AY 2024/2025

EE6222 Machine Vision

Video (Part II)

Cheng Jun

Institute for Infocomm Research, A*STAR

Address: 1 Fusionopolis Way, 138632

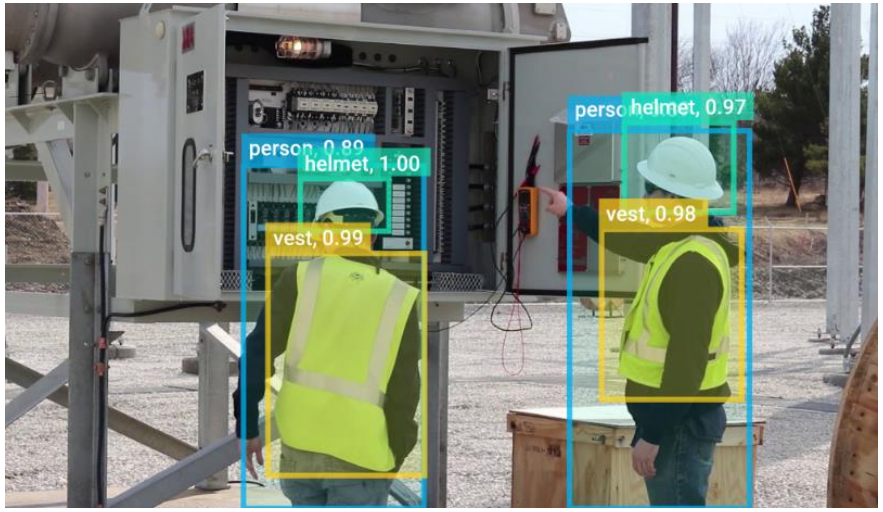
Email: cheng_jun@i2r.a-star.edu.sg; cheng.jun@ntu.edu.sg

Major Topics

- Object Detection & Tracking
- Action Recognition
- Video Event/Anomaly Detection
- Video Enhancement
- Optical Flow

Video Analysis Applications

Counting people in shopping mall



Inspection in power and utility industry

Fall detection

Action recognition: stretching leg

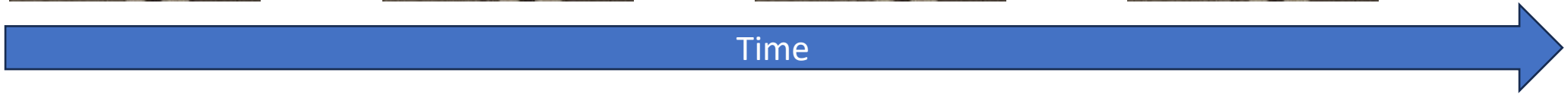

Male
Age:27.56
Anger
Male
Age:23.90
Neutral
Dist:3.28
Dist:3.39
Green: Tree
Purple: Road
Gray: Building
Raw Video Source: goo.gl/LqyMQM

Video

What we see:

What a computer sees:

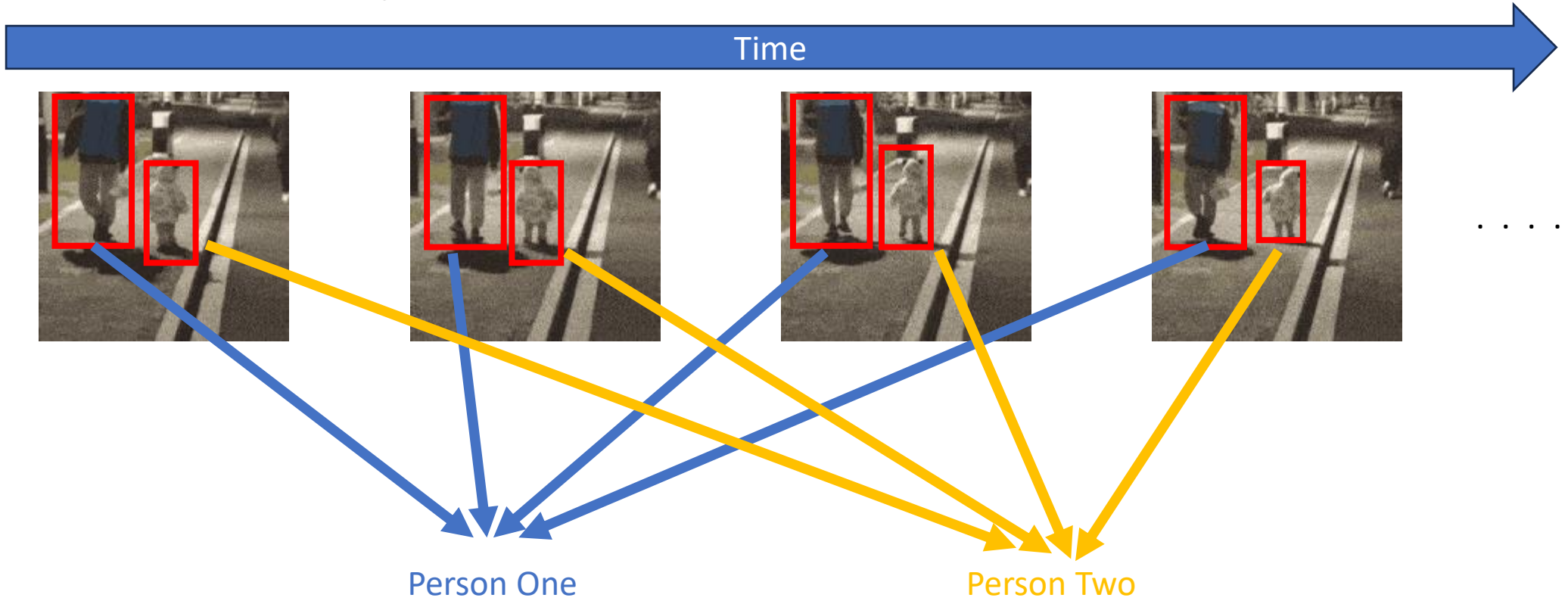
. . . .



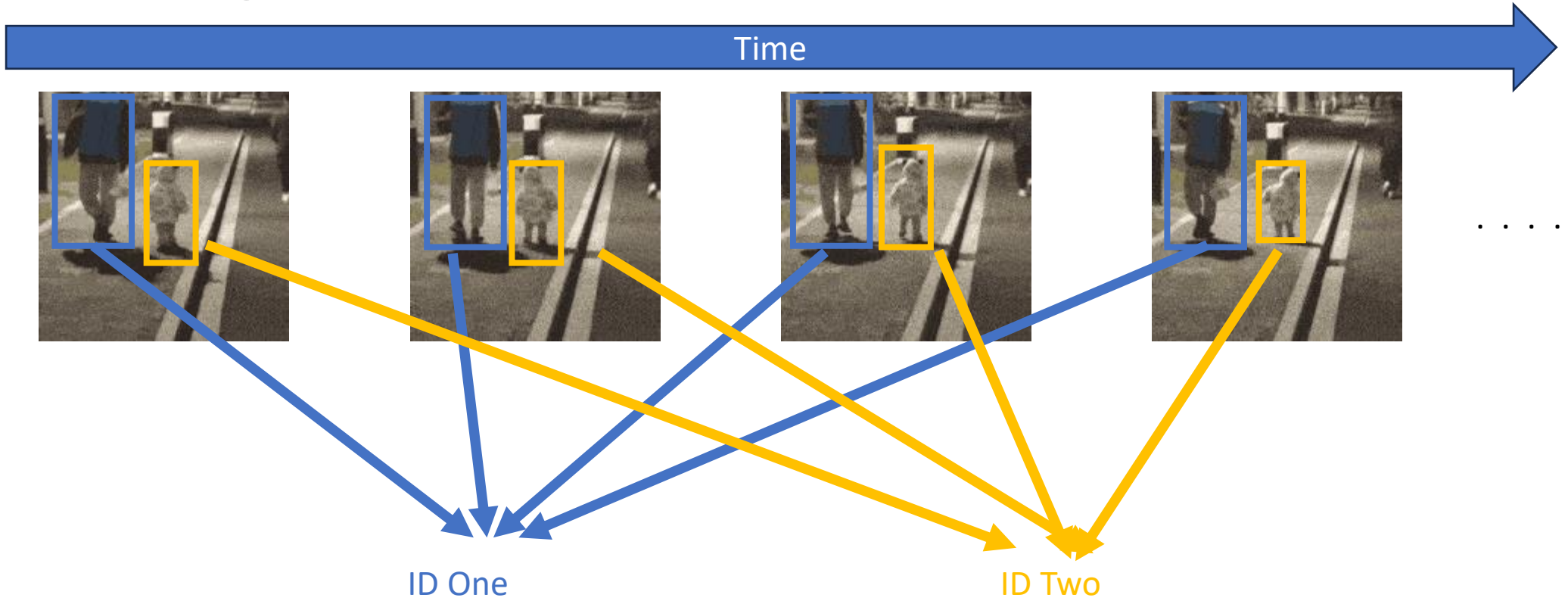
Video is a sequence of images with correlations among the images.

Detection and Tracking

How a computer understands the video



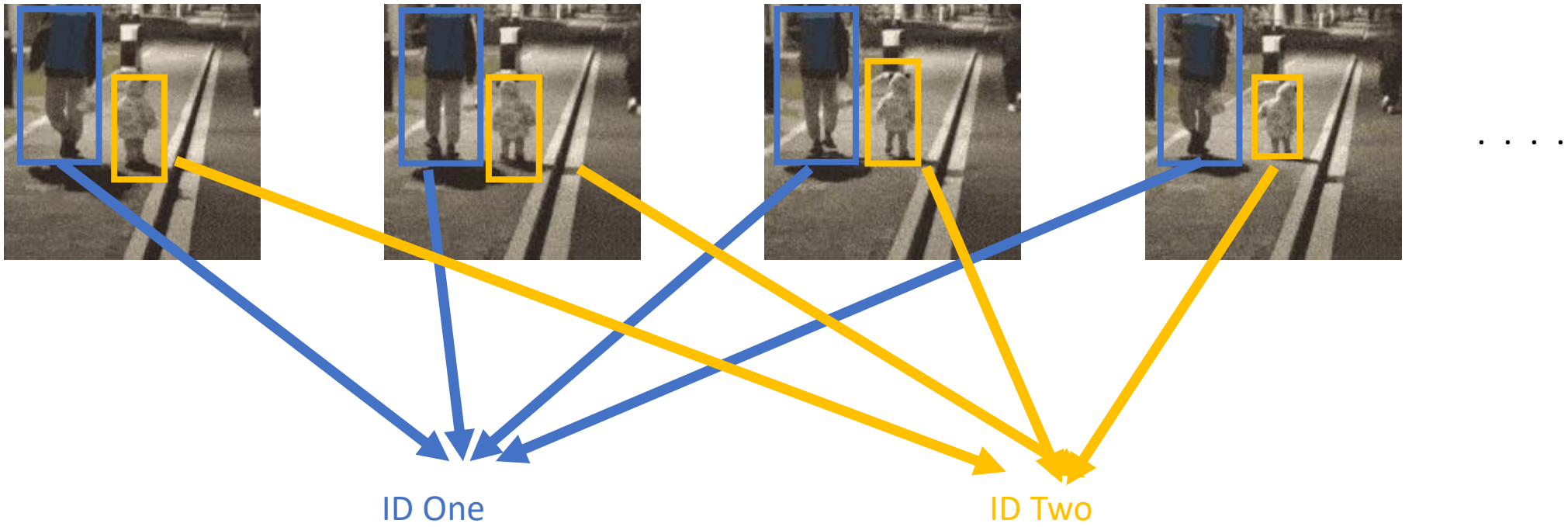
Tracking



Object Detection & Tracking

Typical way:

1. Detect/segment objects
2. Associate detections over time




Male
Age:27.56
Anger
Male
Age:23.90
Neutral
Dist:3.28
Dist:3.39
Green: Tree
Purple: Road
Gray: Building
Raw Video Source: goo.gl/LqyMQM

