facebook
Artificial Intelligence

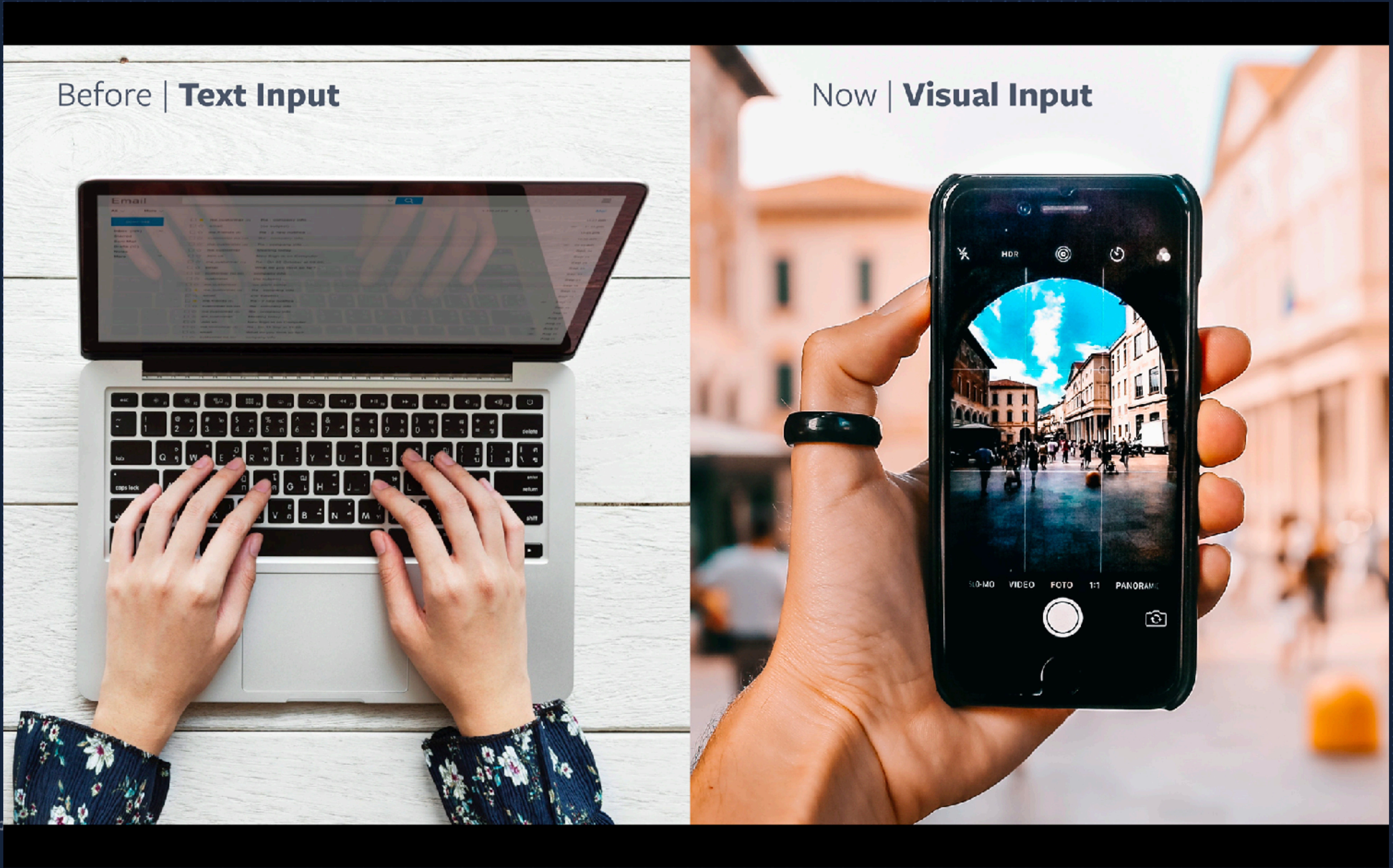
Computer Vision @ Scale

Manohar Paluri
Director, Facebook AI


Connecting People to bring them closer together

AI is an essential piece of the puzzle

And helping machines understand the visual world is an important component!



What will you learn today?

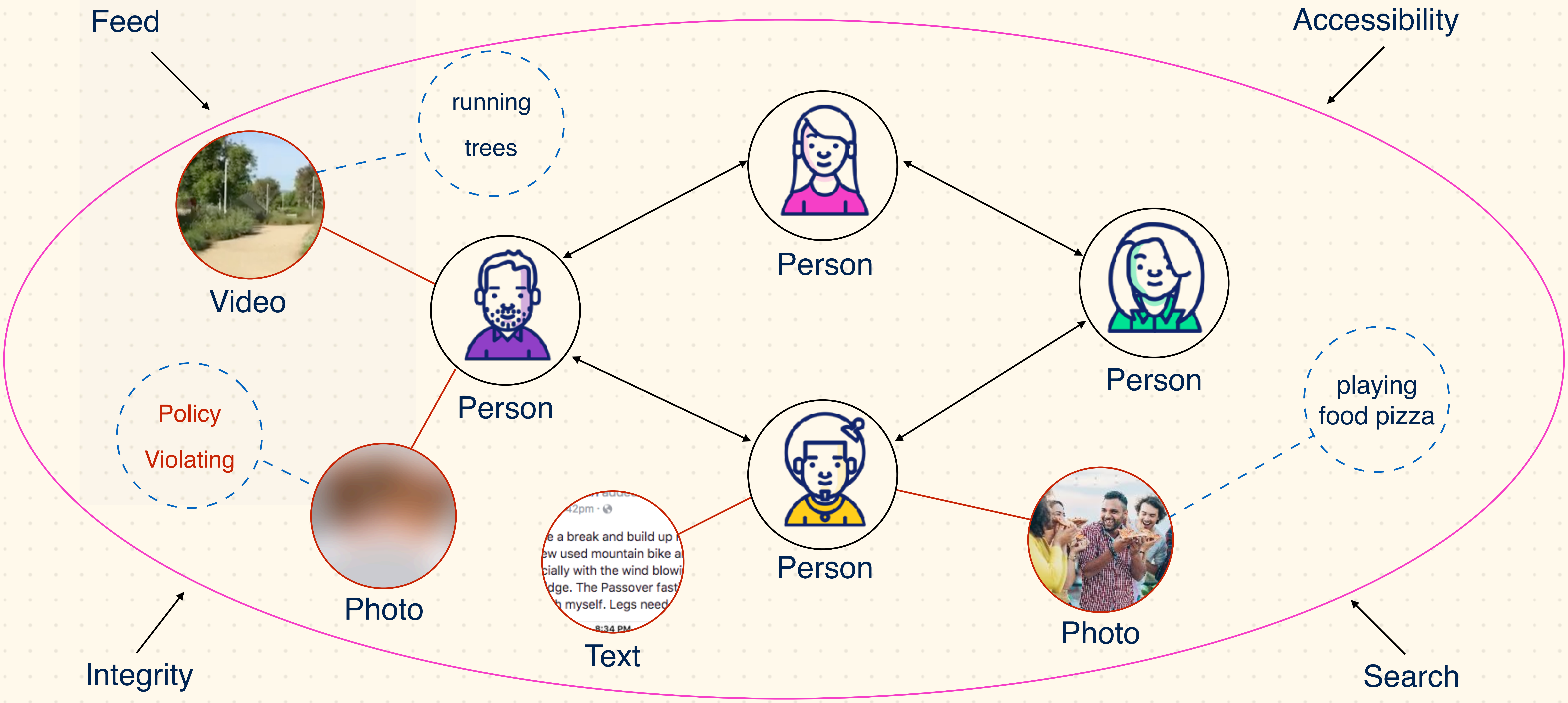
How do you design a image and video recognition system for billion scale?

Can you remove the requirement of annotation to learn best representations?

Can we understand video faster than understanding individual frames?

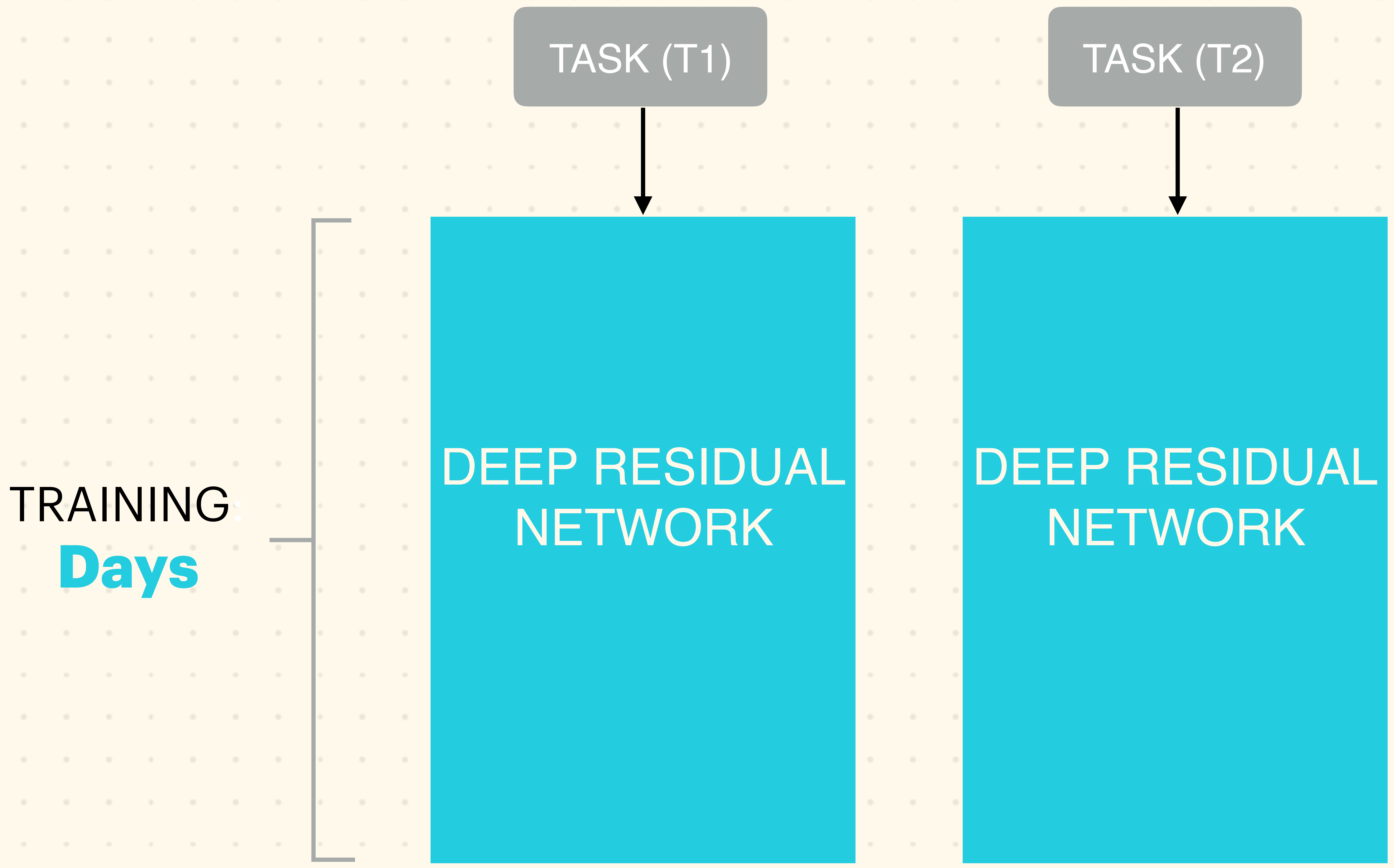
How does pushing state of the art in CV make a meaningful difference to everyone in the world?

