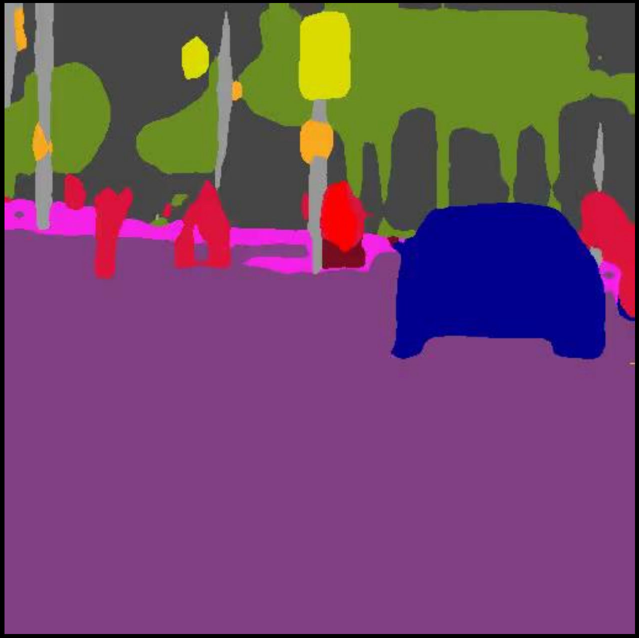
Other correspondence results on arbitrary images