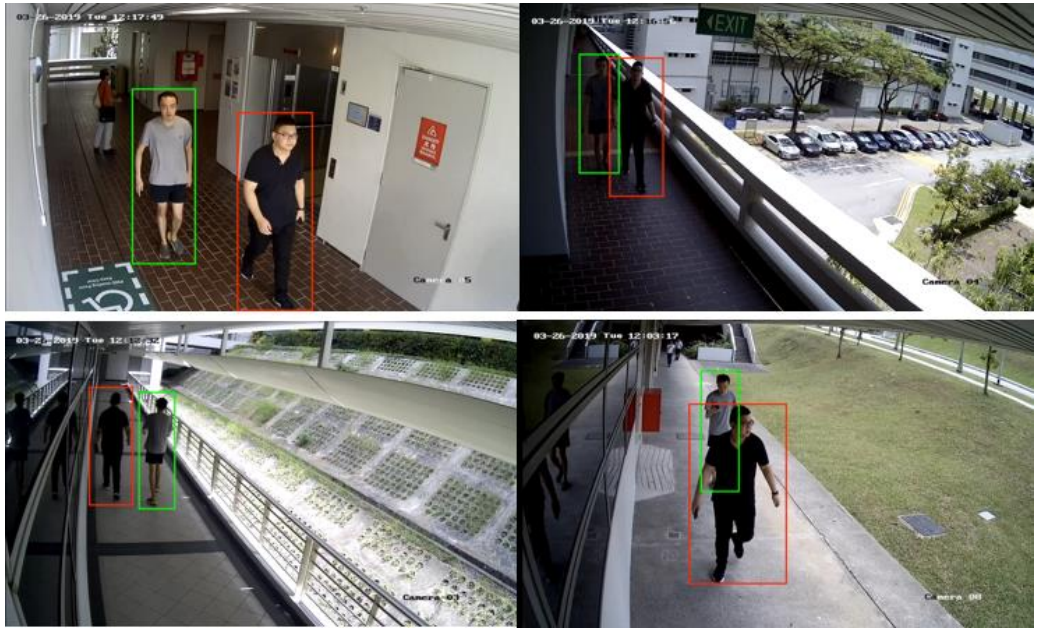
Main Difference between Video and Image

- Temporal Continuity

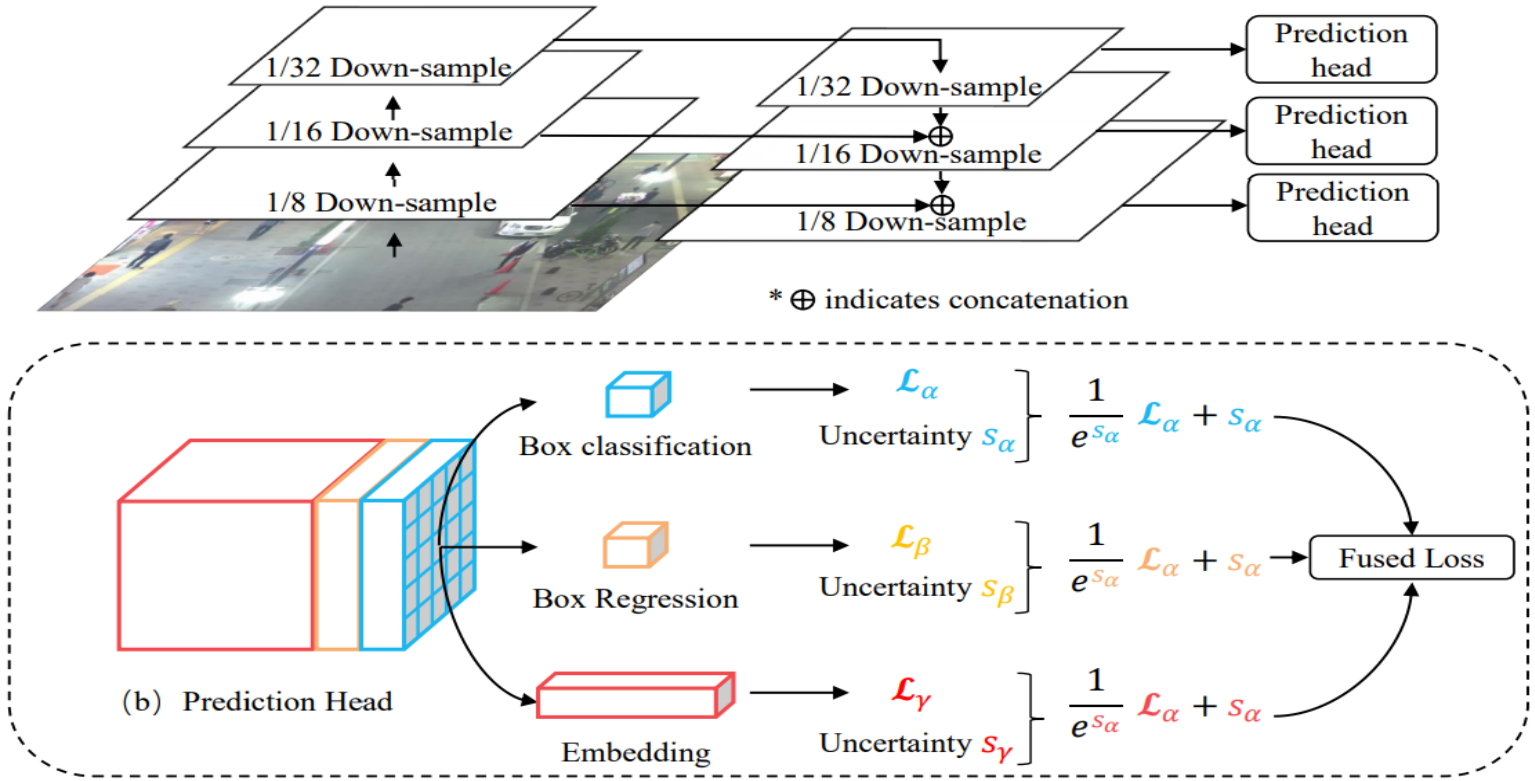
1. Image object detection: each image is independently processed
2. Video object detection: there is a continuity across time. Temporal information can be used to improve detection, tracking etc.

Re-ID

- Re-identification (Re-ID for short) is defined as the problem of matching object/people across disjoint camera views in a multi-camera system.
- It is often achieved by detection and tracking



Joint Detection and Embedding (JDE)



ReID vs. JDE

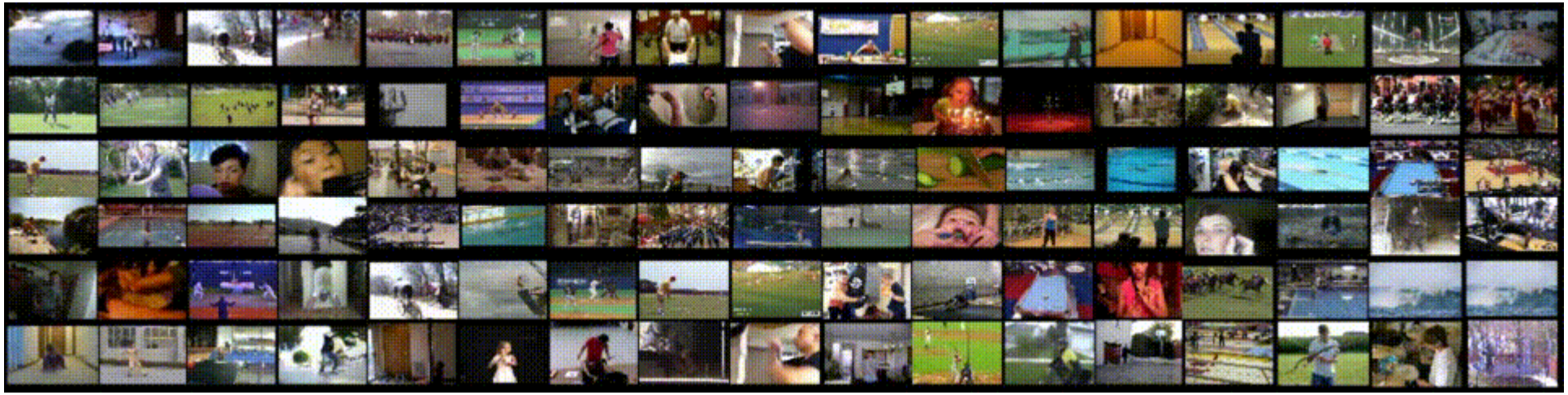
- Dji ReID

FairMOT JDE

What are the differences?

Video(Action, Event) Recognition/Detection

http://www.thumos.info/home.html



Video Events

Basic event or action detection

push

pushup

ride bike

ride horse

run

shake hands

shoot ball

Complex or high-level event detection

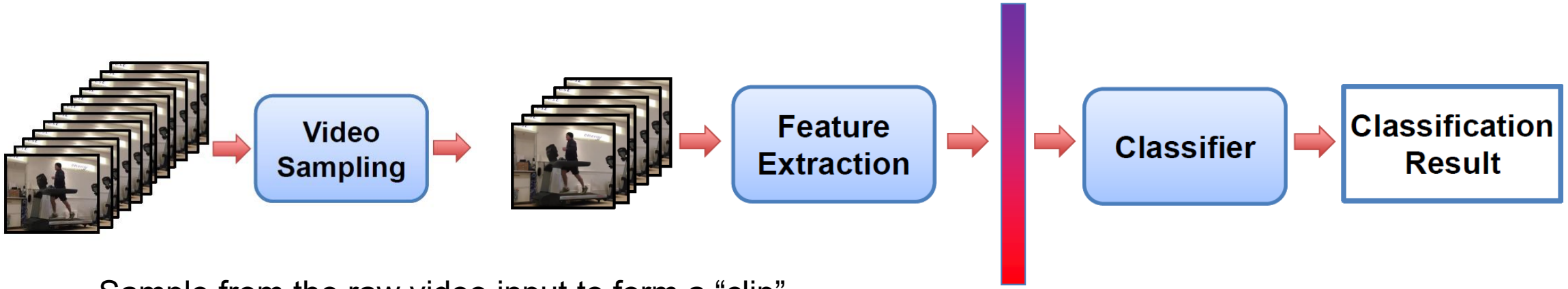
Wedding Ceremony

Video Event

How to detect “Goal”?

Action Recognition

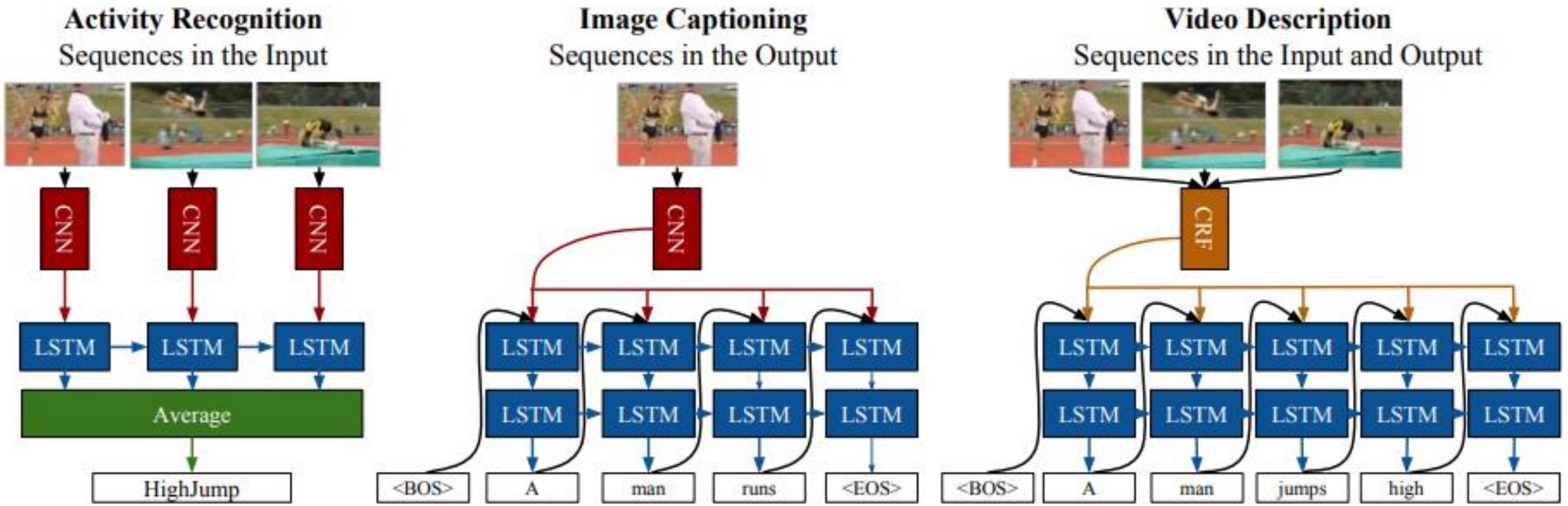
- Action Recognition is a computer vision task that involves recognizing human actions in videos or images.



Sample from the raw video input to form a “clip”.
Feature extraction from video clip for recognition.

Action recognition is similar to object recognition.