Social Graph -> Semantic Graph



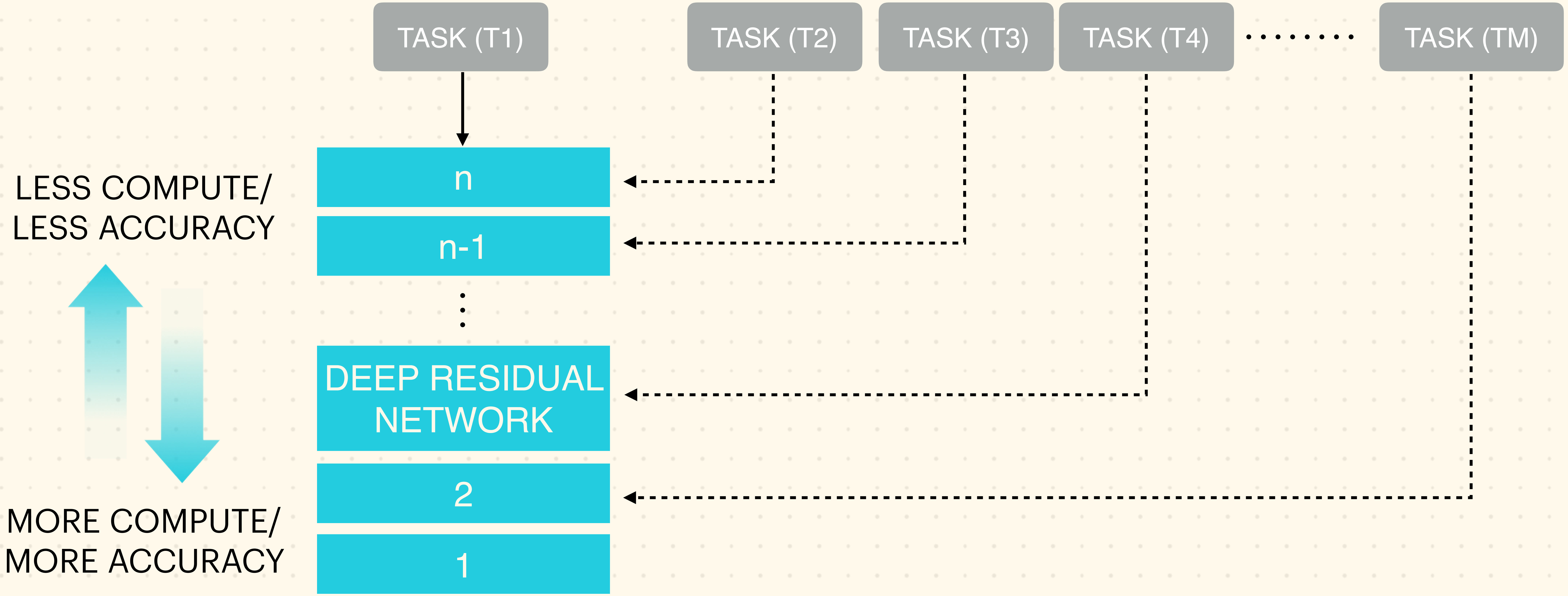
Vision Models In Production

- Multiple vision tasks need to be done
- Cannot afford one separate model for each task
- Explosion in computations cost and resources

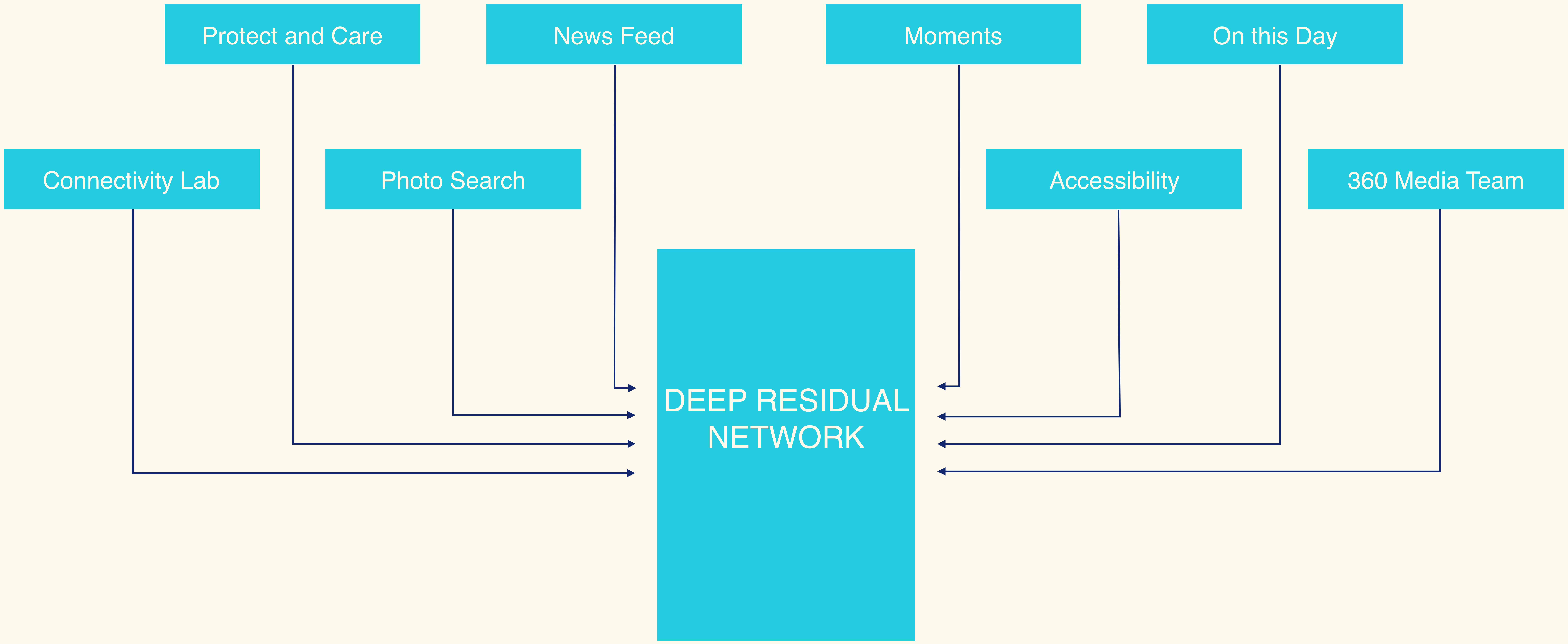


Our Vision - Towards Universal Vision Model

SUBTITLE



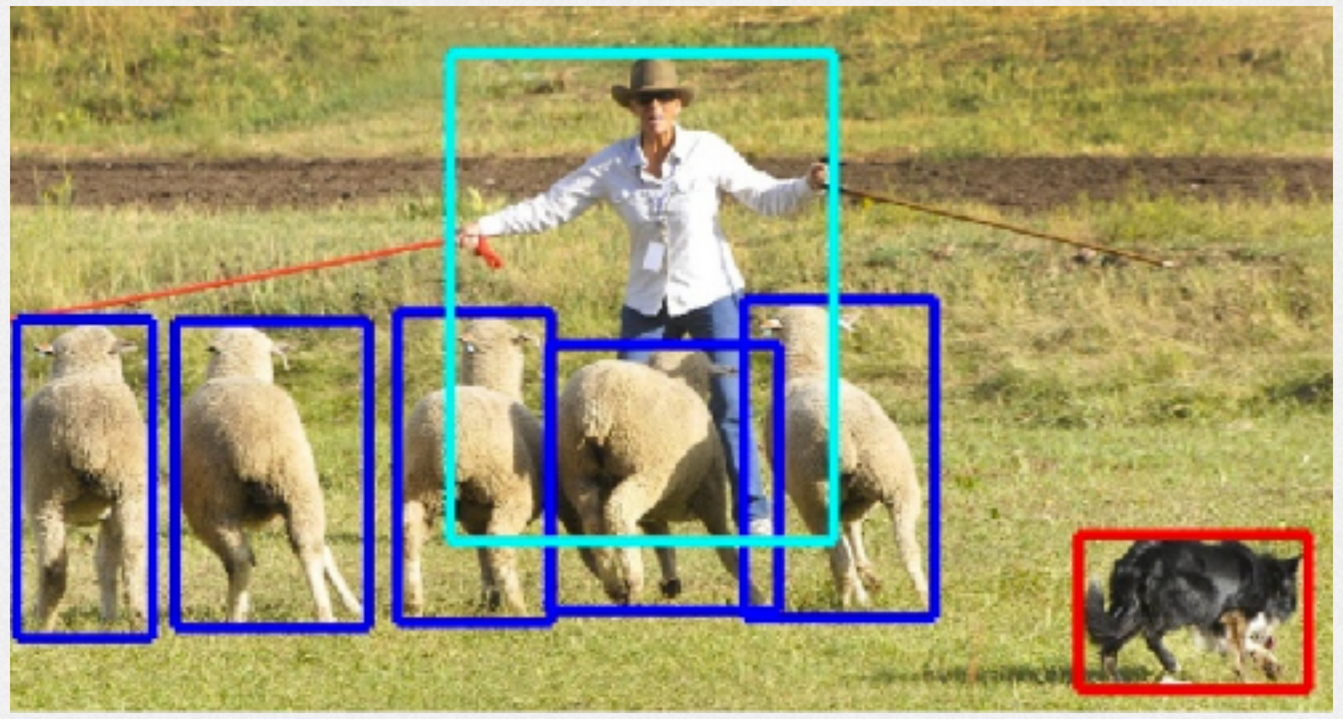
Our work allows to move the tasks towards upper layers


Protect and Care
News Feed
Moments
On this Day
Connectivity Lab
Photo Search
Accessibility
360 Media Team
DEEP RESIDUAL NETWORK