Weakly-supervised
Semantic Video Segmentation



Per Frame

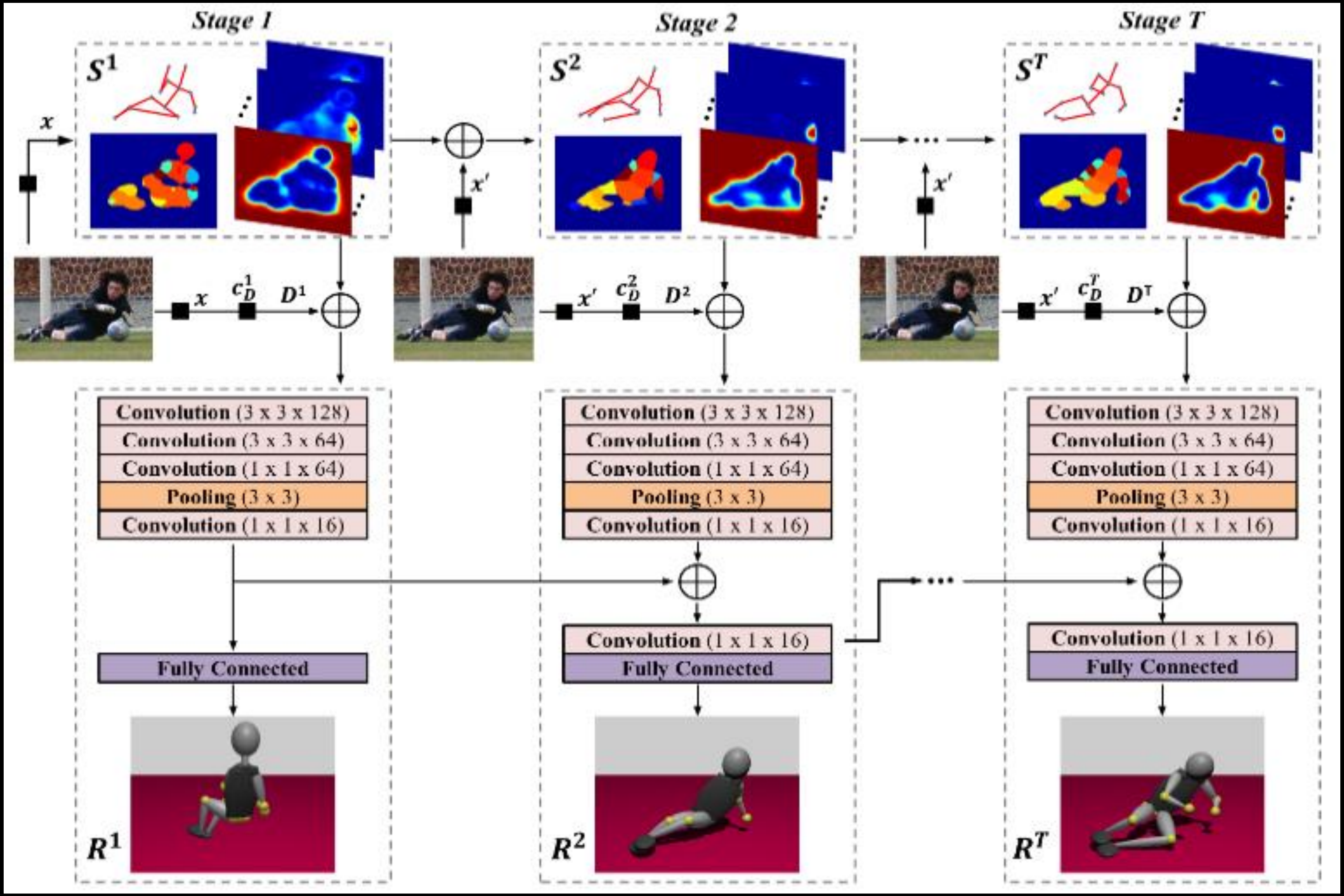
Original Video



Proposed (GRFP)