Long-term Recurrent Convolutional Network

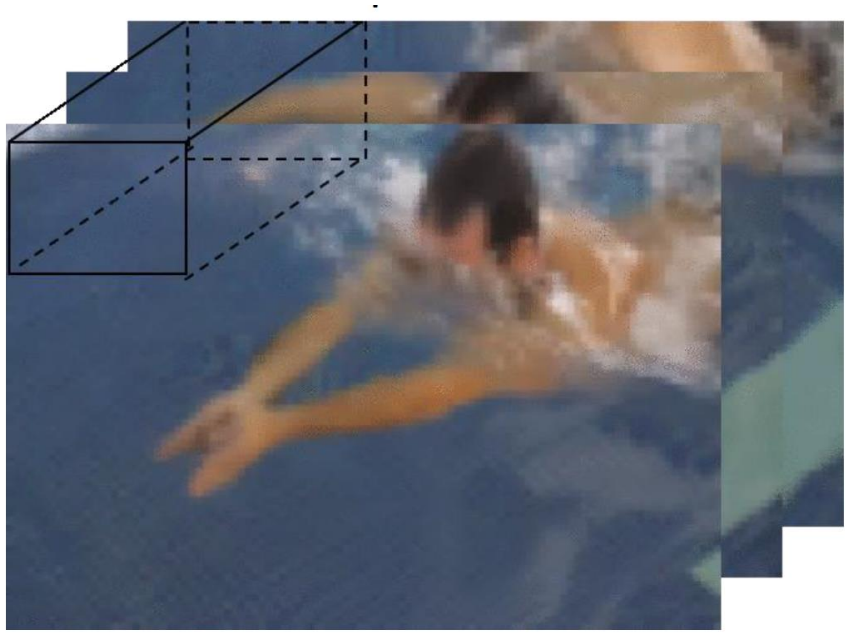


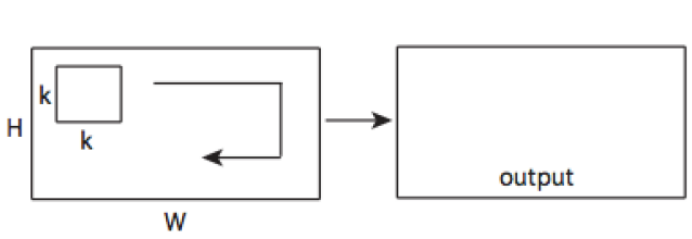
J. Donahue *et al*., "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 677-691, 1 April 2017, doi: 10.1109/TPAMI.2016.2599174.

C3D: 3D Convolutional Networks

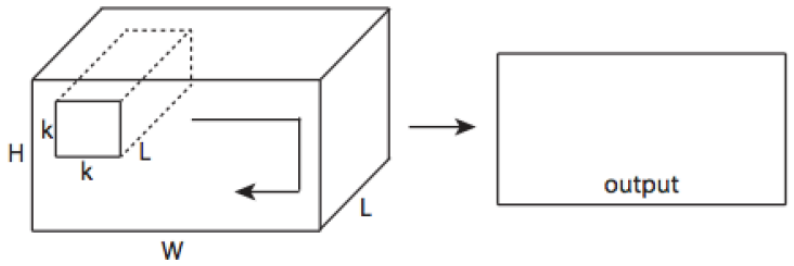
- Du Tran et al., “Learning Spatiotemporal Features with 3D Convolutional Networks,” 2014 (ArxivLink)

- Repurposing 3D convolutional networks as feature extractors
- Extensive search for best 3D convolutional kernel and architecture
- Using deconvolutional layers to interpret model decision

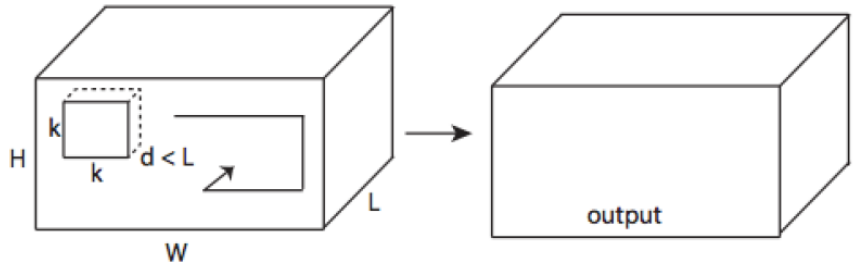




(a) 2D convolution



(b) 2D convolution on multiple frames

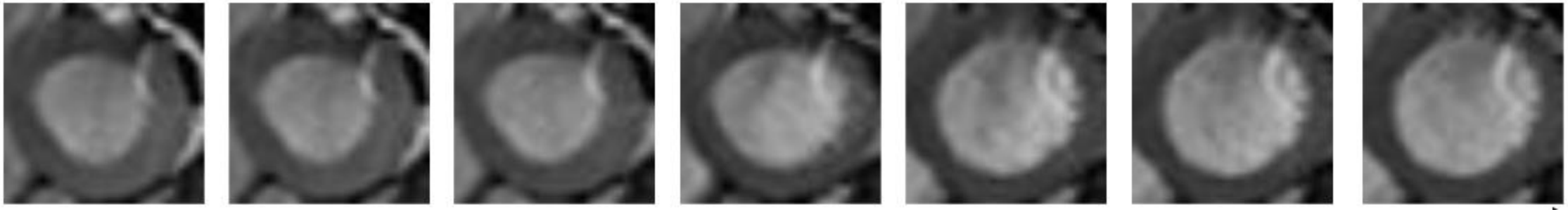


(c) 3D convolution

Cine MRI for Microvascular Obstruction Identification

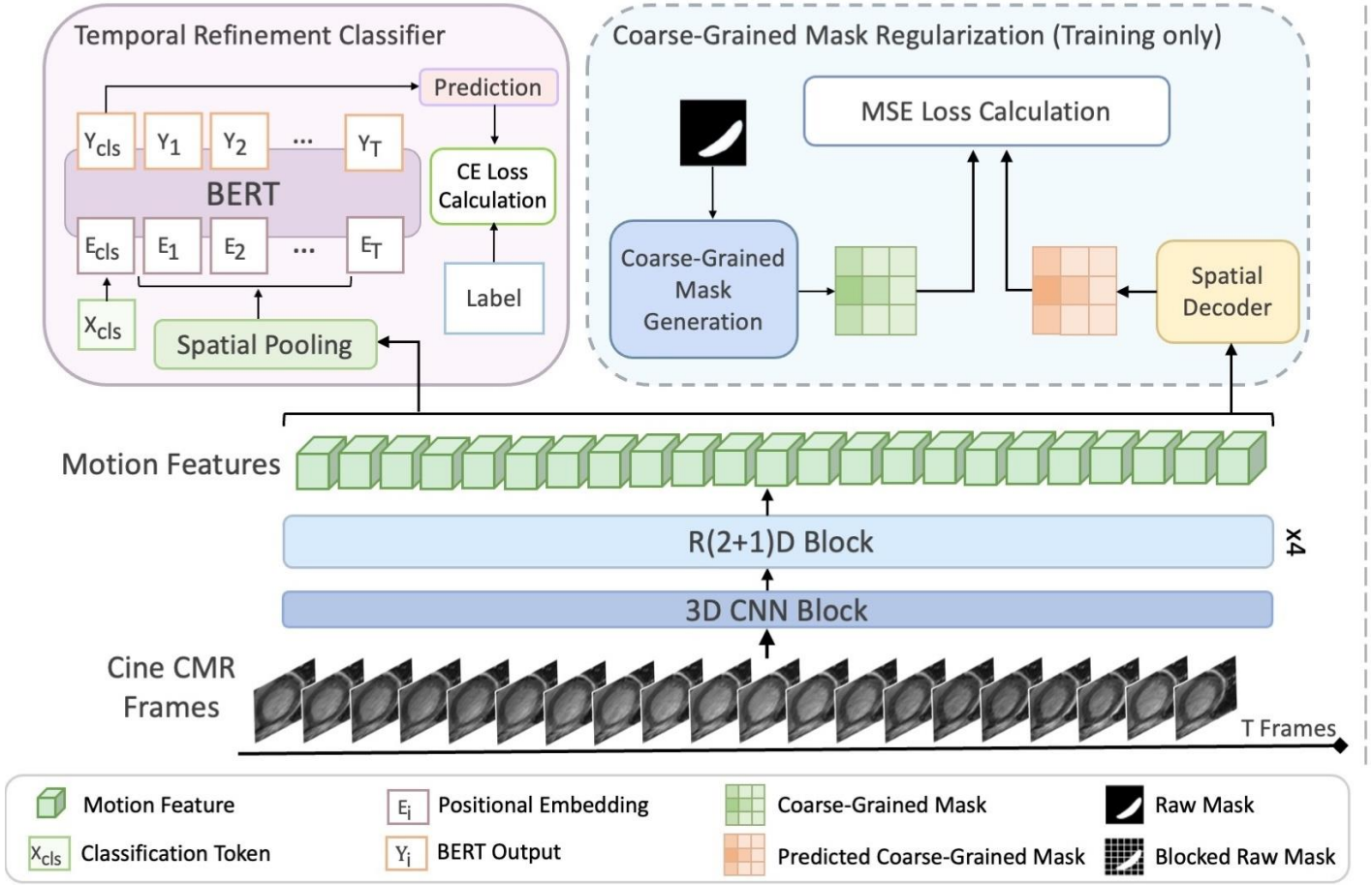
- What we see:

- What a computer sees:



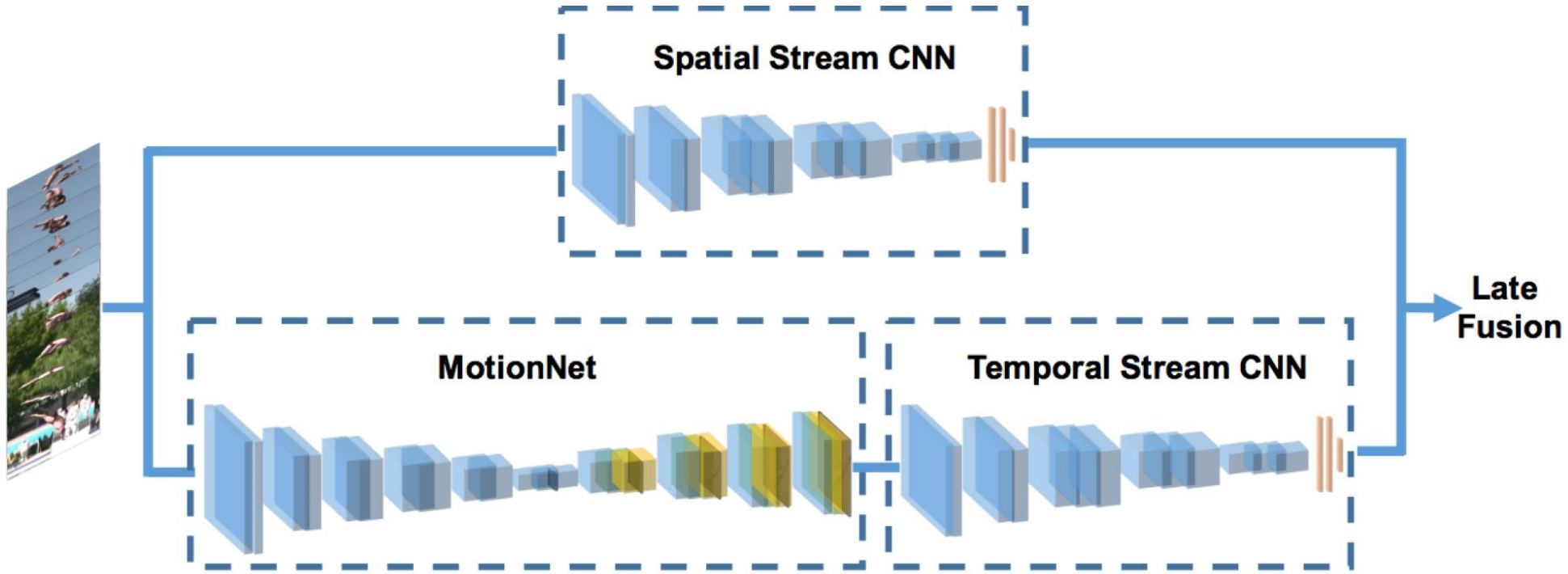
Healthy or Unhealthy?