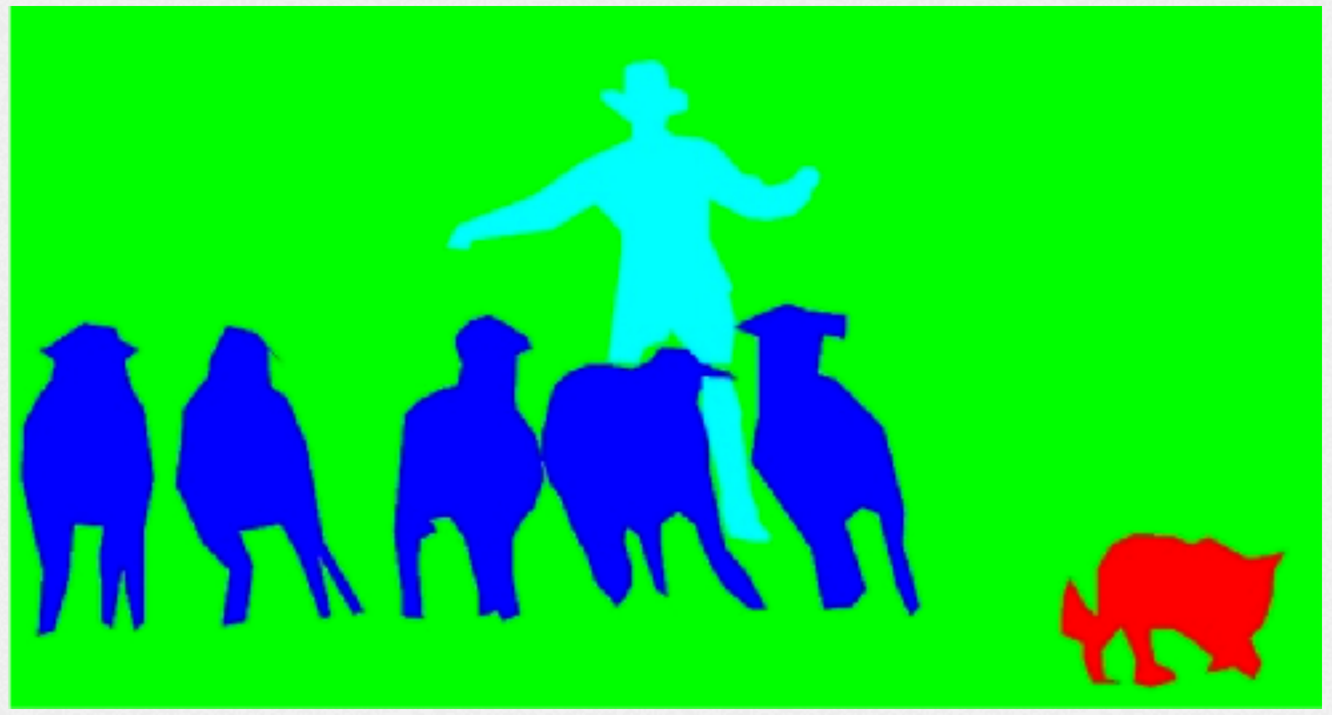
Branches

Progress in Image Understanding

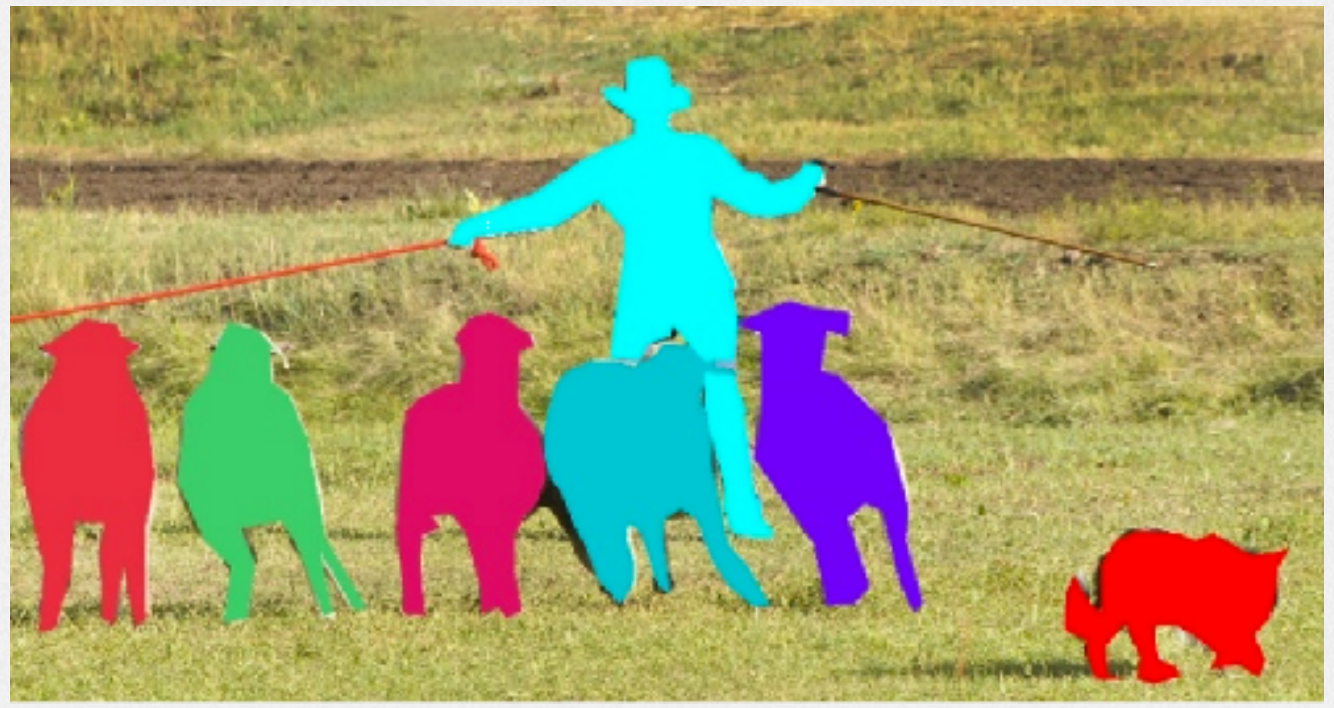
Classification



Detection



Semantic Segmentation



Instance Segmentation

Attributes

Relationships

Fine grained

Aesthetics

.

.