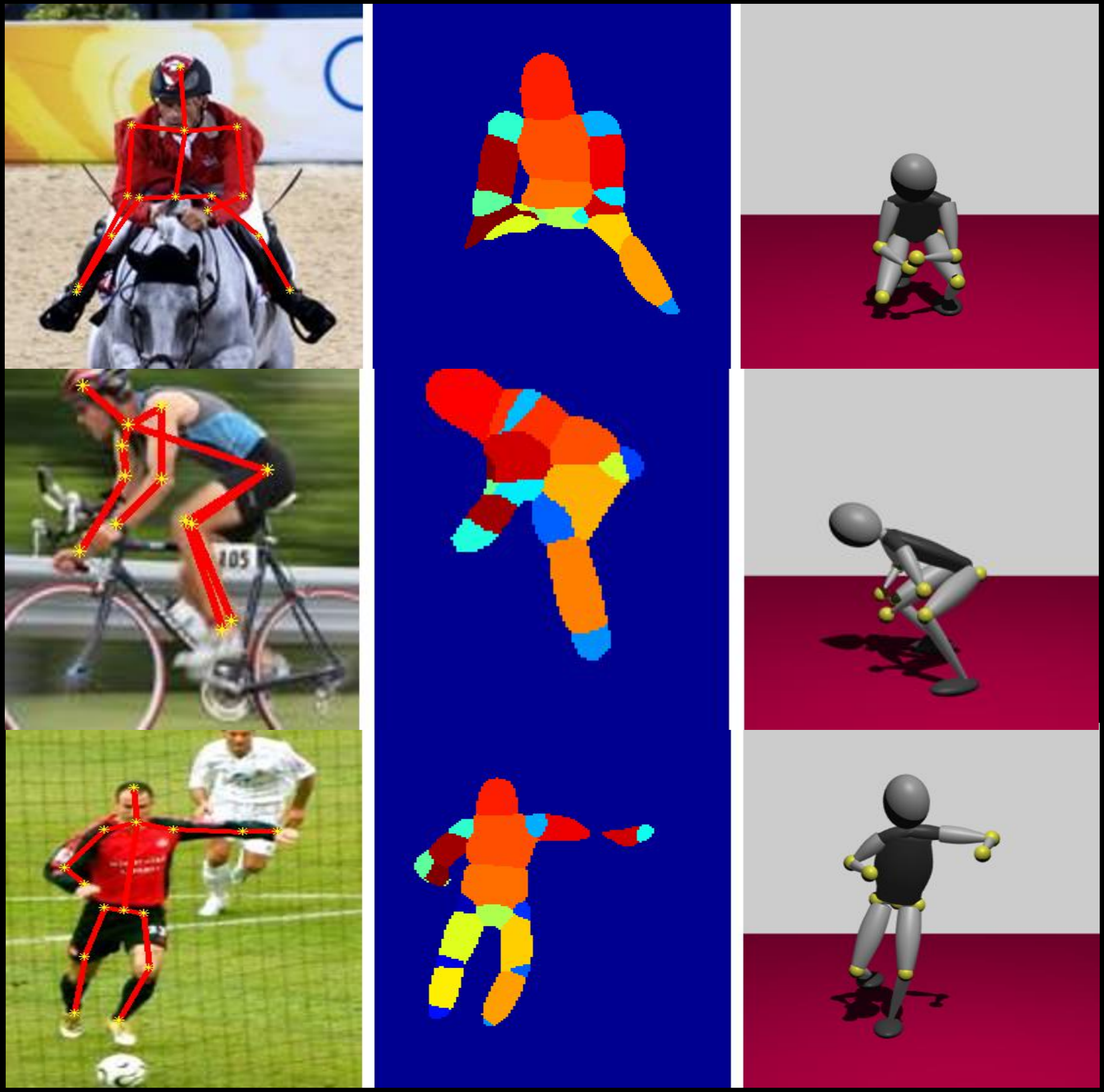
Nilsson, Sminchisescu, CVPR 2018

Multitask Sensing Architecture

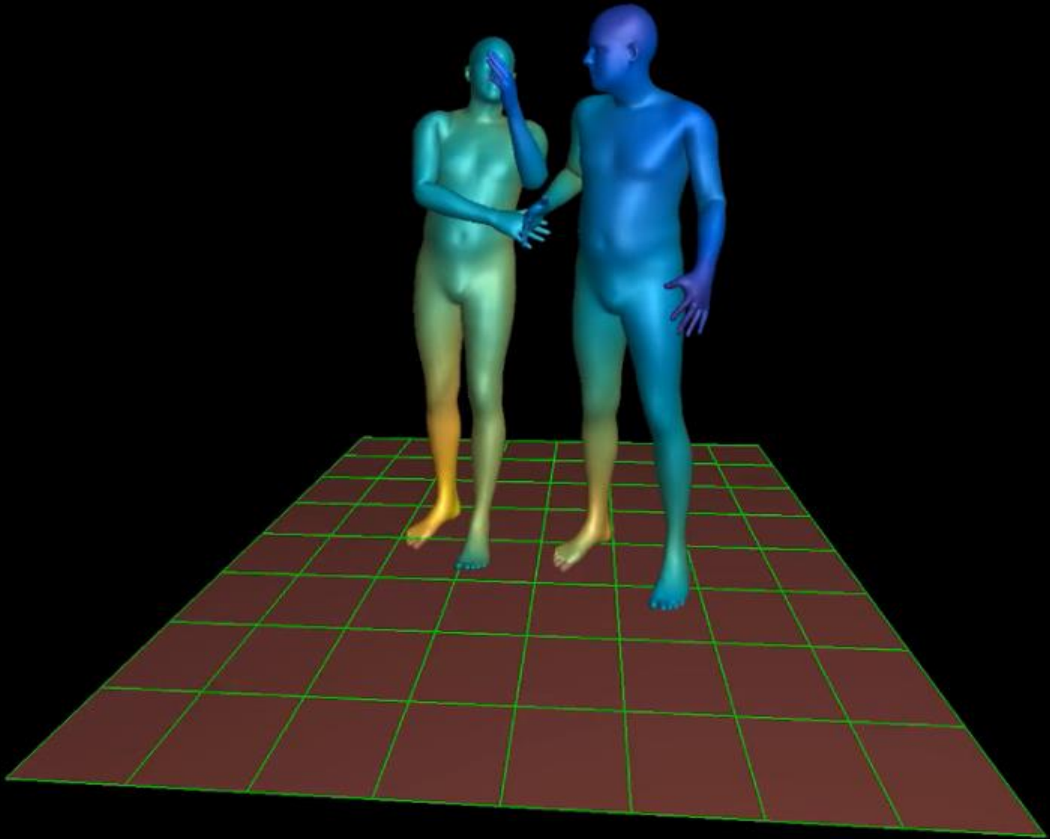


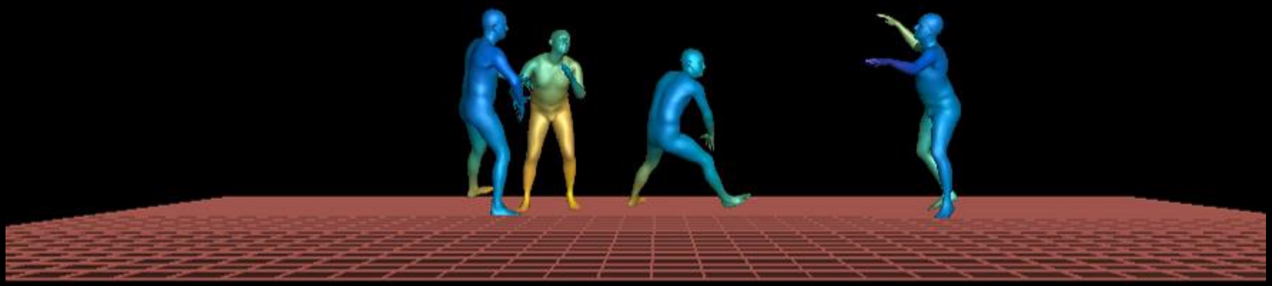
Popa, Zanfir, Sminchisescu, CVPR 2017

Semanic Segmentation + 3D



Multiple People, 3D Pose and Shape





Autism Treatment Automation

Camera placed behind robot to avoid interference with therapy

Field of view must avoid robot

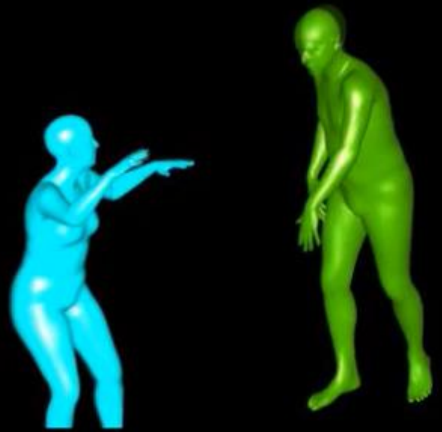
Chairs of child and therapist need to be close to table to use cards