Cine MRI for Microvascular Obstruction Identification



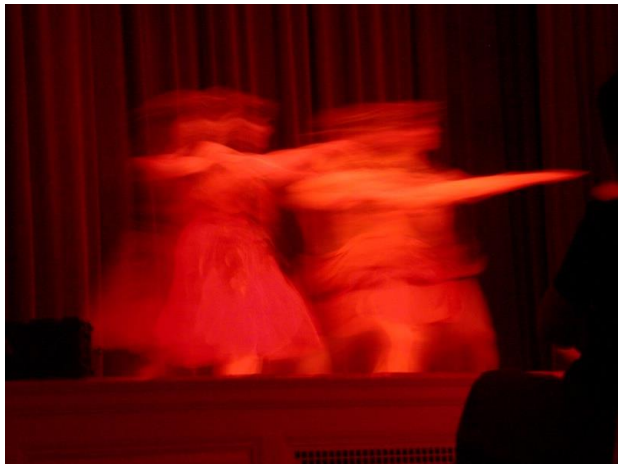
Hidden Two Stream

- Zhu et al., “Hidden Two-Stream Convolutional Networks for Action Recognition,” 2017
- •Novel architecture for generating optical flow input on-the-fly using a separate network



Video Enhancement

Motion blur is everywhere



Object Motion

Camera Motion

Blur model

Point-spread function

$$g(n_1, n_2) = d(n_1, n_2) * f(n_1, n_2) + w(n_1, n_2)$$

$$= \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} d(i,j) f(n_1 - i, n_2 - j) + w(n_1, n_2)$$

- d(x, y) takes on nonnegative values only, because of the physics of the underlying image formation process;
- when real-valued images are dealt with the point-spread function d(x, y) is real-valued too;
- the imperfections in the image formation process are modeled as passive operations on the data, i.e, no “energy” is absorbed or generated.

$$\sum_{n_1=1}^{N-1} \sum_{n_2=1}^{M-1} d(n_1, n_2) = 1$$

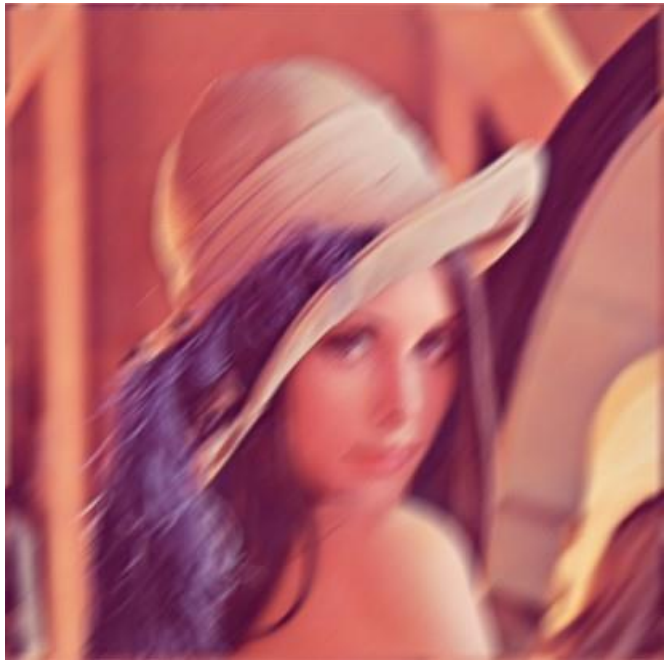
Blur model

No blur:

PSF is modeled as a unit pulse:

$$d(n_1, n_2) = \delta(n_1, n_2) = \begin{cases} 1 & \text{if } n_1 = n_2 = 0 \\ 0 & \text{elsewhere} \end{cases}$$

$$g(n_1, n_2) = \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} d(i, j) f(n_1 - i, n_2 - j) + w(n_1, n_2)$$
$$= f(n_1, n_2) + w(n_1, n_2)$$



Blind Image Deconvolution

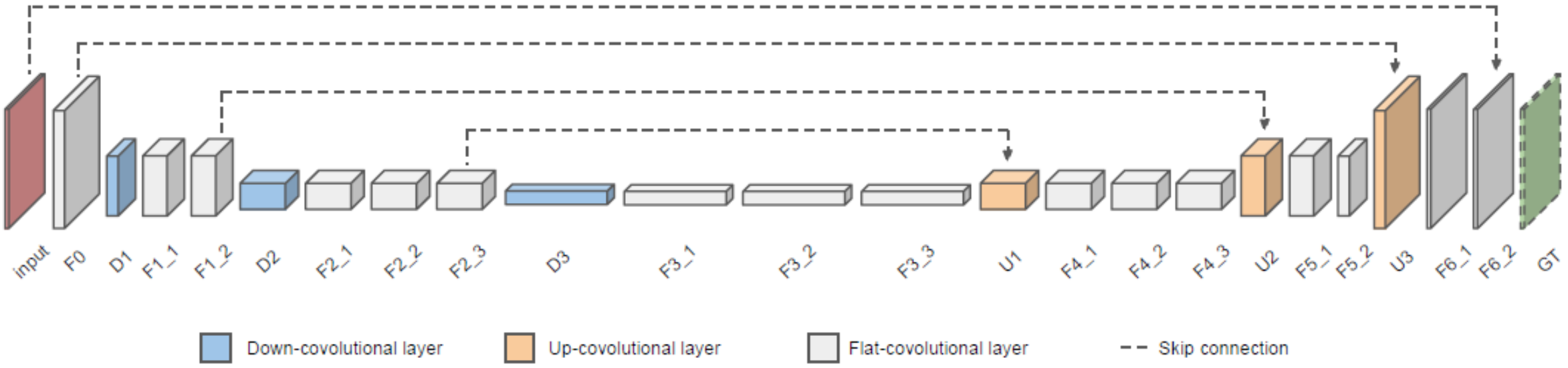
- Accurate Point Spread Function (PSF) Needed.

Deep video deblur

Input: the stacked nearby frames

Structure: Encoder – Decoder

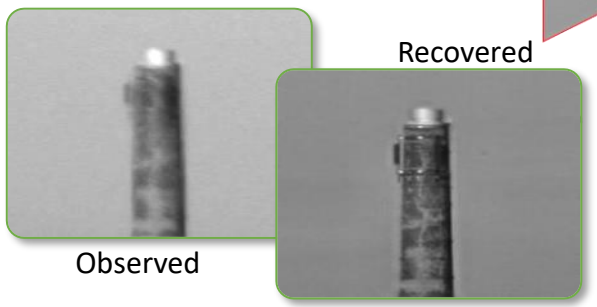
Project home: https://www.cs.ubc.ca/labs/imager/tr/2017/DeepVideoDeblurring/

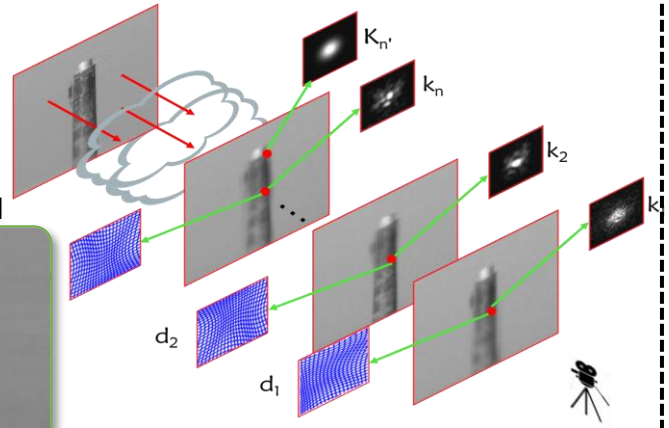


S. Su, et al., "Deep Video Deblurring for Hand-Held Cameras," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 237-246.

Video Turbulence Effect Remove

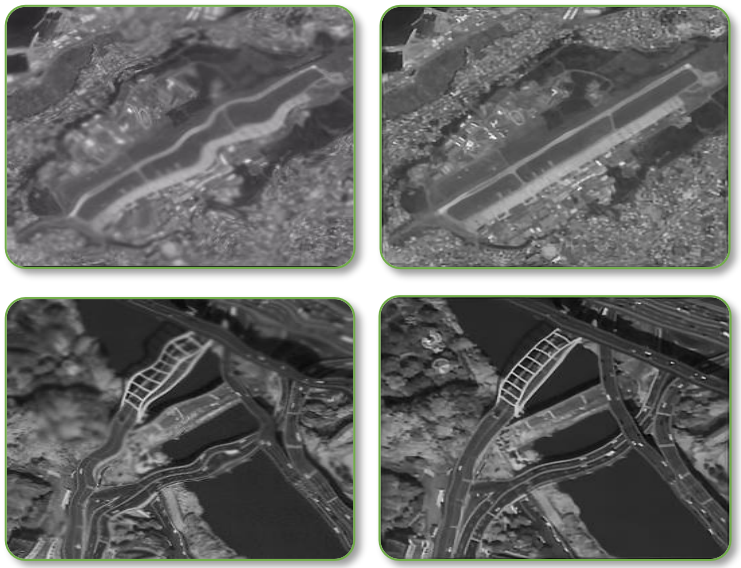
- **Local non-rigid deformation**
- **Space-time-varying blur**