ALEXNET

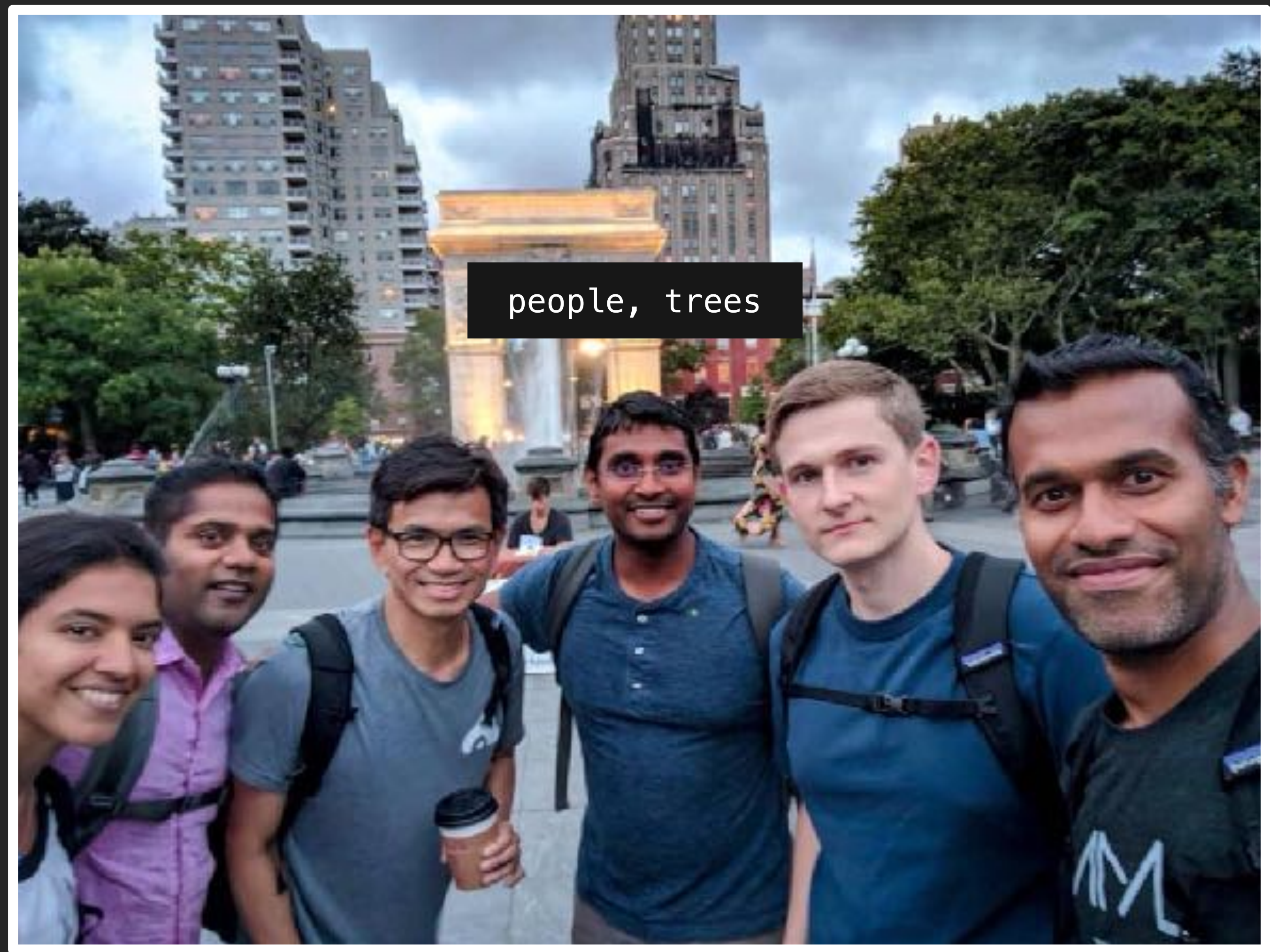
2012

ALEXNET

2012

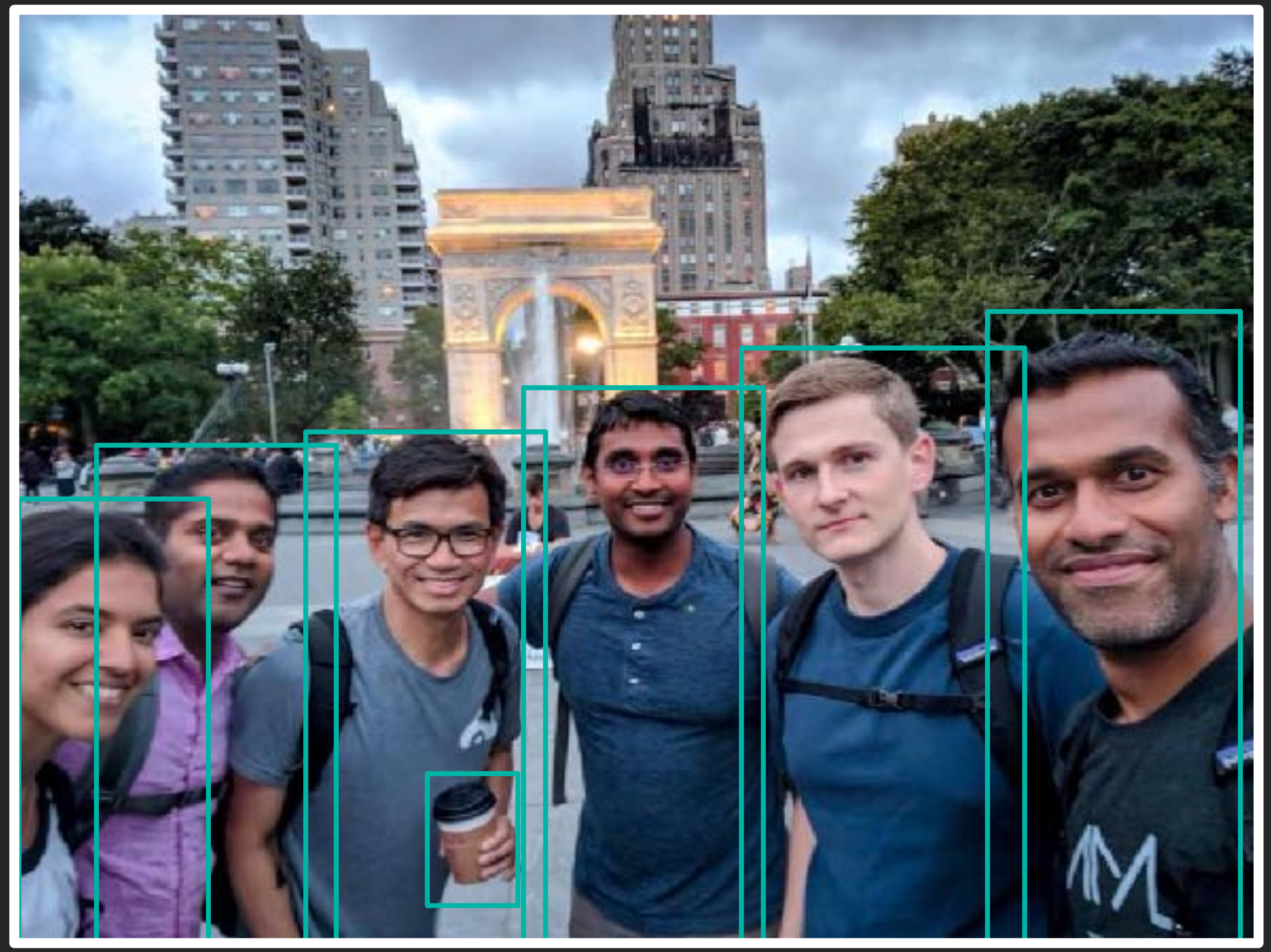
FASTER R-CNN

2015



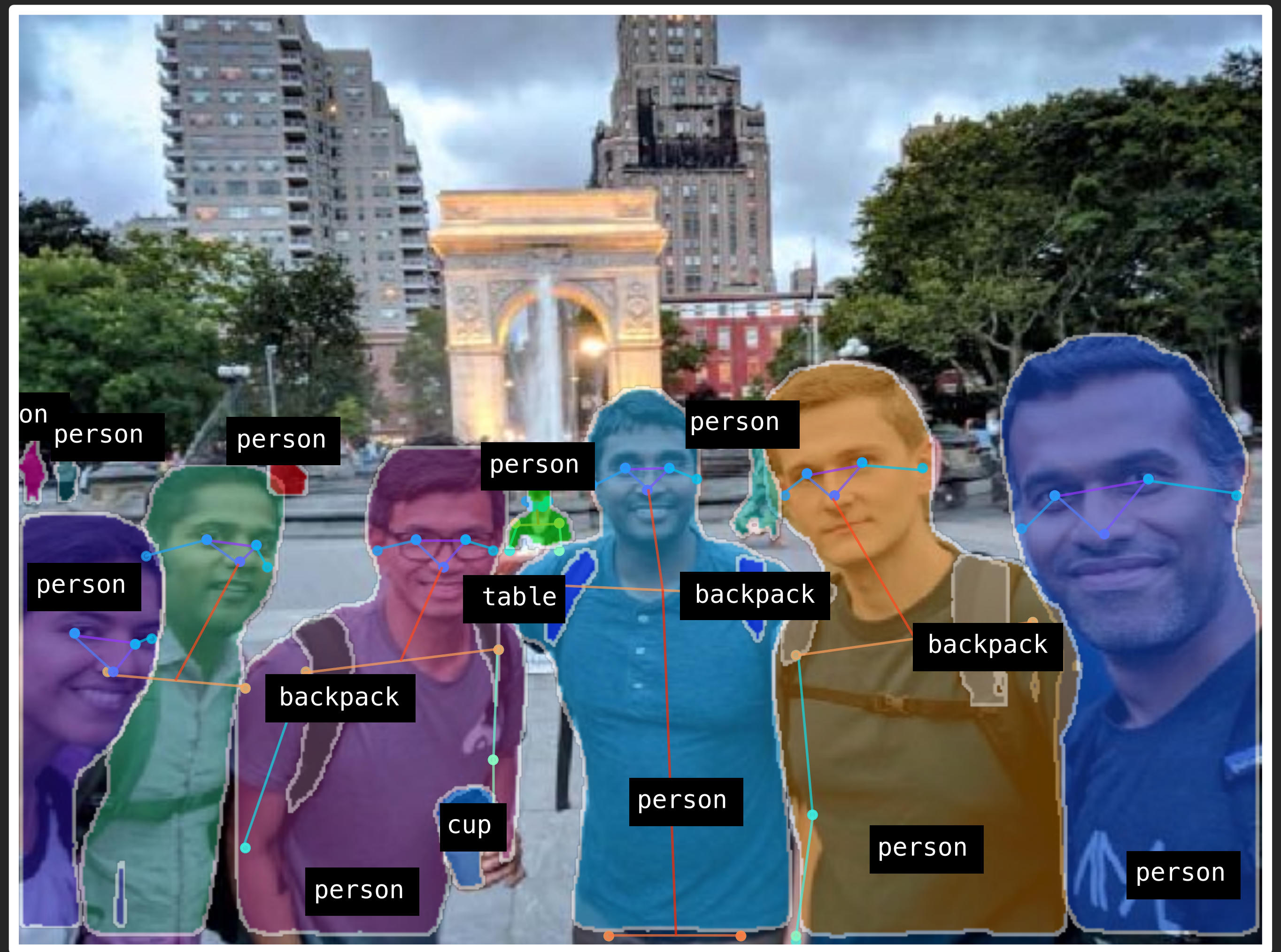
ALEXNET

2012



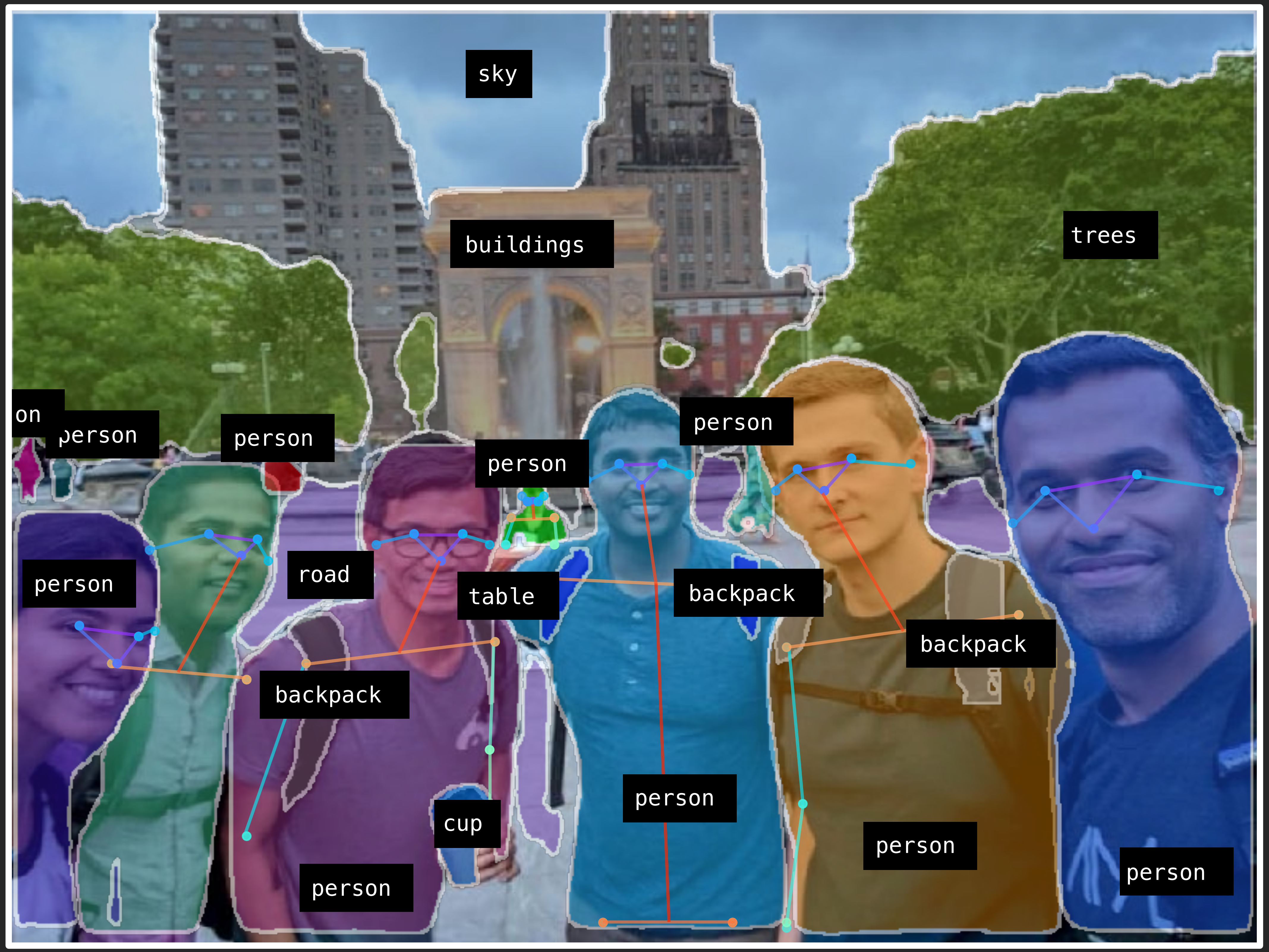
FASTER R-CNN

2015



MASK R-CNN

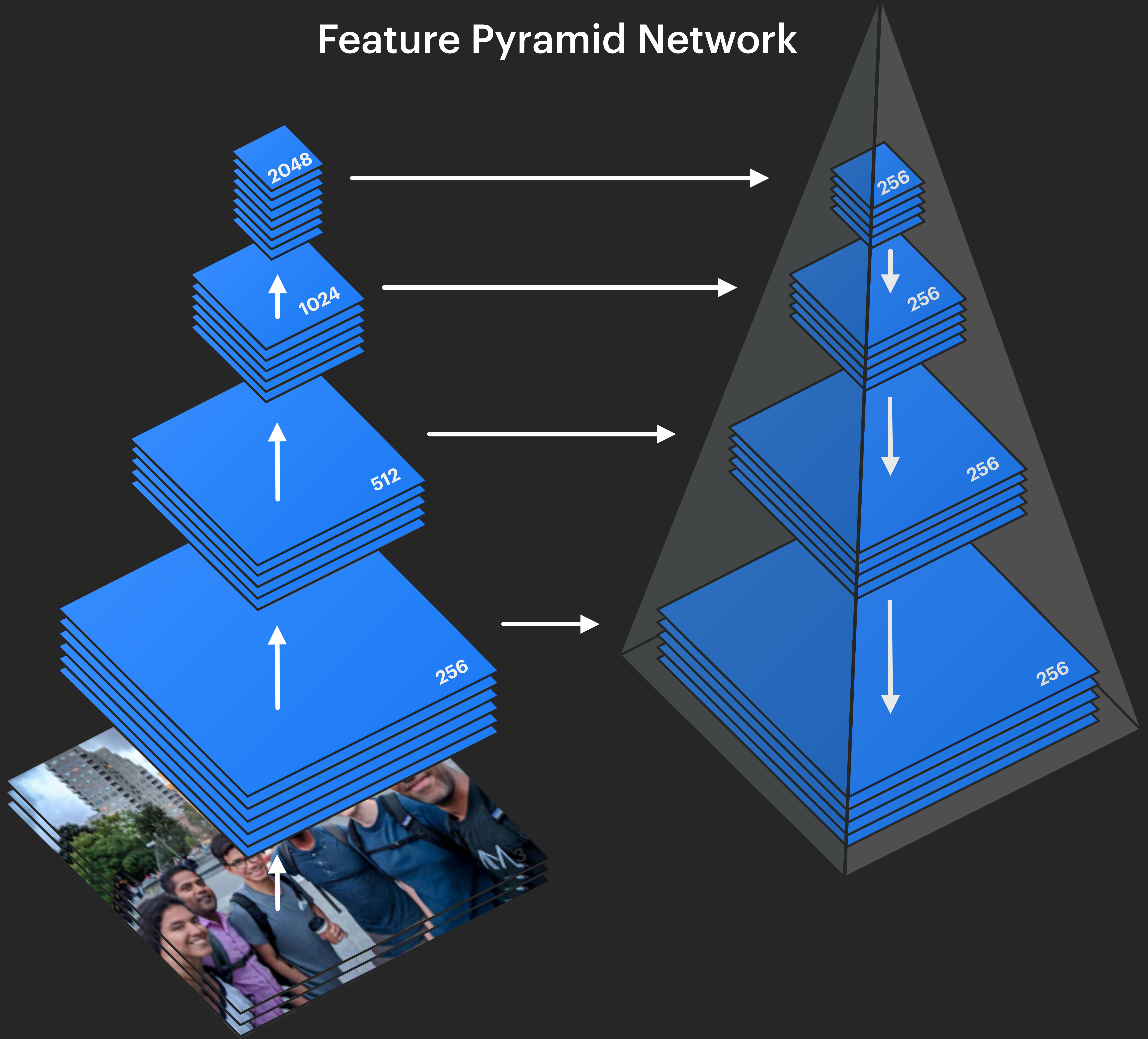
2017



PANOPTIC FPN

2019

Feature Pyramid Network



Feature Pyramid Network
2048
1024
512
256
256
256
256
256
Instance Segmentation Branch
network head
MASK R-CNN
Semantic Segmentation Branch
segment head 1/32
segment head 1/16
segment head 1/8
segment head 1/4
PANOPTIC FPN