Child and therapist out of field of view

Interaction results in occlusions

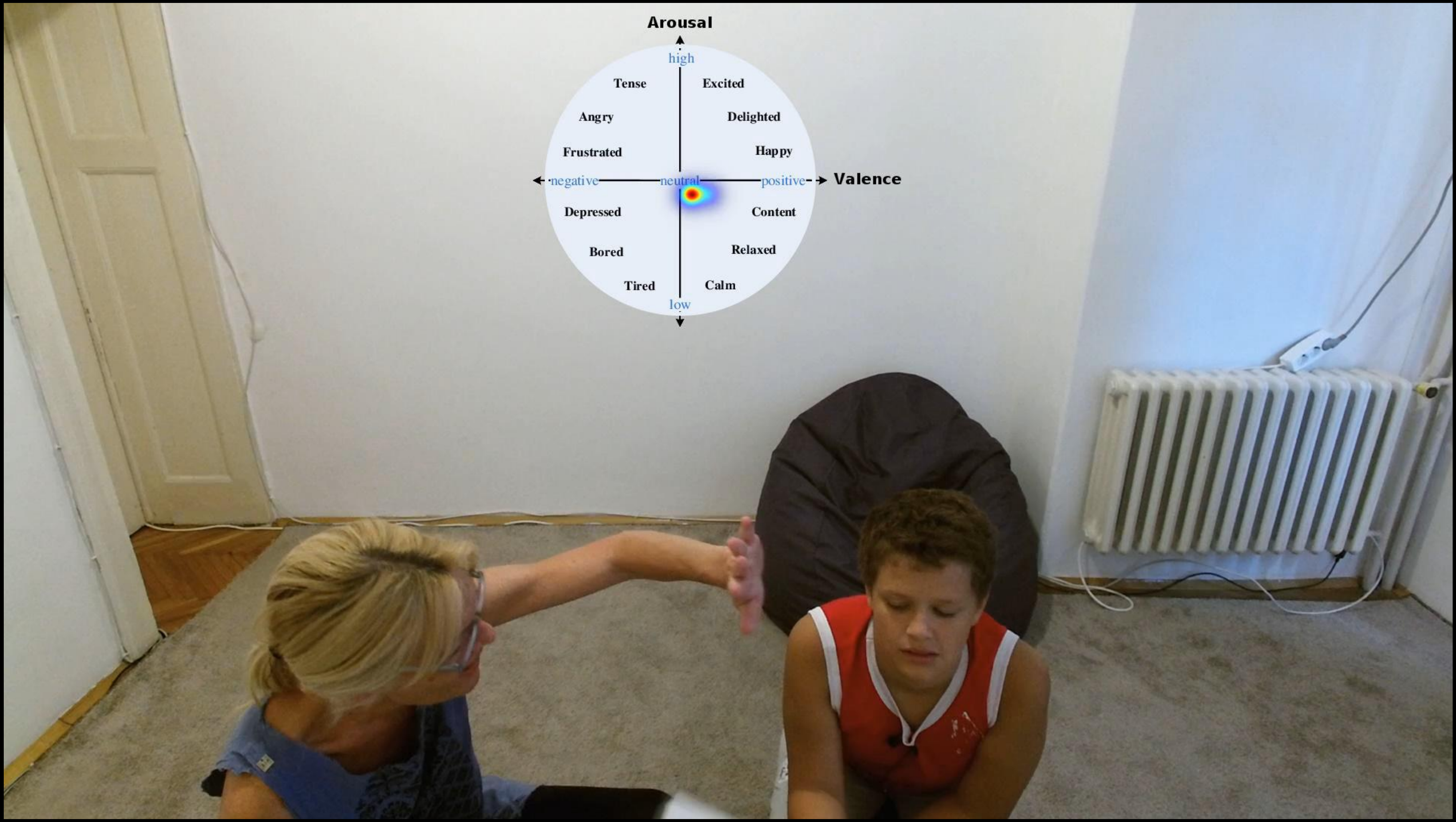
EU project DE-ENIGMA

Reconstructing Human Interactions



Marinoiu, Zanfir, Olaru, Sminchisescu, CVPR 2018

Behavior: Valence Arousal



Marinoiu, Zanfir, Olaru, Sminchisescu, CVPR 2018

AI Aspects

- Societal
  - Adapted new degree [programs and continuous instruction, well-balanced regulatory framework
- Scientific enablers
  - Learning theory (dynamics and nonlinear systems), weak supervision, perception and action in non-trivial environments
- How can pitfalls be avoided?
  - Critical understanding of potential and limitations/evolution. Adapt and communicate, adequate data collection (unbiasedness), privacy
- Multidisciplinary aspects
  - Huge potential but domain knowledge is key
- Time perspective
  - Active/robotics, personalization

Thank you!

Human Appearance Transfer

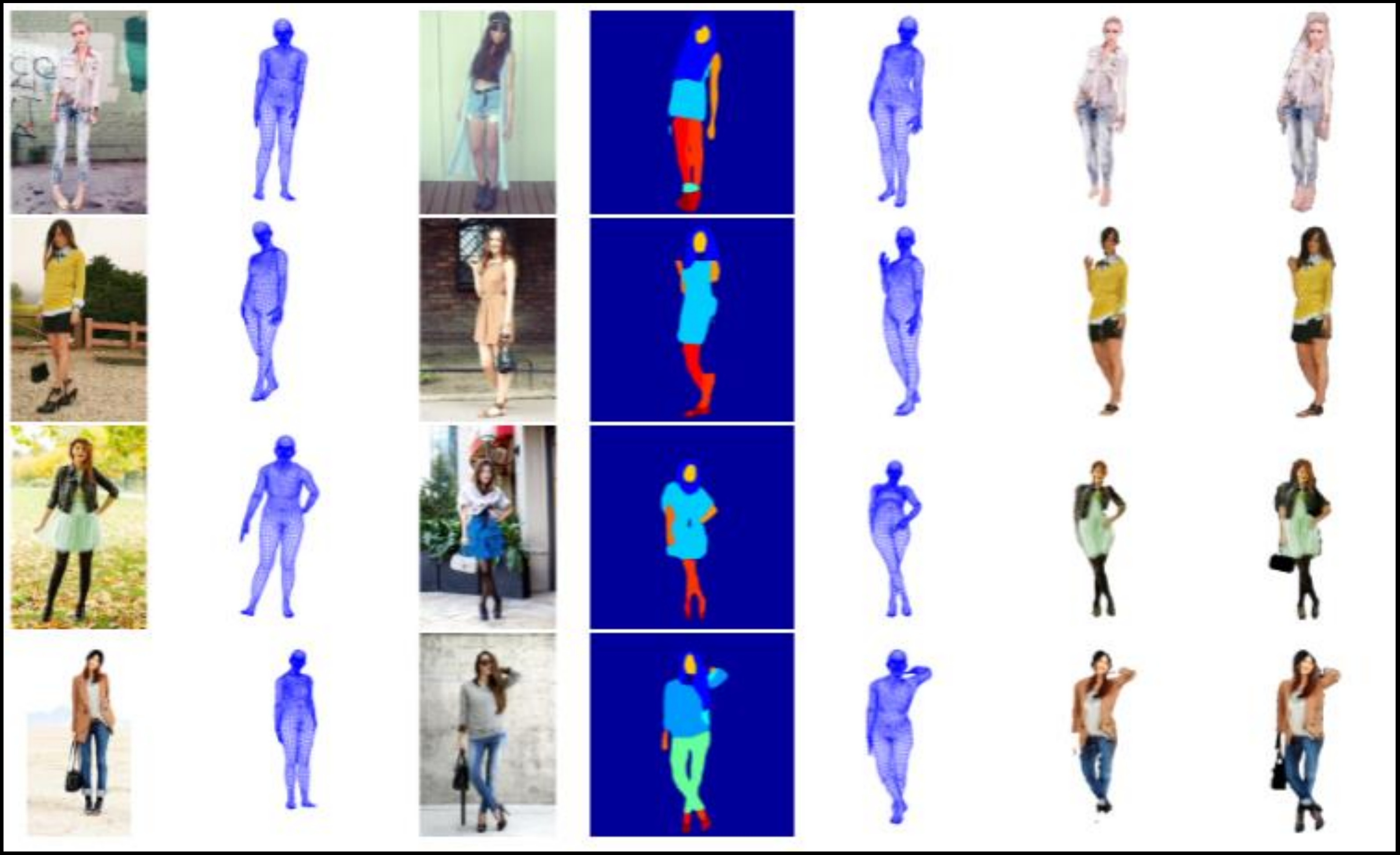
Given two people differently dressed and in different poses, swap appearance