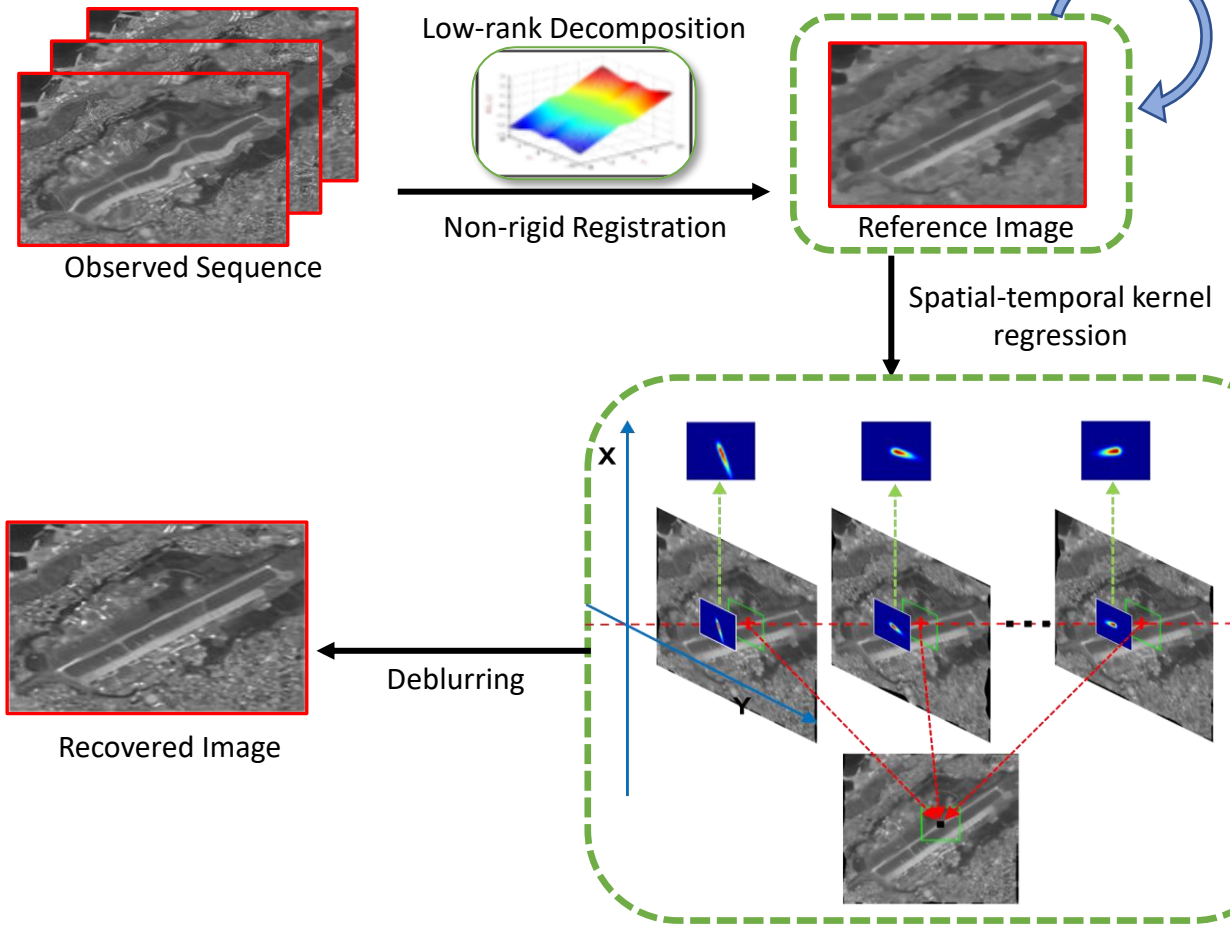


Air Turbulence

Water Turbulence



The Proposed Restoration Framework



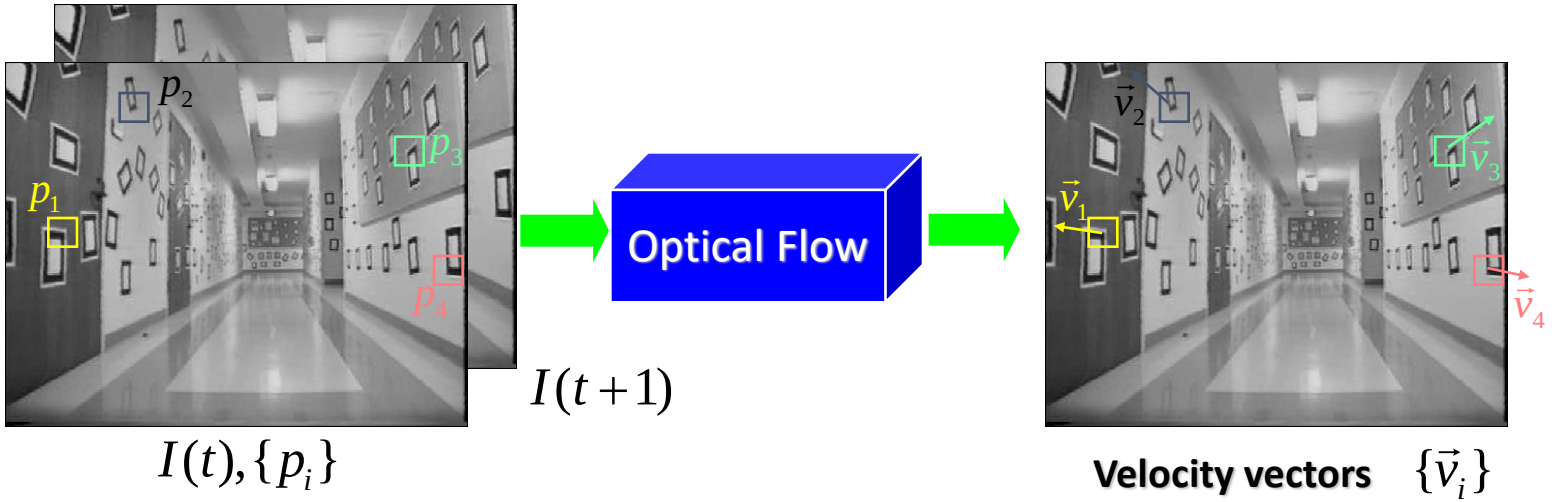
Xie, Zhang, Tao, Hu, Qu, Wang: Removing Turbulence Effect via Hybrid Total Variation and Deformation-Guided Kernel Regression. IEEE T-IP, vol. 25, no. 10, 2016.

Optical Flow

Image sequence
(single camera)

Tracked sequence

What is Optical Flow?

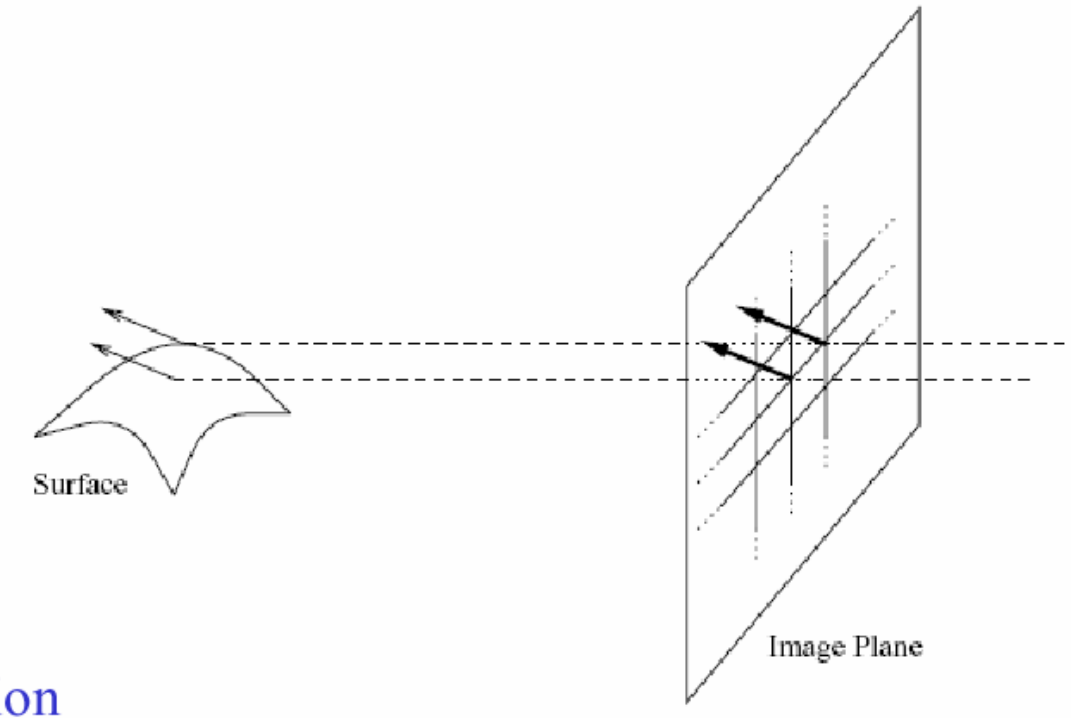


Optical flow is the relation of the motion field

- *the 2D projection of the physical movement of points relative to the observer* to 2D displacement of pixel patches on the image plane.

Optical Flow Assumptions:

Spatial Coherence

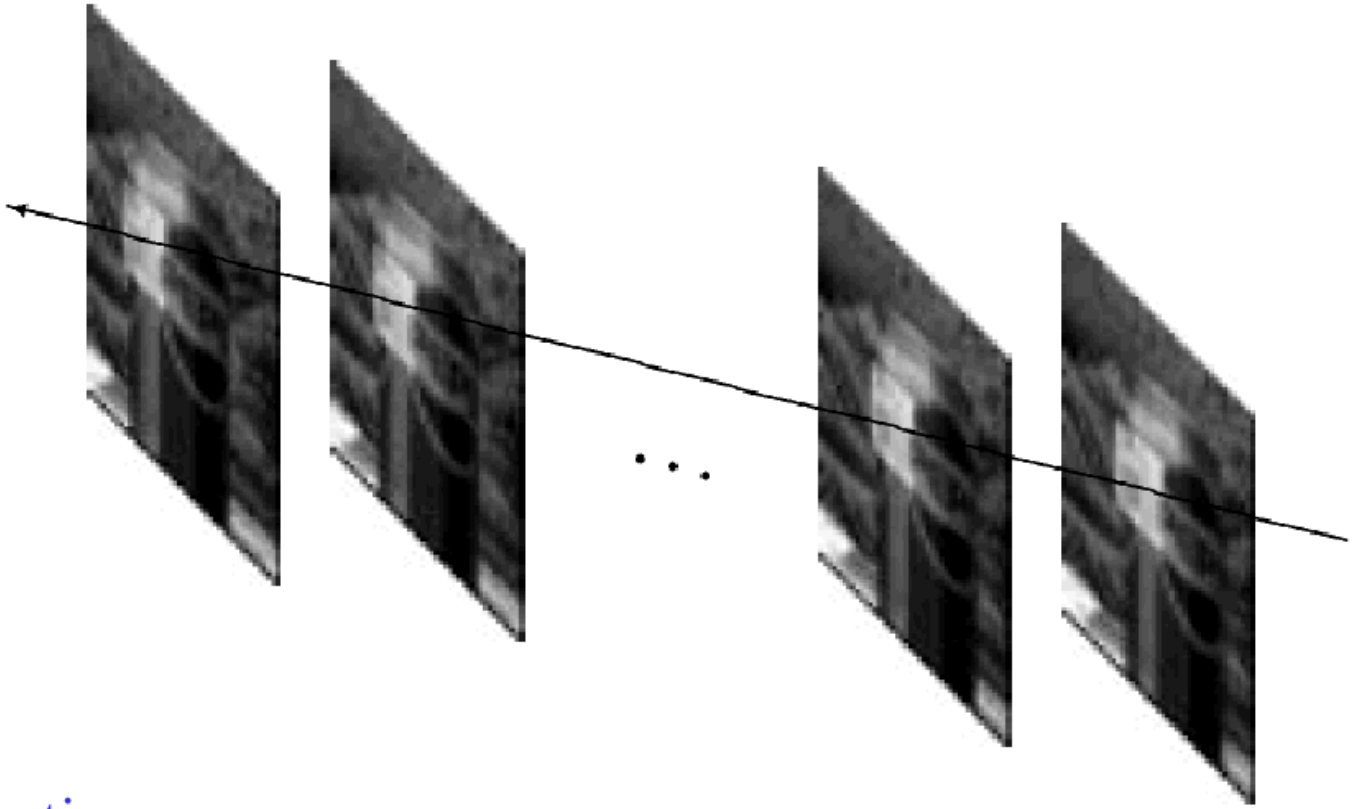


Assumption

* Neighboring points in the scene typically belong to the same surface and hence typically have similar motions.
* Since they also project to nearby points in the image, we expect spatial coherence in image flow.

Optical Flow Assumptions:

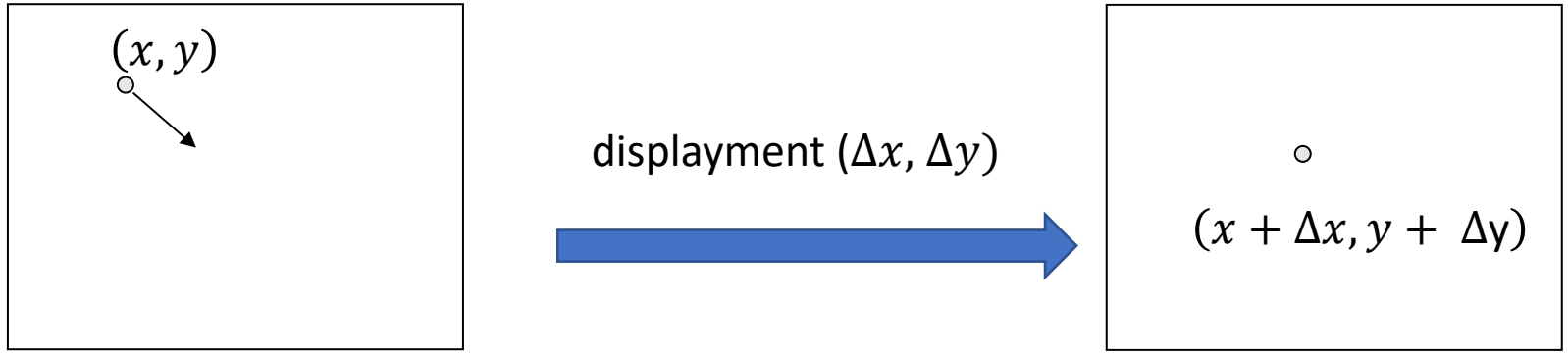
Temporal Persistence



Assumption:

The image motion of a surface patch changes gradually over time.