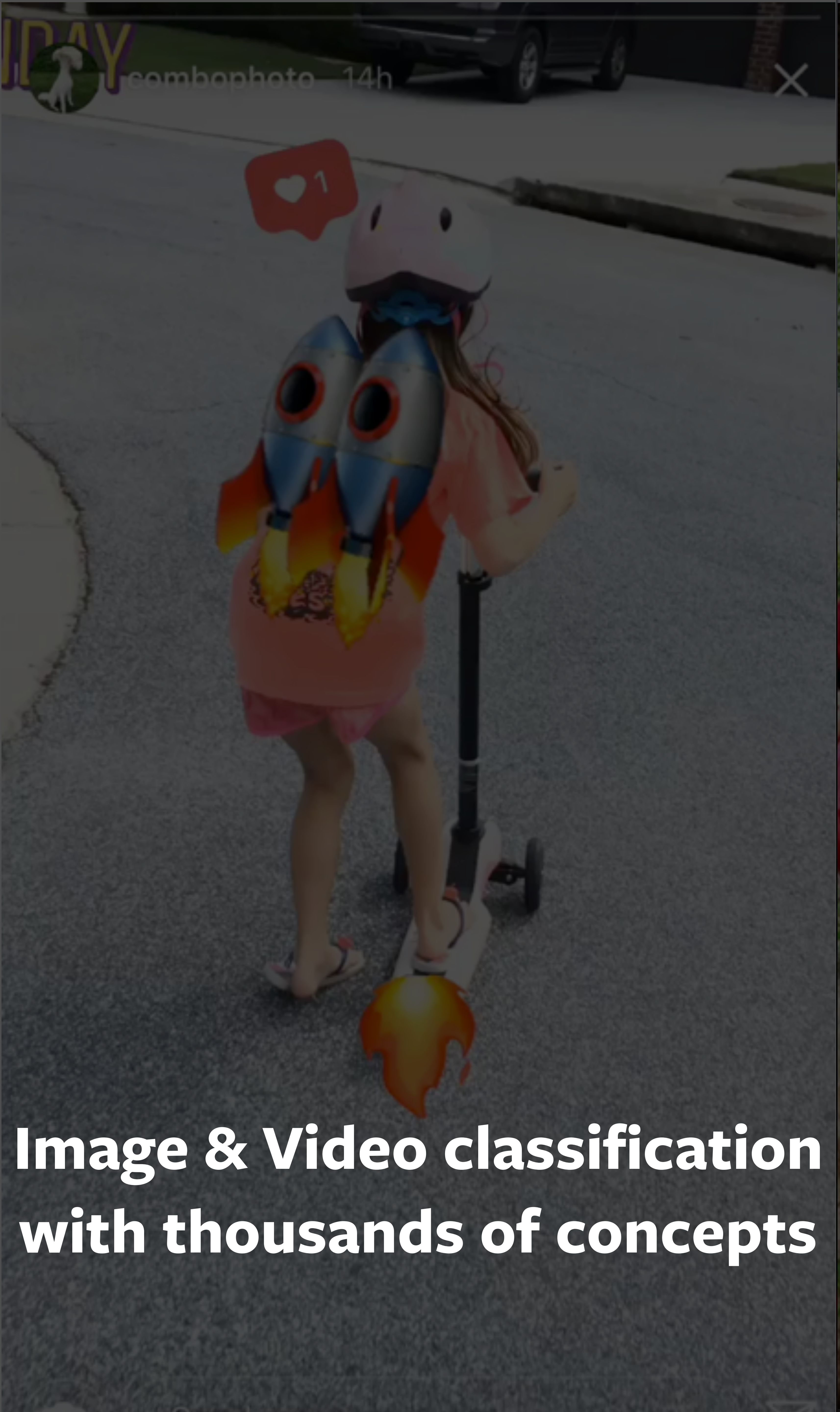
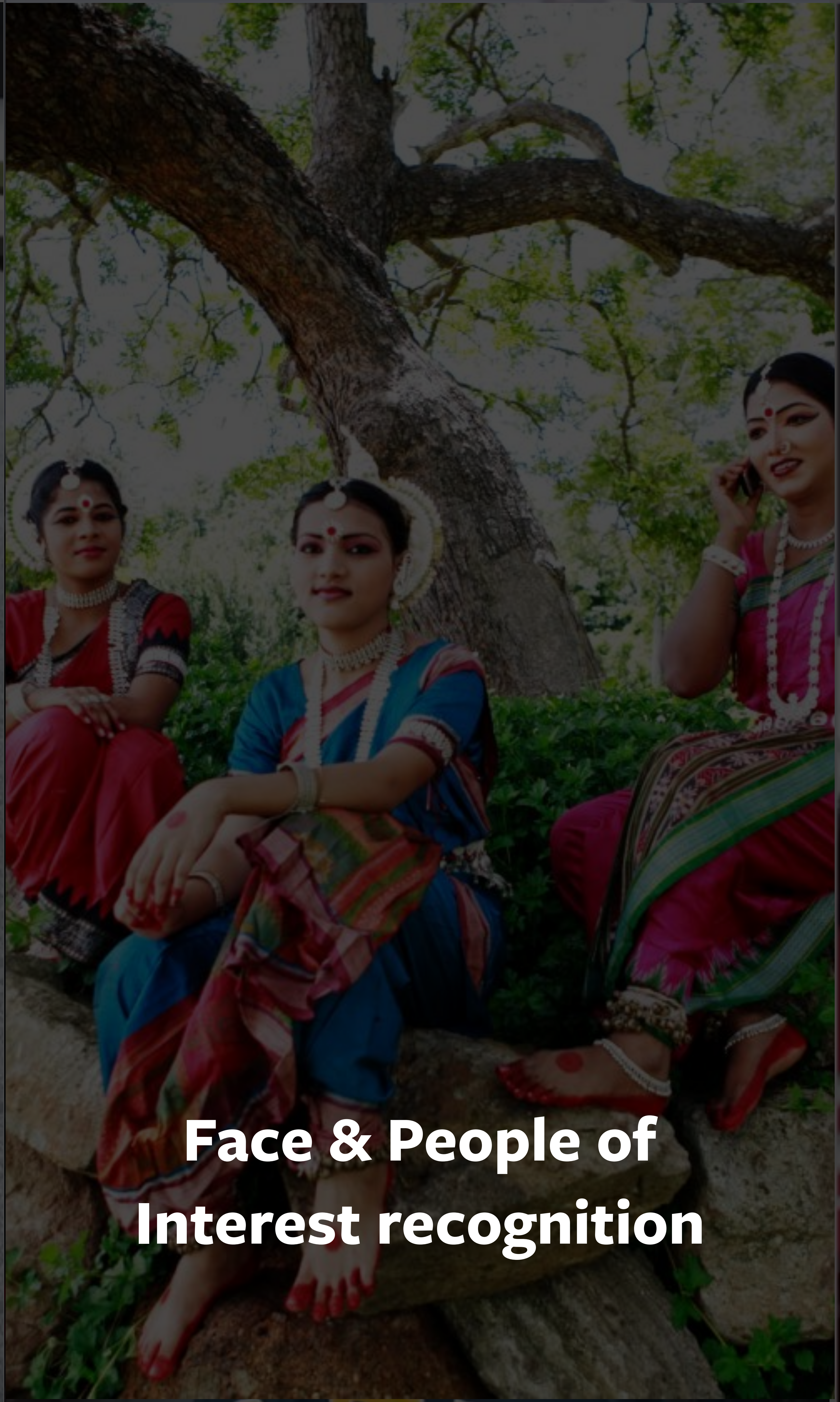


Image & Video classification with thousands of concepts



Face & People of Interest recognition

Text Detection & Recognition

Levels of Supervision


Levels of Supervision
Car
Tree
Fully-Supervised
Weakly-Supervised
Car
Tree
Semi-Supervised
Self-Supervised