Human Appearance Transfer

Dress a person with clothing from another

Popa, Zanfir, Sminchisescu, CVPR18

Human Appearance Transfer



Dress Like a Celebrity

Celebrity

Regular person

Regular person dressed as celebrity

Popa, Zanfir, Sminchisescu, CVPR18

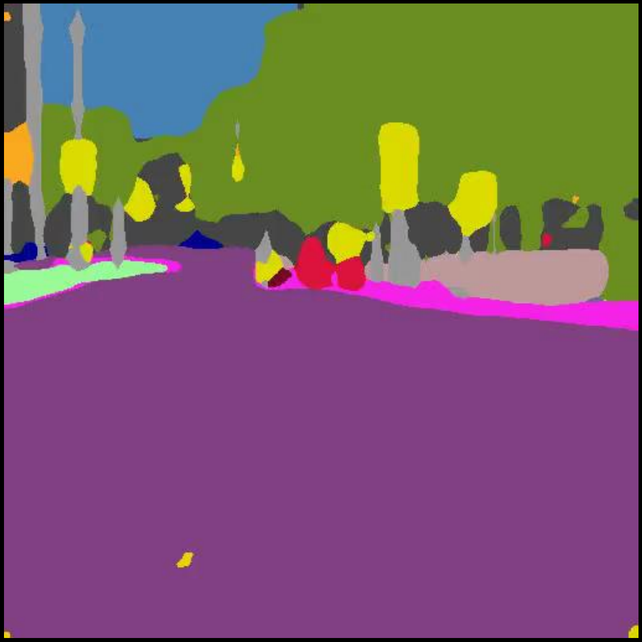
Impact

- ERC team of 4 members established (part of a larger research group of 10 people)
- 10 publications at CVPR, NIPS, AISTATS
  - 1 best paper award honorable mention (CVPR18, highest impact conference in AI)
- ECCV 2018 organization (Program Chair)
- Presentations at ETHZ, INRIA, Stanford, Heidelberg, Max Planck, Edinburgh, etc.

Semantic Video Segmentation (II)

Per Frame

Original Video



Proposed (GRFP)

3D Human Training Data

- Labeled 2d and 3d human training data difficult to acquire
- Accurate 2d can be captured and labeled by hand (e.g. body joints) – not really ground truth but good enough
- Motion capture synchronized 2d-3d is extremely accurate but backgrounds/clothing not as diverse
- Mixed reality training dates back at least to 2001 but realism of both body and geometric scene is essential

Do we need fully labeled data… or can work with whatever available within a multitask architecture?