Optical Flow Assumptions: The brightness constancy constraint



Brightness Constancy Assumption

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

Brightness constancy equation

- Take Taylor expansion of $I(x+\Delta x, y+\Delta y, t+\Delta t)$ at (x, y, t) :

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

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

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

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

$$\frac{\partial I}{\partial x}\frac{dx}{dt} + \frac{\partial I}{\partial y}\frac{dy}{dt} + \frac{\partial I}{\partial t} = 0 \quad \text{or} \quad I_x \cdot \frac{dx}{dt} + I_y \cdot \frac{dy}{dt} + I_t = 0 \quad \text{or} \quad \nabla I \cdot \left[\frac{dx}{dt} \ \frac{dy}{dt}\right]^T + I_t = 0$$

The brightness constancy constraint

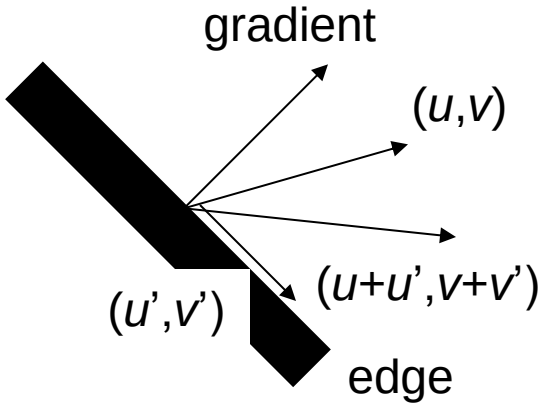
- How many equations and unknowns per pixel? $\nabla I \cdot \begin{bmatrix} u & v \end{bmatrix}^T + I_t = 0$

  One equation (this is a scalar equation!), two unknowns (u,v)

The component of the motion perpendicular to the gradient (i.e., parallel to the edge) cannot be measured

If (u, v) satisfies the equation, so does (u+u', v+v') if

$$\nabla I \cdot \begin{bmatrix} u' & v' \end{bmatrix}^T = 0$$



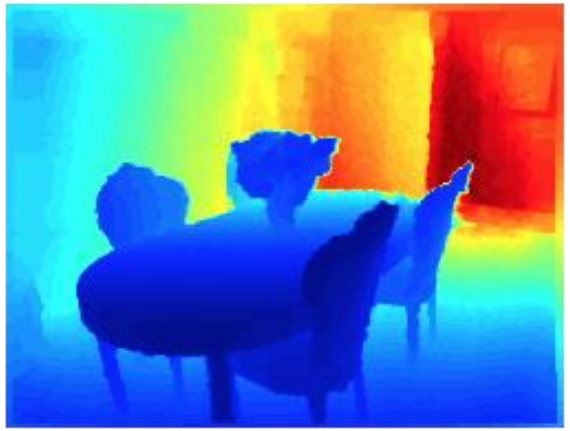
Video Segmentation

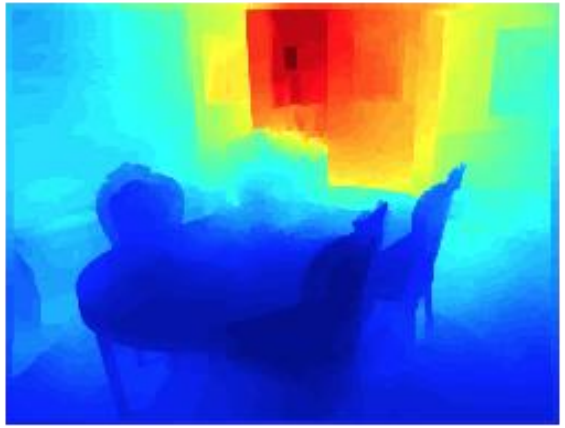
The process of segment objects from video sequences.

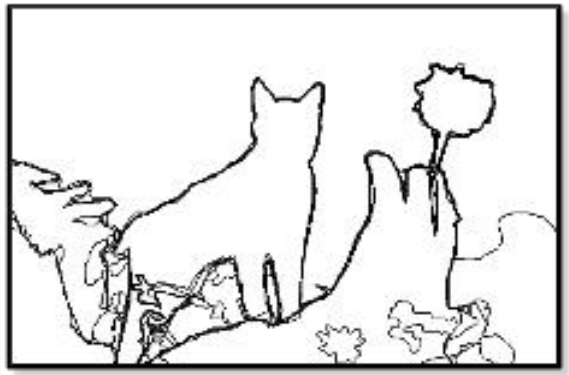
Segmentation: Pixels in, pixels out

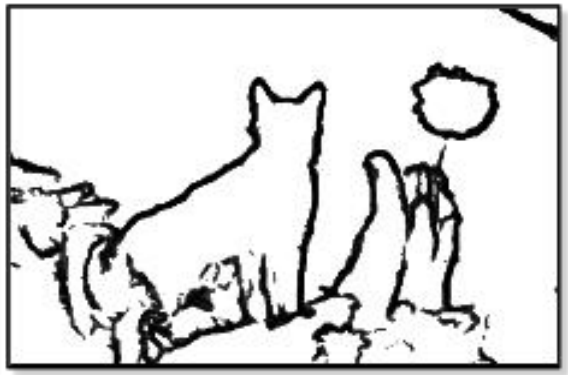
semantic segmentation

monocular depth estimation (Liu et al. 2015)



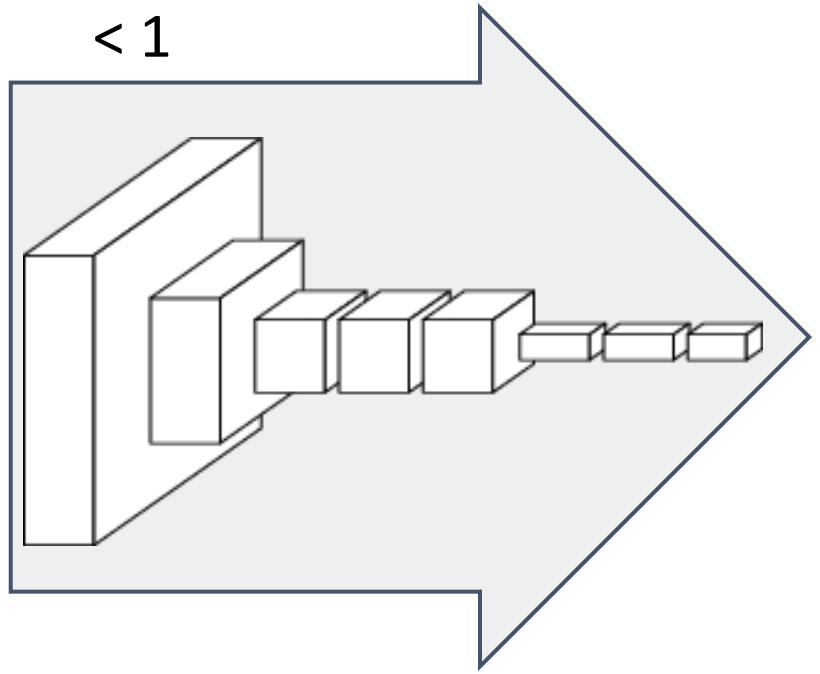


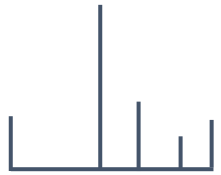




boundary prediction (Xie & Tu 2015)

Convnets perform classification



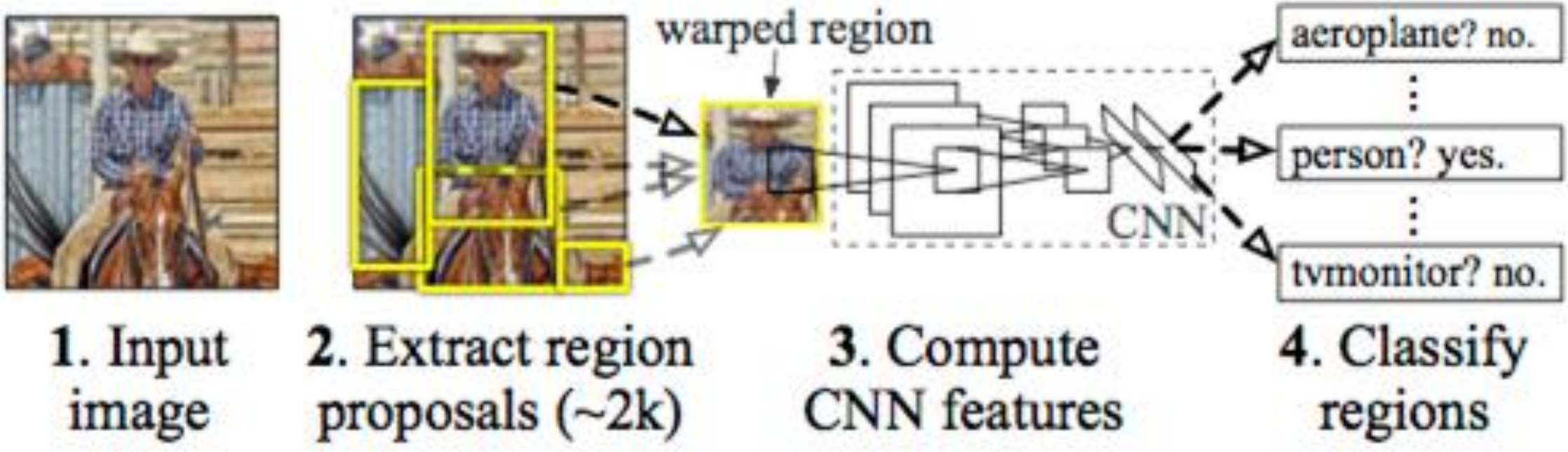


1000-dim vector

“tabby cat”

end-to-end learning

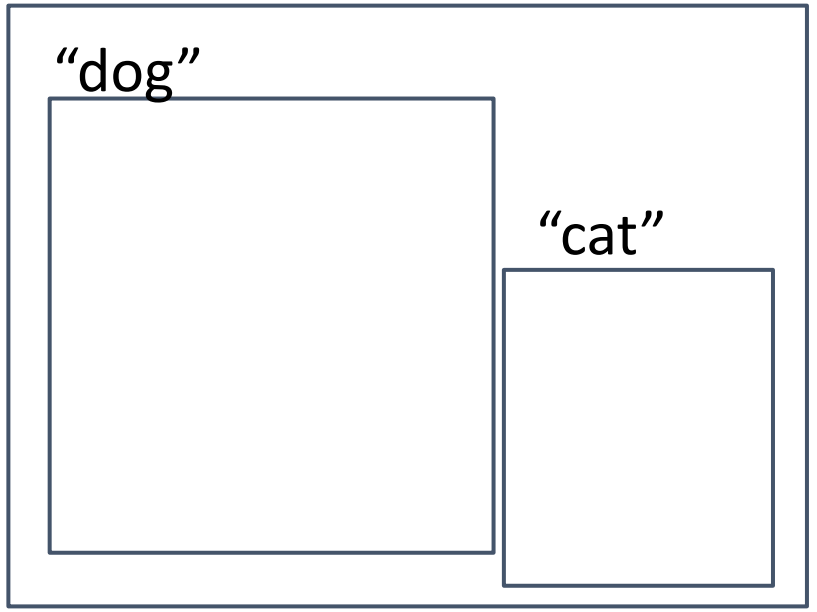
R-CNN


warped region
CNN
aeroplane? no.
person? yes.
tvmonitor? no.
1. Input image
2. Extract region proposals (~2k)
3. Compute CNN features
4. Classify regions

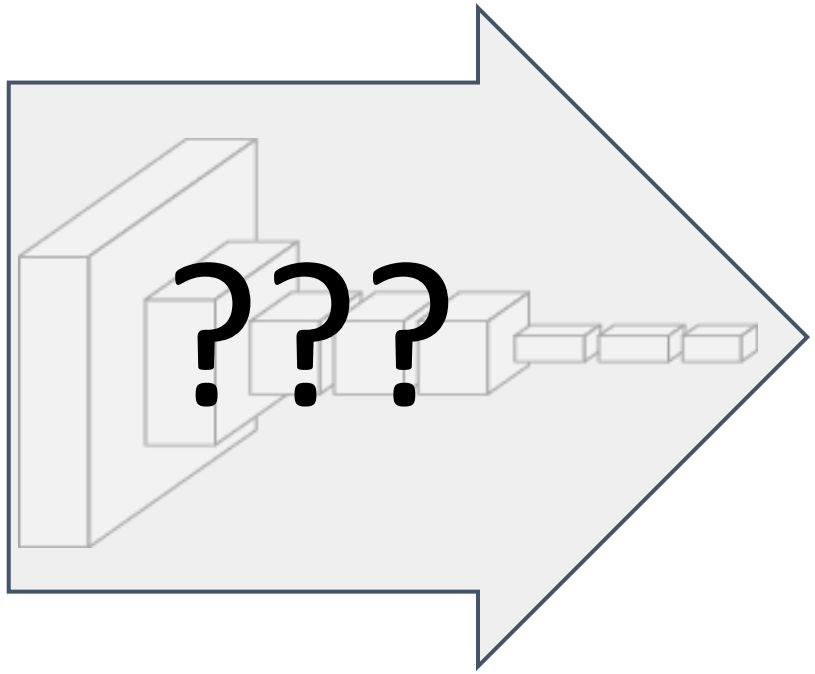
R-CNN does detection

many seconds

R-CNN

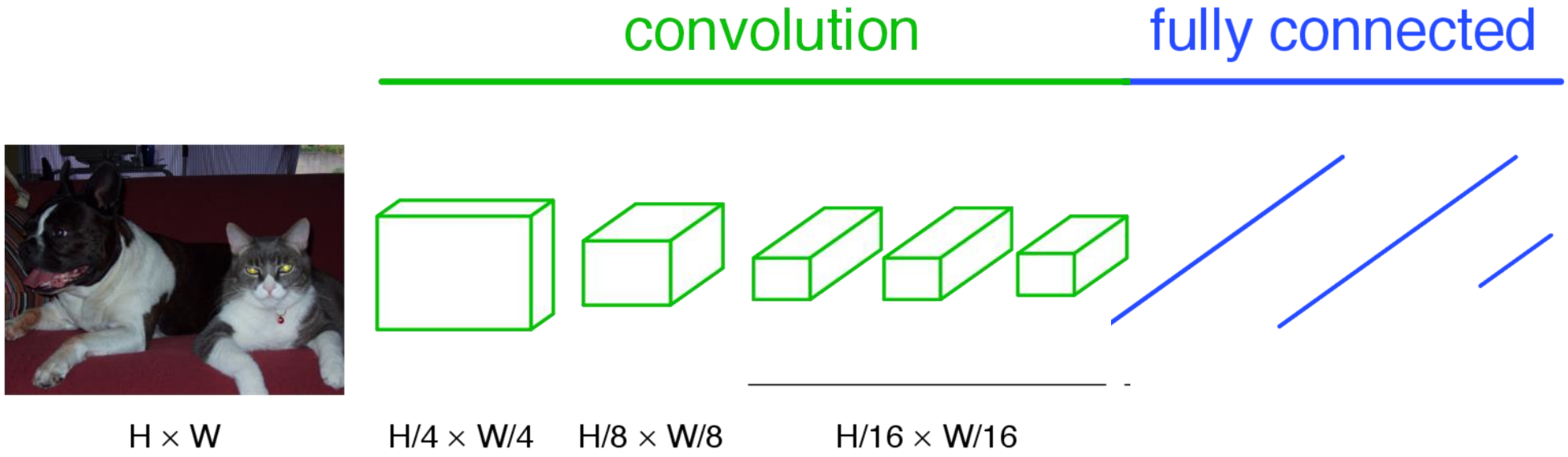


Segmentation?