Levels of Supervision

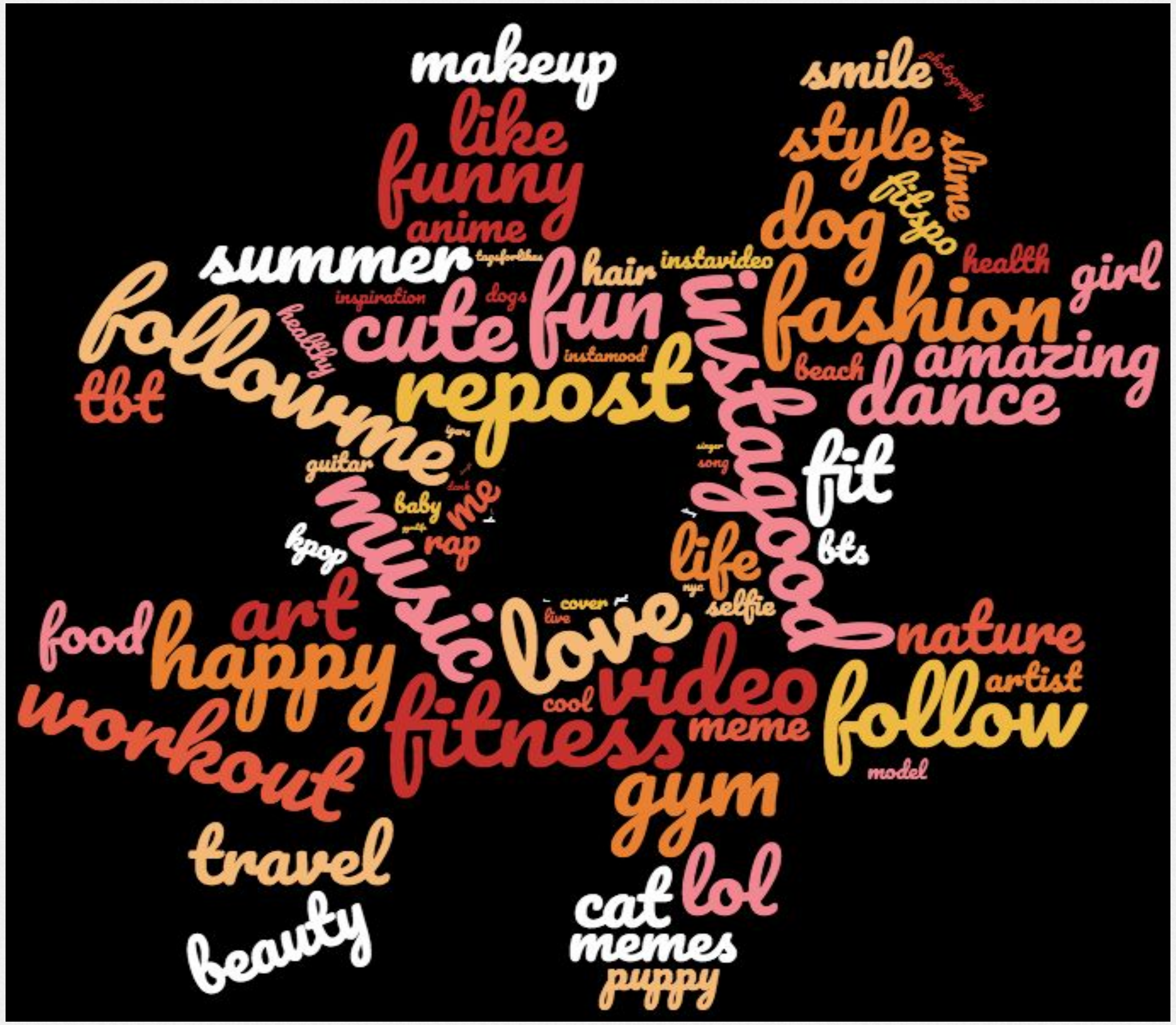
Levels of Supervision

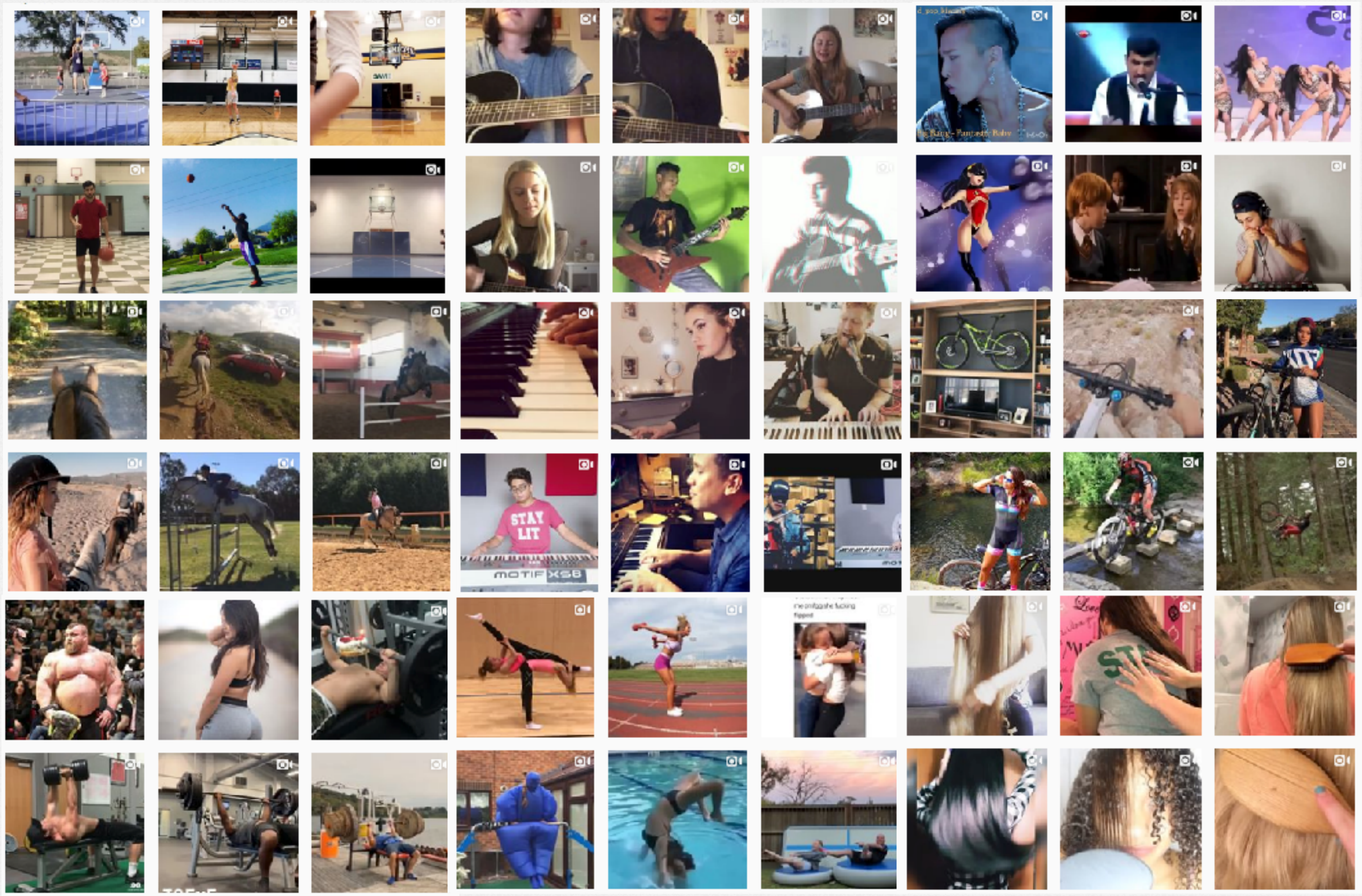
Levels of Supervision

Weakly Supervised Learning @ Billion Scale

Leverage large-scale, extremely noisy

hashtags for **weak-supervision**





Challenges of Training at Billion Scale

LEVELS OF SUPERVISION

Fully Supervised **Weakly Supervised** **Un-supervised**

CAT, DOG, WOODEN FLOOR

#CAT

???

A CUTE CAT COUPLE

Challenges of Training at Billion Scale

NOISY DATA

Non-Visual Labels → #LOVE #CAT ~~#DOG~~ #HUSKY ← **Incorrect Labels**

Missing Labels


Hello
From FB NY!
CHICHARRON
HAPPY

Large Weakly Supervised Training

BILLIONS OF
UNIQUE IMAGES

HUMUNGOUS
MODELS

THOUSANDS OF
LABELS

DISTRIBUTED
TRAINING

ImageNet in **one hour**

- ImageNet 1K has
  - 1.28 Million sample
  - 1000 categories
  - ResNet50 architecture
  - P100 GPUs
  - Caffe2

#machines	#gpus	Training time (mins)	Top-1 error
1	**8**	**1726.88**	**23.56**
2	16	905.22	23.35
4	32	464.23	23.48
8	64	231.42	23.39
16	128	117.76	23.58
32	**256**	**60.93**	**23.74**
36	288	54	23.76
40	320	48.84	24.08
44	352	44.36	24.23

Billion Scale Training at FB

IMAGENET-1K: STATE OF THE ART RESULTS

85.1%

OUR 3.5B TRAINING
RESNEXT101-32X32 MODEL

83.1%

PREVIOUS SOA

Before

Food

Food

Landmark

???

After

Cupcake

Apple pie

Statue of Liberty

Taj Mahal

WSL

IMAGENET

A simple, minimalist kitchen with very low lighting.

Black cat next to toy mouse on carpet.

WSL

A very modern bathroom with green glass tile work in the shower.

A cat sleeping next to a vehicle's tire on top of pavement.

PyTorch Models Are Available

https://pytorch.org/hub/facebookresearch_WSL-Images_resnext/

```
import torch
model = torch.hub.load('facebookresearch/WSL-Images', 'resnext101_32x8d_wsl')
# or
# model = torch.hub.load('facebookresearch/WSL-Images', 'resnext101_32x16d_wsl')
# or
# model = torch.hub.load('facebookresearch/WSL-Images', 'resnext101_32x32d_wsl')
# or
#model = torch.hub.load('facebookresearch/WSL-Images', 'resnext101_32x48d_wsl')
model.eval()
```

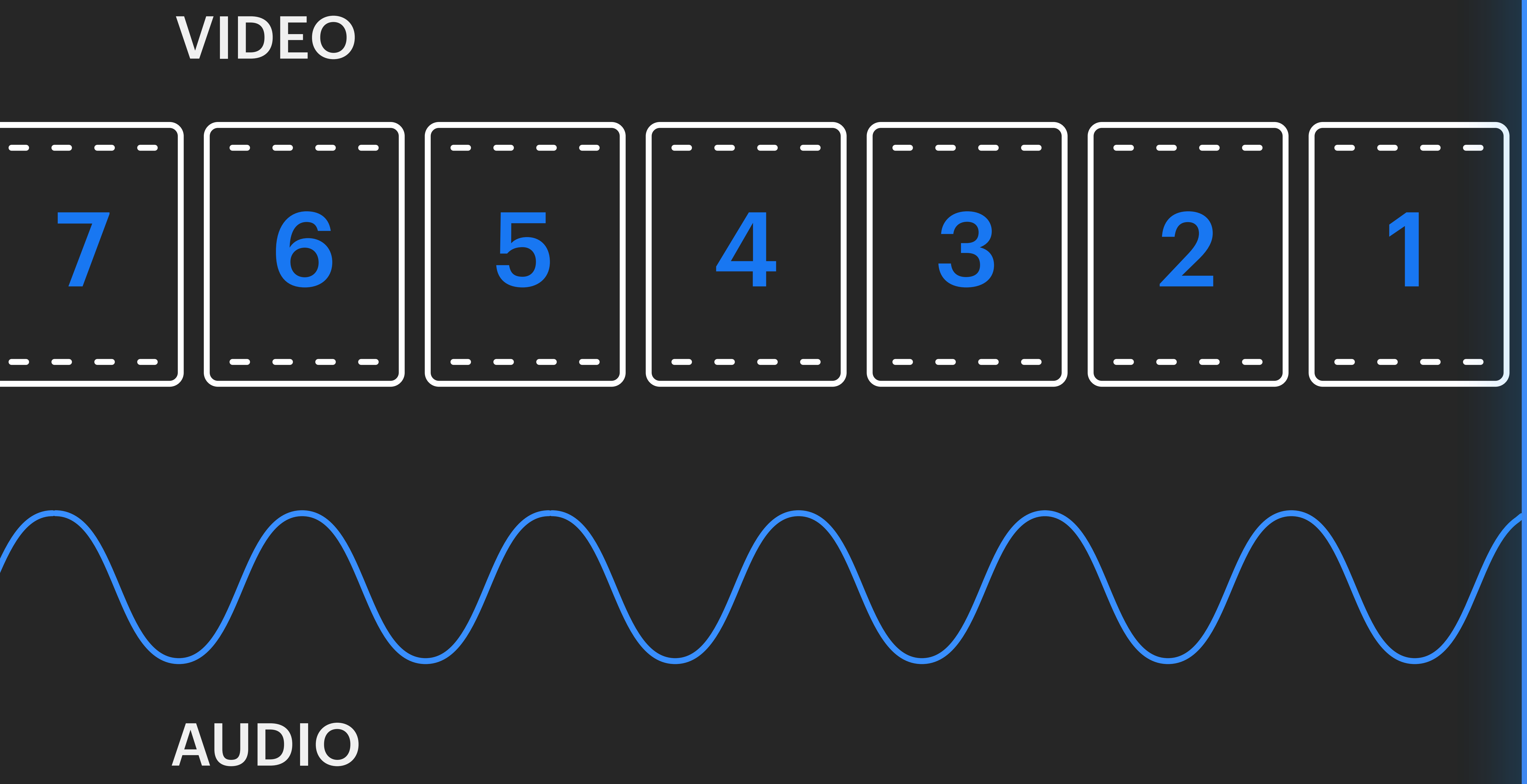
How about Videos?

ENRICH OUR USER'S EXPERIENCE

Did We Miss a Great Moment?

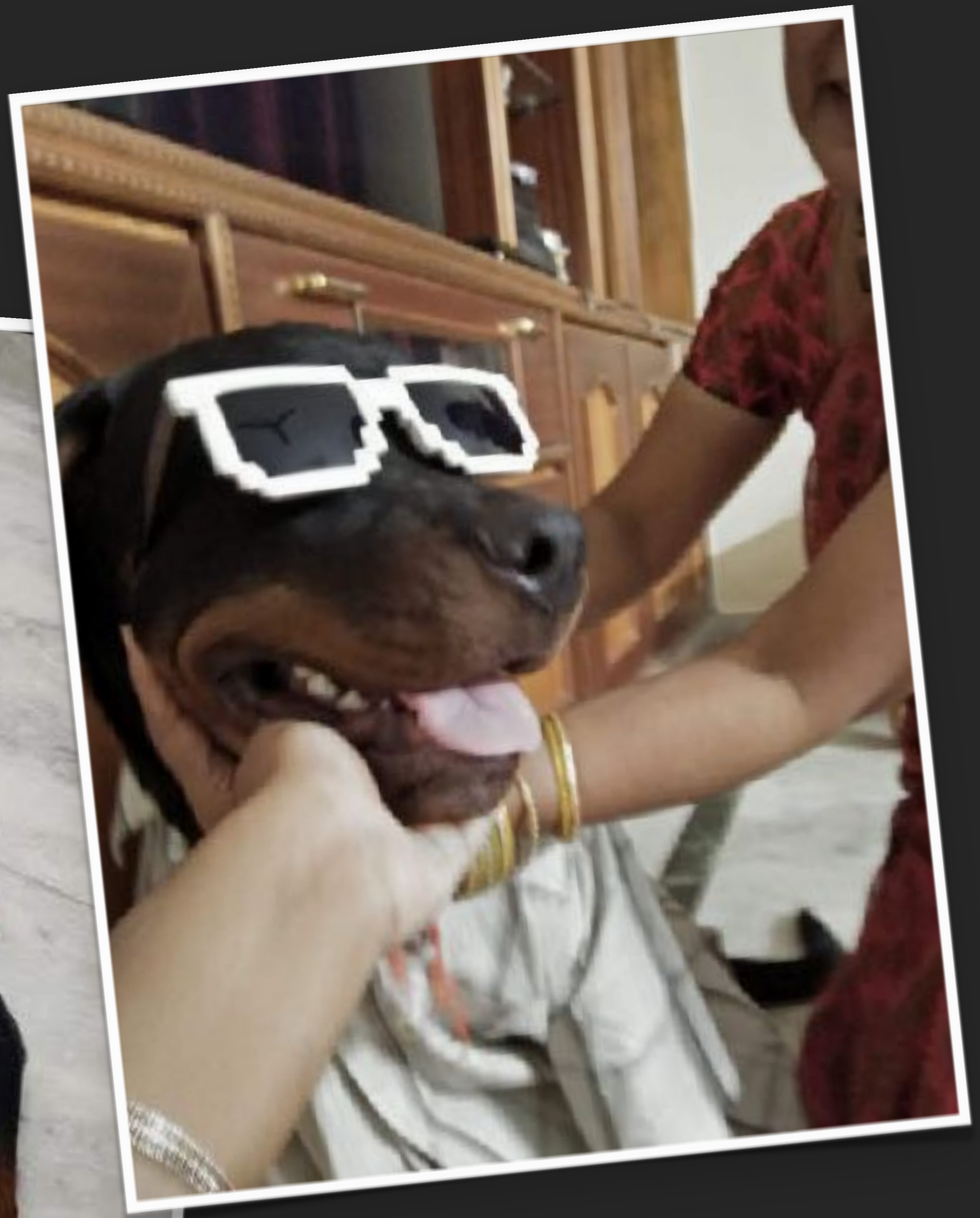
Spatiotemporal Visual Modeling

SPACE AND TIME HAVE DIFFERENT STATISTICS


VIDEO
7
6
5
4
3
2
1
AUDIO


ACTION OF INTEREST

INFORMATION REDUNDANCY
TEMPORAL NOISE



dog 0.99 door 0.8 wag 0.5 walk 0.8

stick 0.7 pickup 0.7 drop 0.8 tail 0.7

Sampling Salient Clips

Spatiotemporal CNN

dog 0.99	door 0.8	wag 0.5	walk 0.8
stick 0.7	pickup 0.7	drop 0.8	tail 0.7

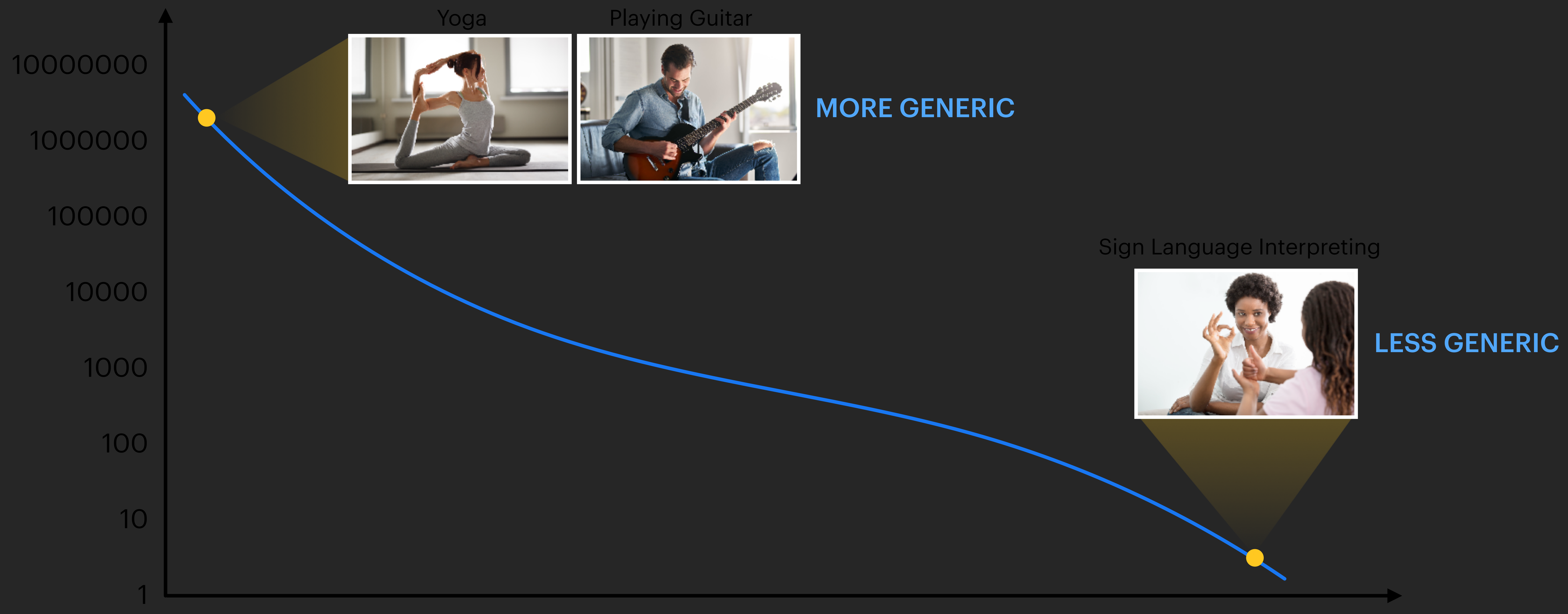
Levels of Supervision

Leveraging Rich Metadata

HASHTAGS

Challenges With Training at Scale

SKEWED (LONG-TAIL) DISTRIBUTION



65M

PUBLIC VIDEOS

65M

PUBLIC VIDEOS

6M

PUBLIC VIDEOS

YouTube-8M Dataset

110M

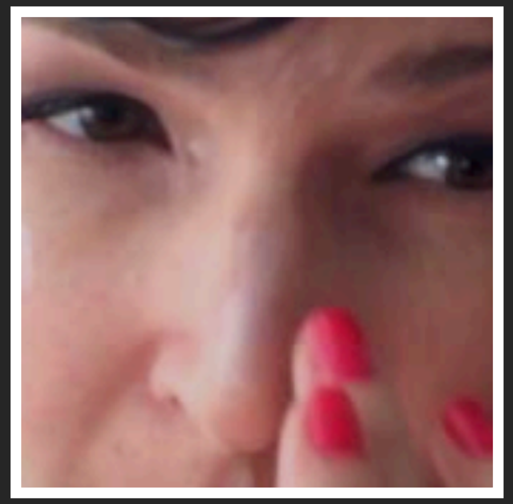
PARAMETERS

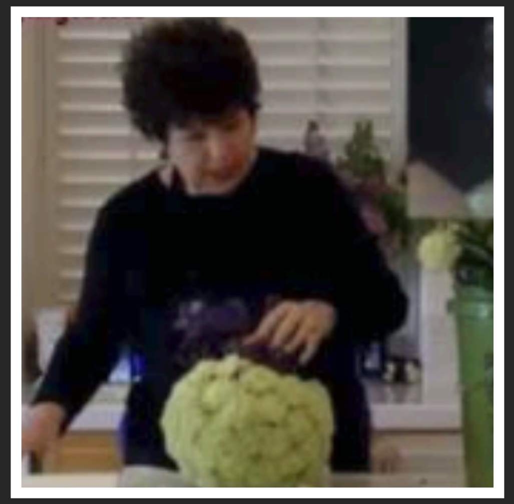
1280

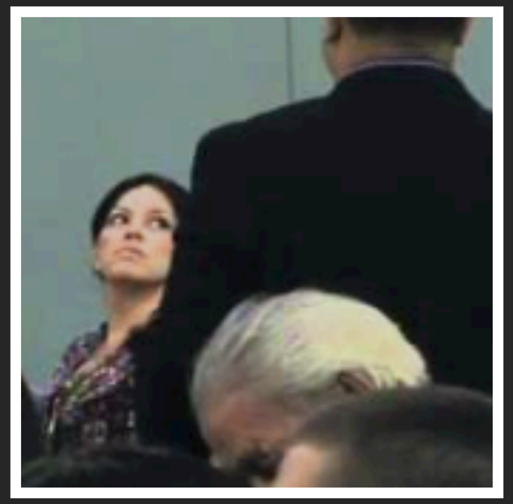
GPUs

State of the Art Results

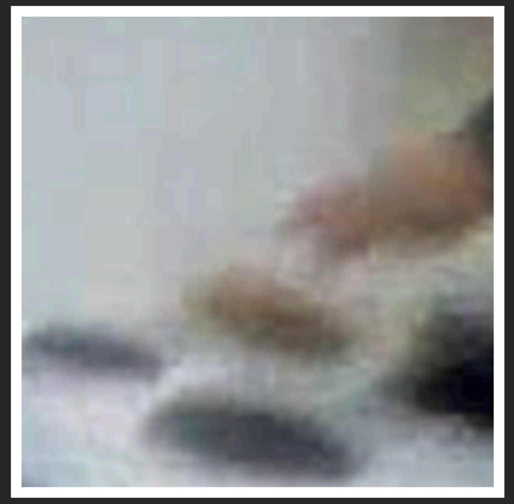

SPEAKING

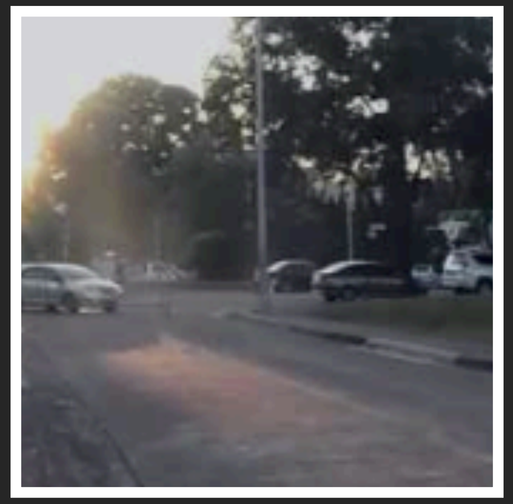

APPLYING CREAM

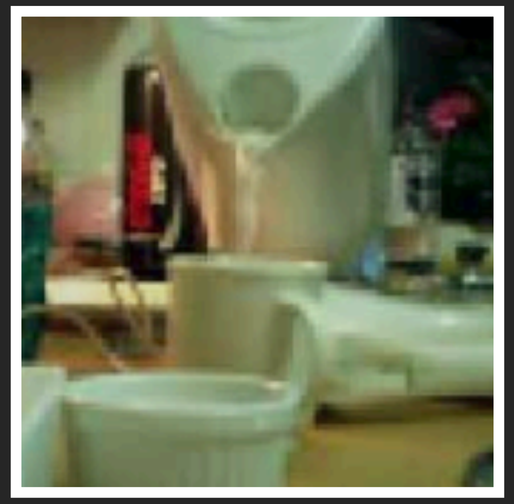

ARRANGING FLOWERS


AUCTIONING

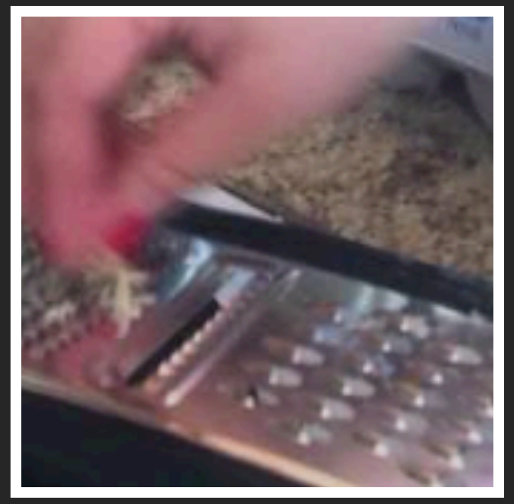

BIKING THROUGH SNOW


FLIPPING PANCAKE


JOGGING


MAKING TEA


MOWING


SCRAMBLING EGGS

KINETICS: 300K VIDEOS, 400 ACTIONS

Y. Bian et.al, ArXiv 2017

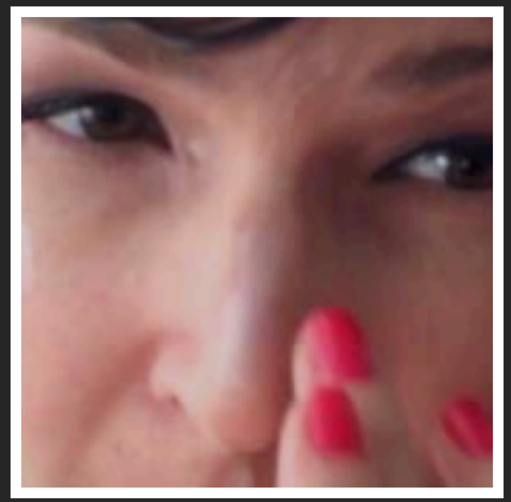
Metric: Top-1 Accuracy

77.7%

PREVIOUS SOA

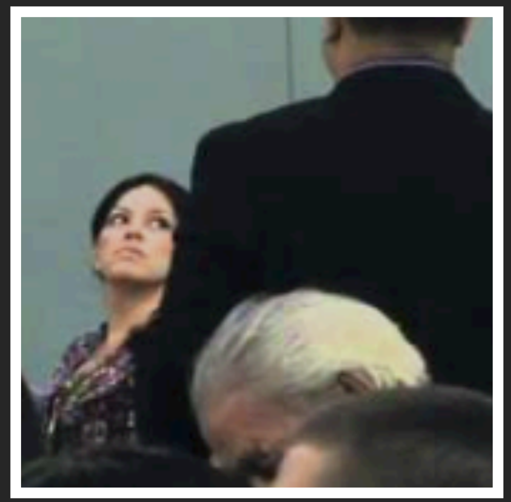