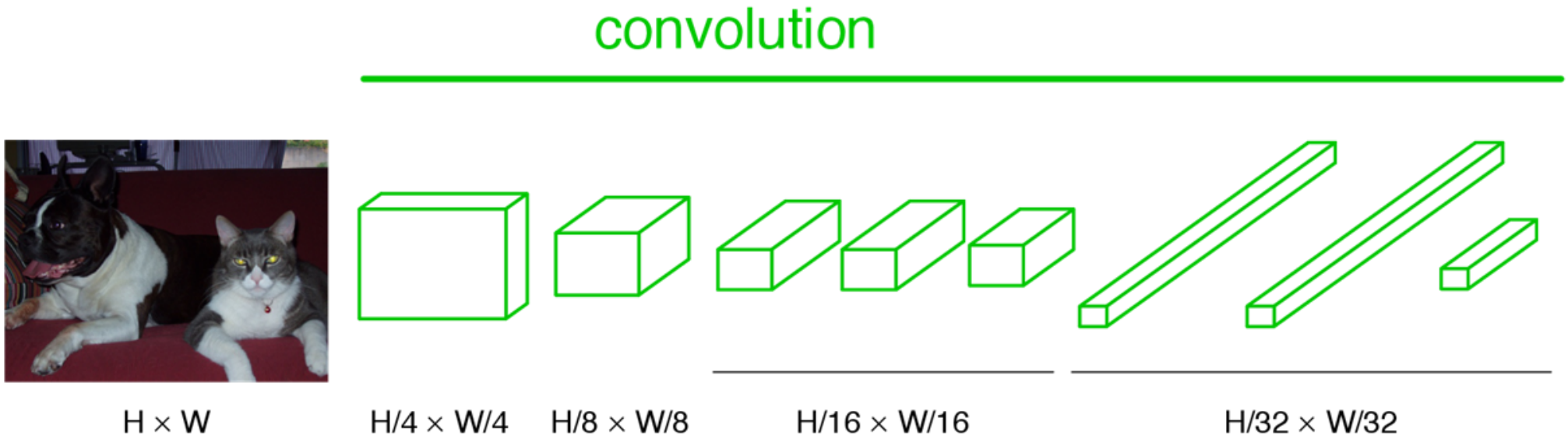
end-to-end learning

A classification network

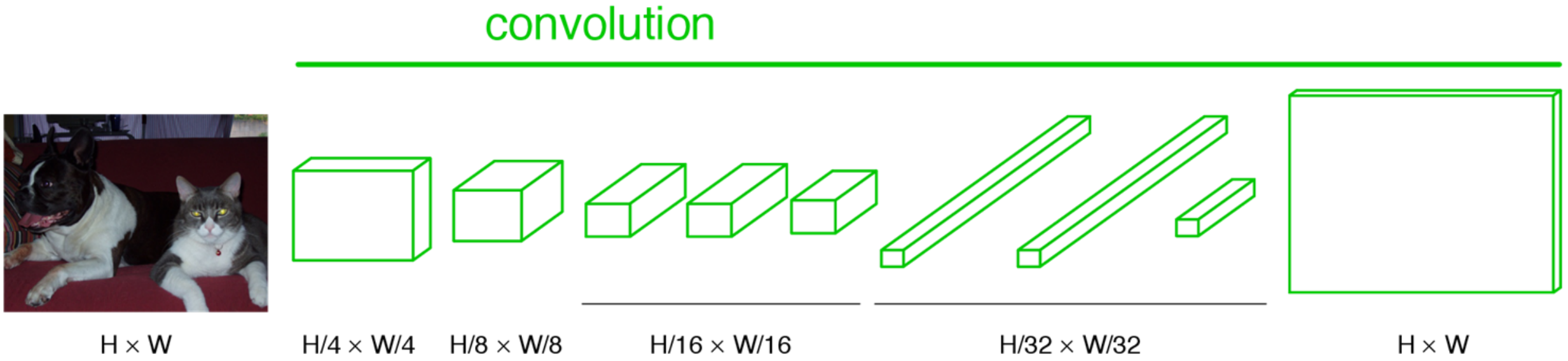


“tabby cat”

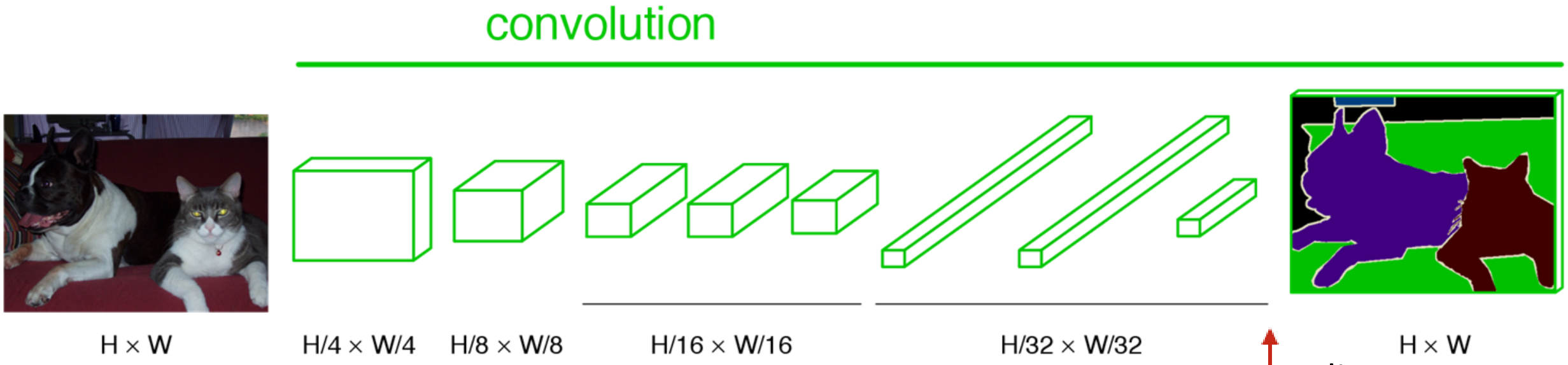
Change to fully convolutional



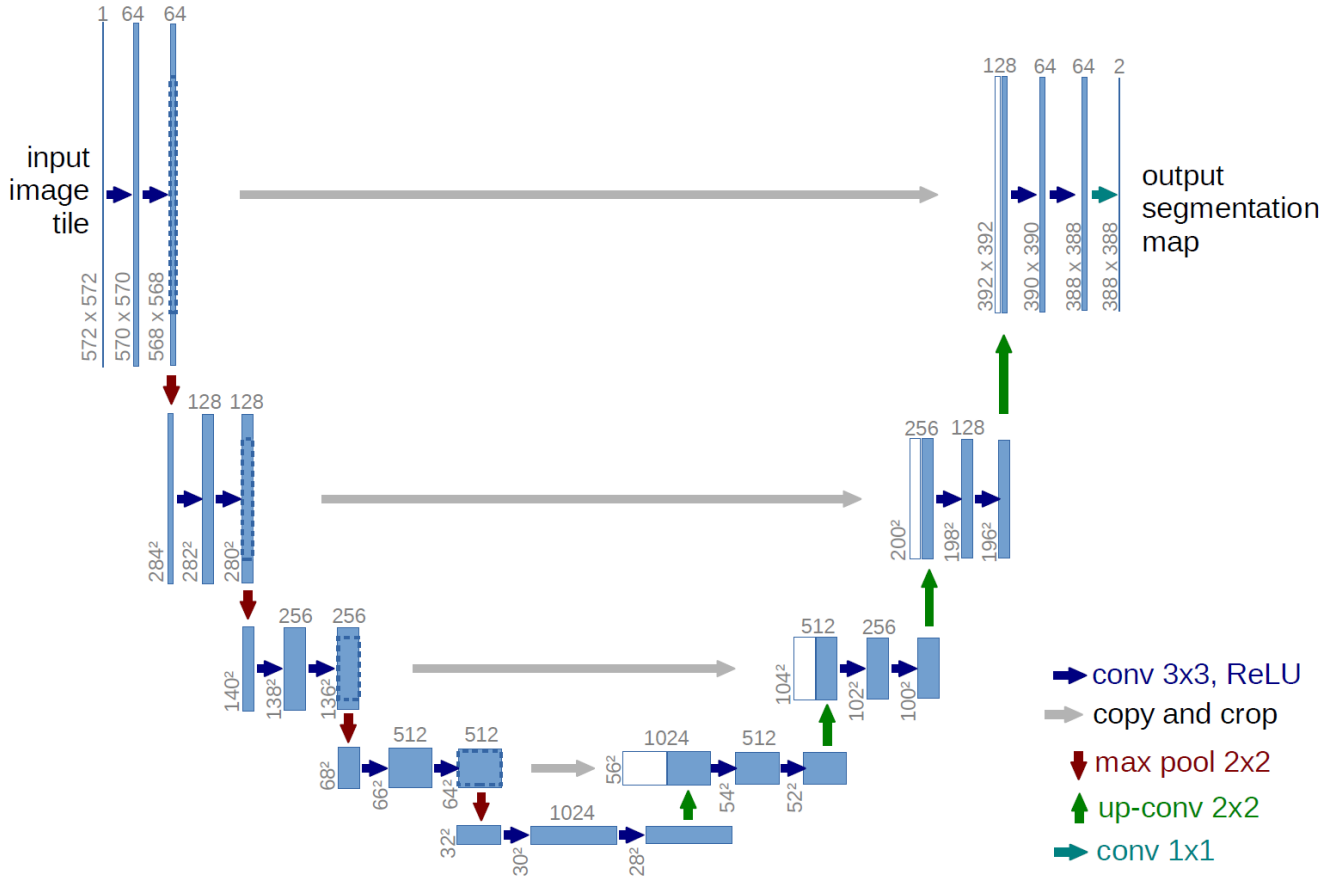
Upsampling output



End-to-end, pixels-to-pixels network

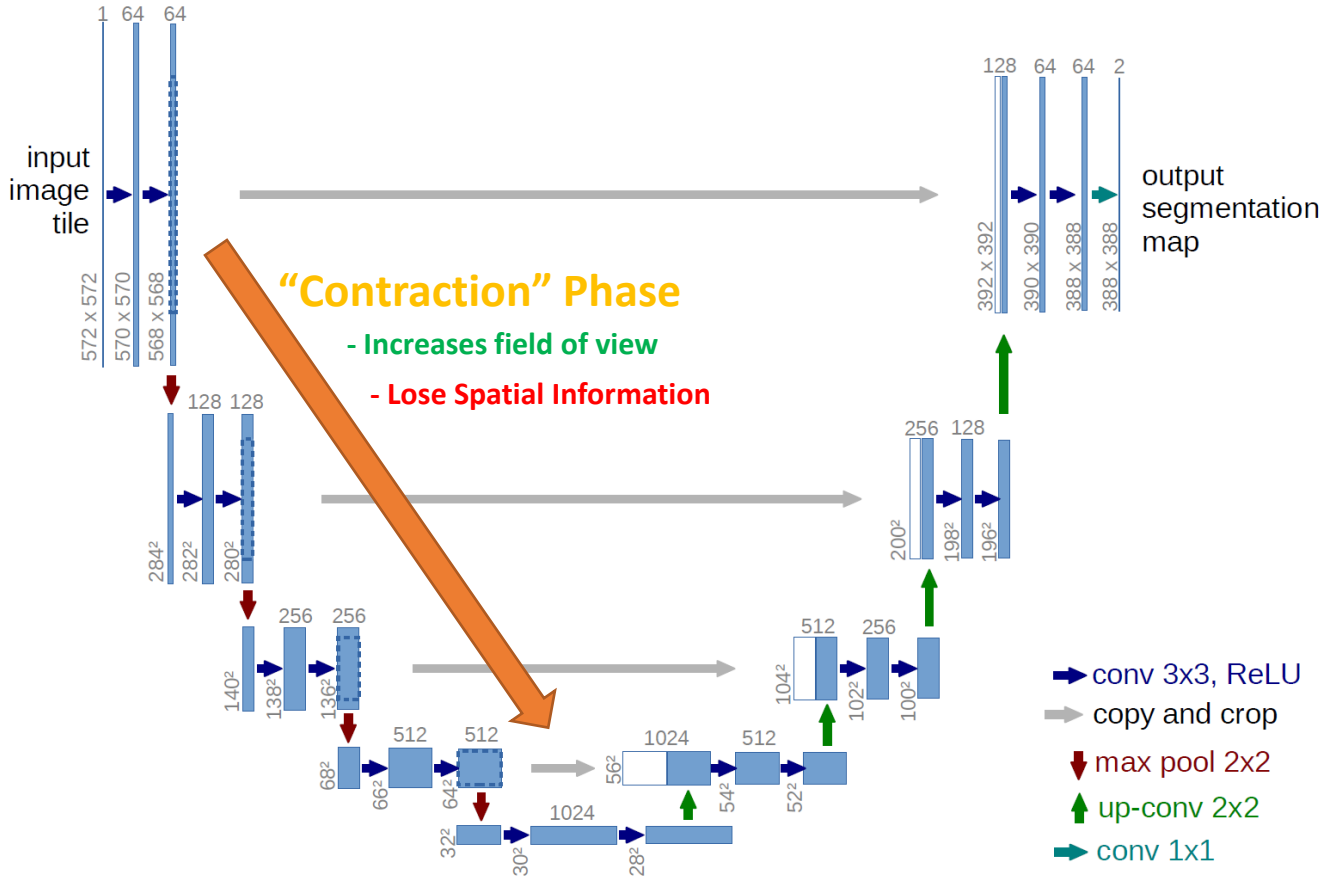


U-Net Architecture

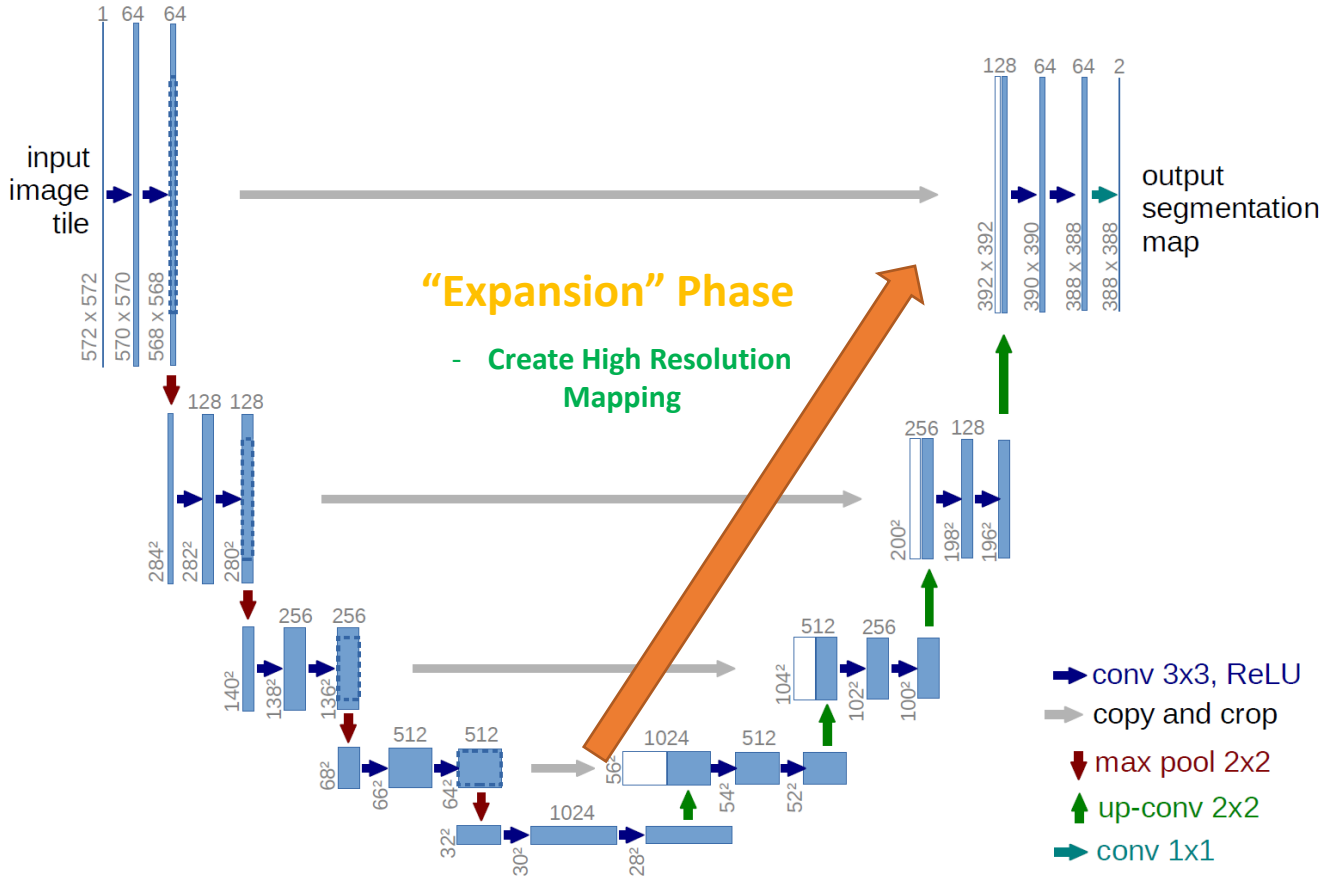


Ronneberger et al. (2015) U-net Architecture

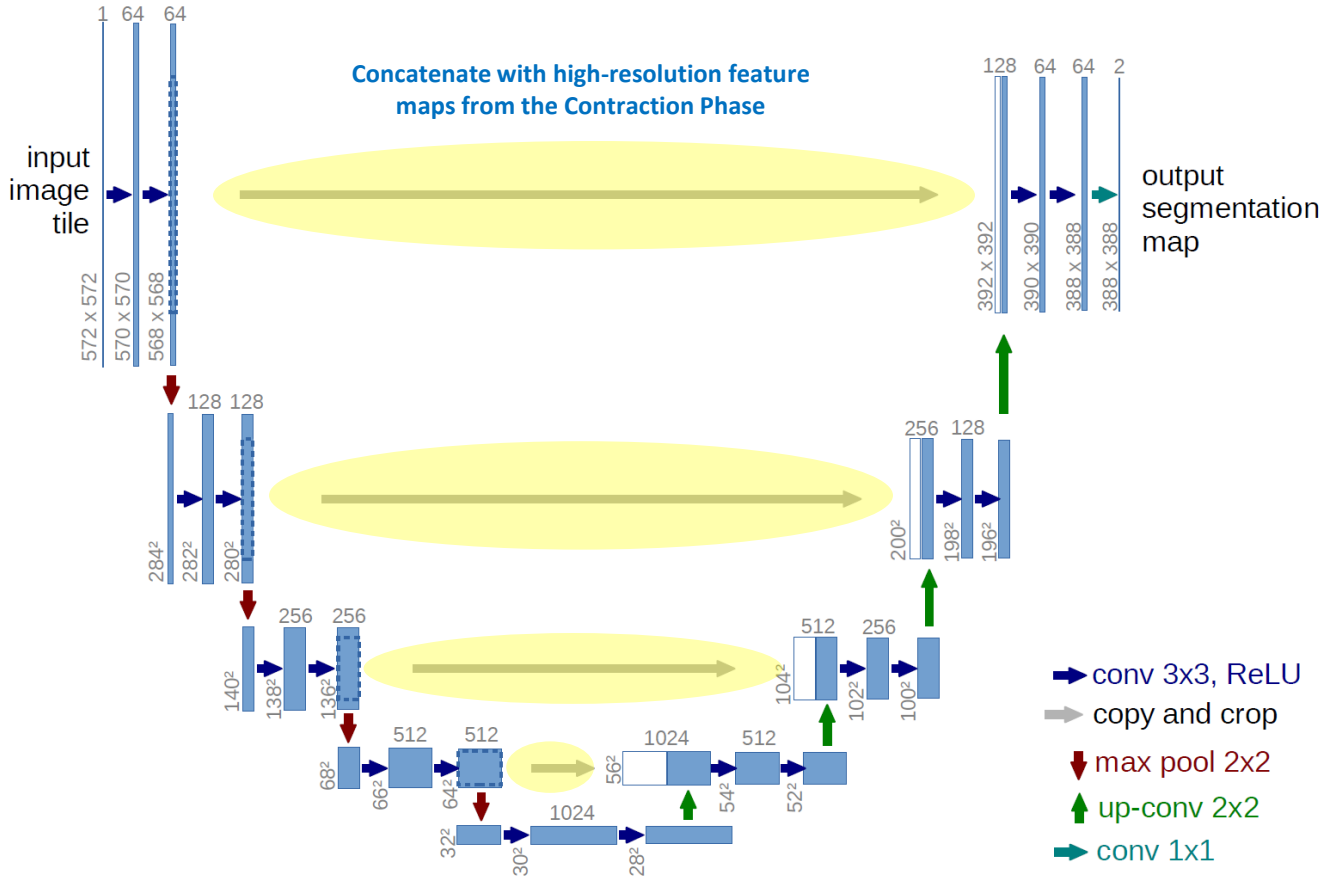
U-Net Architecture



U-Net Architecture



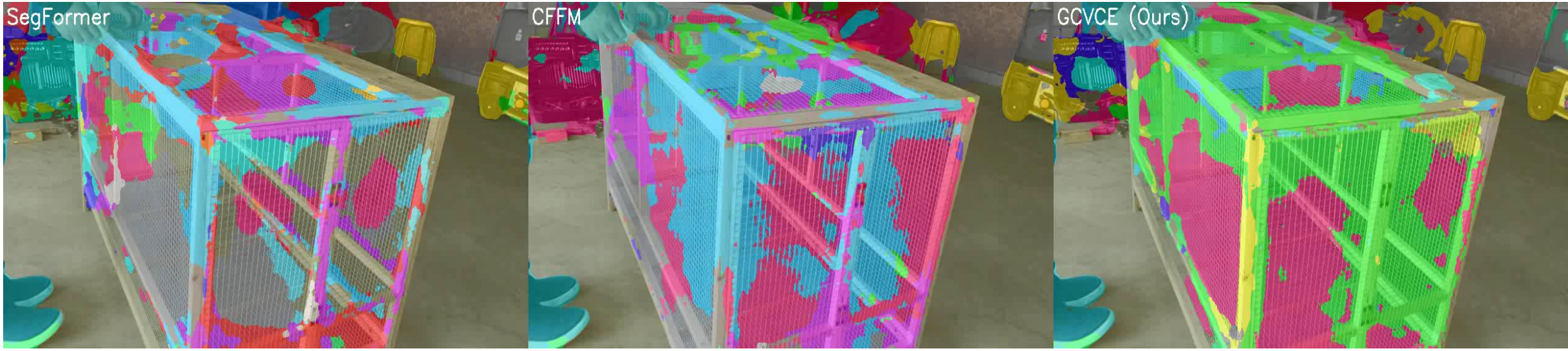
U-Net Architecture



U-Net Summary

- Contraction Phase
  - Reduce spatial dimension, but increases the “what.”
- Expansion Phase
  - Recovers object details and the dimensions, which is the “where.”
- Concatenating feature maps from the Contraction phase helps the Expansion phase with recovering the “where” information.

Image Segmentation vs. Video Segmentation



SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers, NeurIPS 2021
CFFM: Coarse-to-Fine Feature Mining for Video Semantic Segmentation, CVPR 2022

- Image Segmentation
  - ➢ Frame based
  - ➢Flicker artefact
- Video Segmentation
  - ➢Frame-frame association
  - ➢Smooth in continues frames.

https://www.youtube.com/watch?v=hGrJ3zuuvRQ

Businesswoman in China caught ‘jaywalking’

What is missing?

1. Real human vs photo
2. 2D vs 3D
3. Realistic motion