State of the Art Results

SPEAKING



APPLYING CREAM

ARRANGING FLOWERS

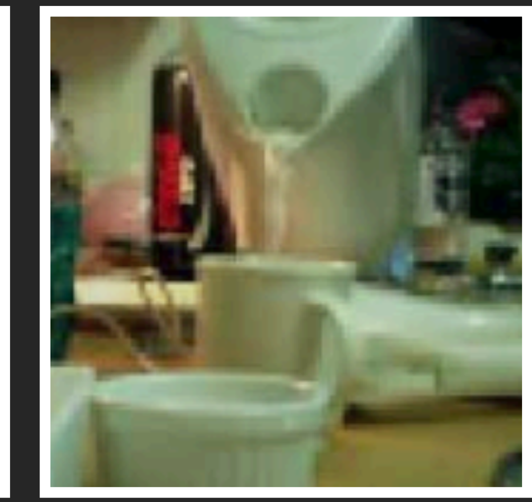


AUCTIONING

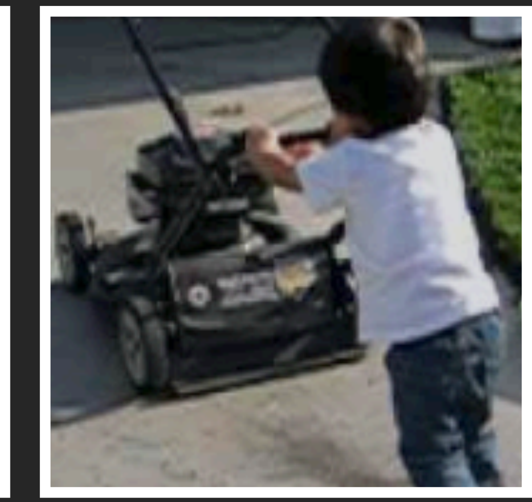
BIKING THROUGH SNOW

FLIPPING PANCAKE

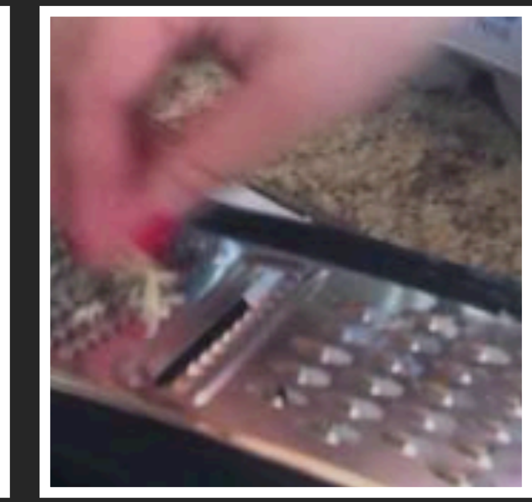
JOGGING



MAKING TEA



MOWING



SCRAMBLING EGGS

KINETICS: 300K VIDEOS, 400 ACTIONS

Metric: Top-1 Accuracy

77.7%

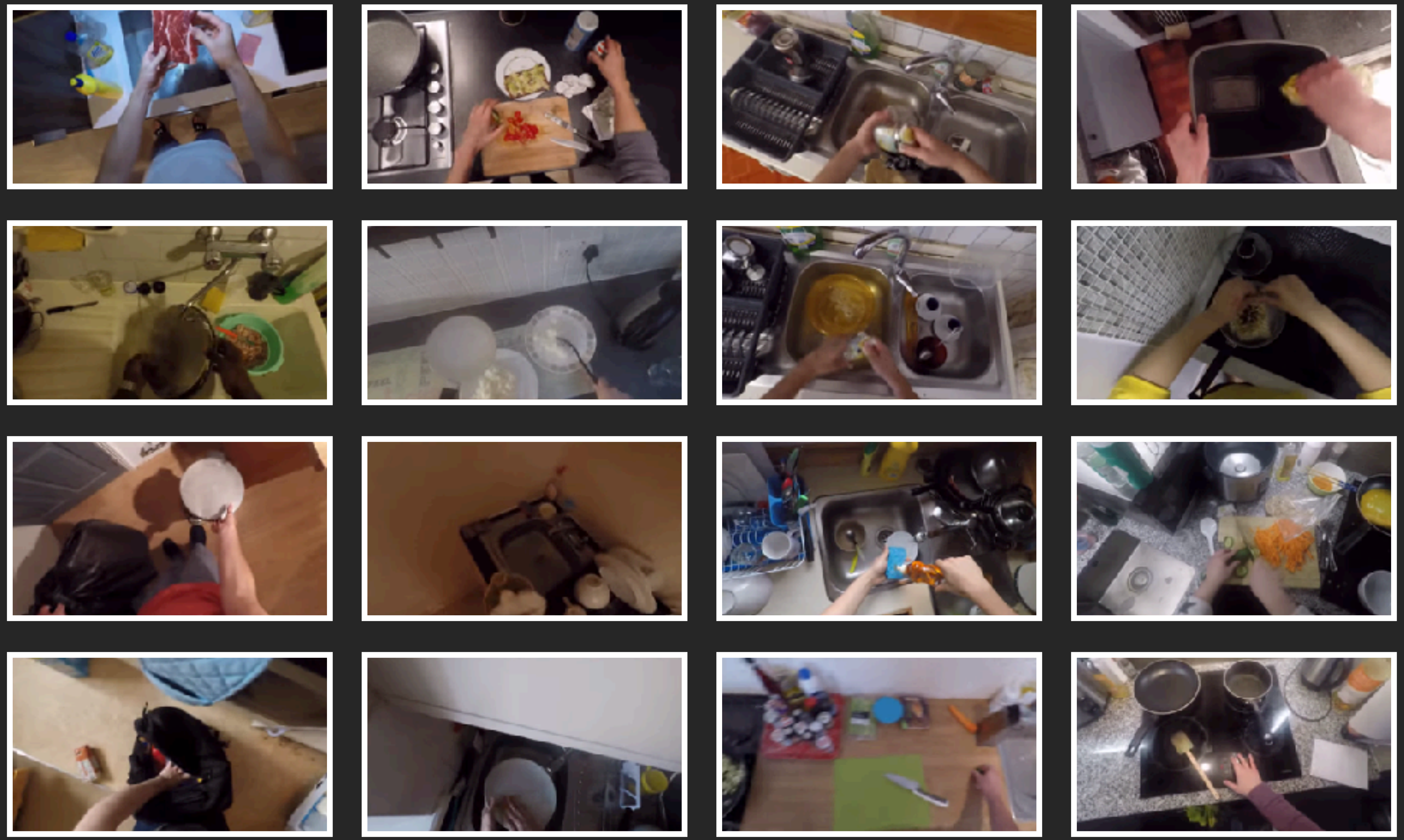
PREVIOUS SOA

+5.1%

82.8%

OUR 65M TRAINING

State of the Art Results



EPIC-KITCHENS: 28K VIDEOS, 2337 ACTIONS

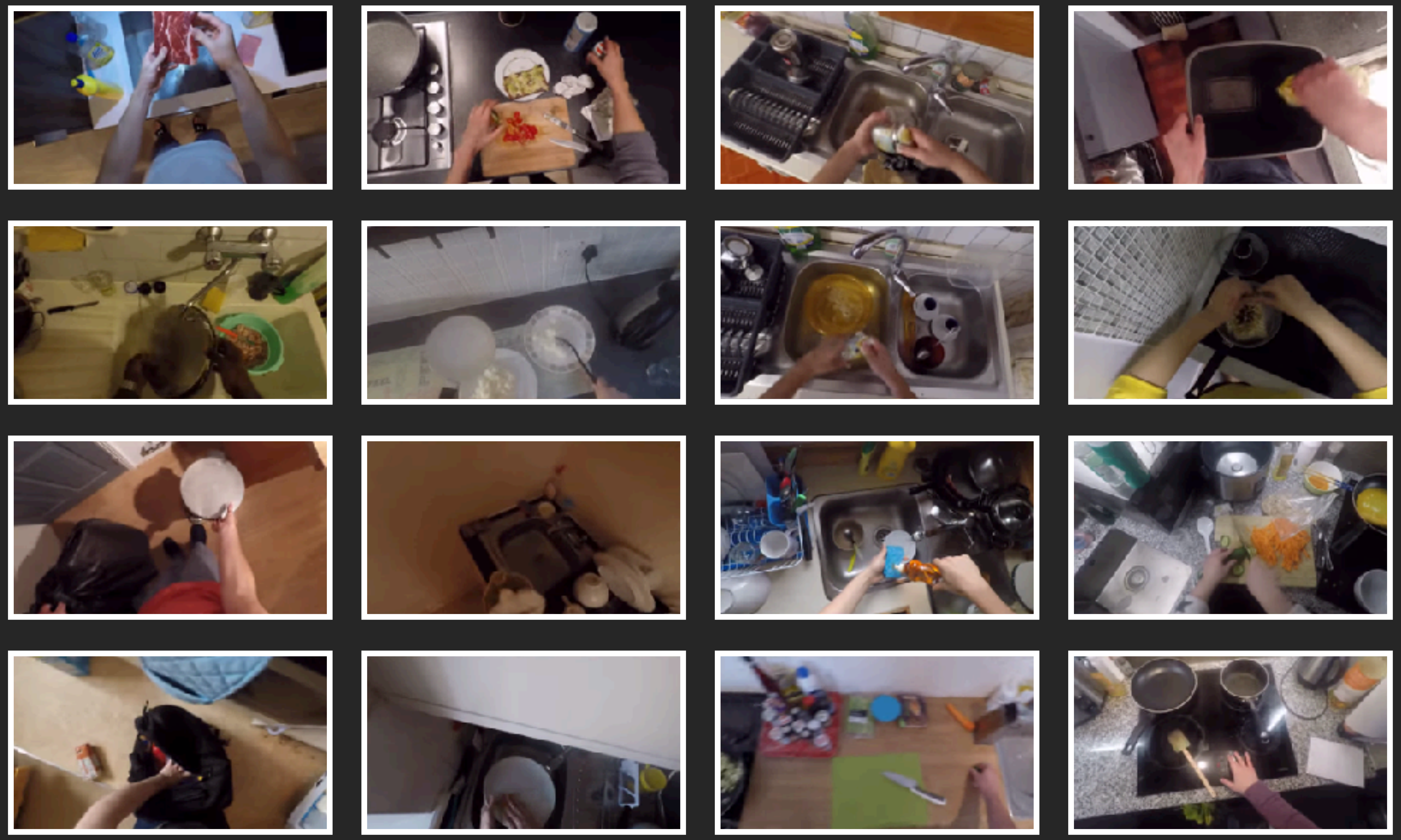
EPIC-Kitchens Action Recognition Challenge

Metric: Top-1 Accuracy

21.0%

PREVIOUS SOA

State of the Art Results



EPIC-KITCHENS: 28K VIDEOS, 2337 ACTIONS

Metric: Top-1 Accuracy

21.0%

PREVIOUS SOA

+4.6%

25.6%

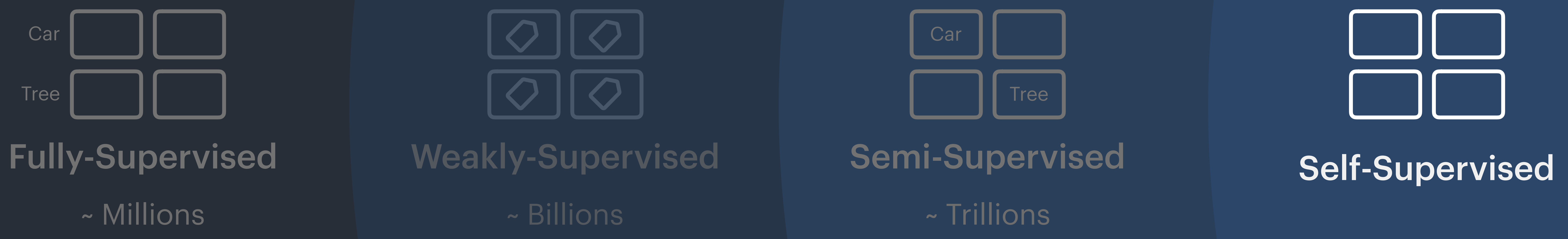
OUR 65M TRAINING

References for the video understanding efforts

- SCSampler: Sampling Salient Clips from Video for Efficient Action Recognition - https://arxiv.org/abs/1904.04289
- Video Classification with Channel-Separated Convolutional Networks - https://arxiv.org/abs/1904.02811
- Large-scale weakly-supervised pre-training for video action recognition - https://arxiv.org/abs/1905.00561

Pushing State of the Art helps the world in significant ways

Levels of Supervision



From Satellite Imagery to Disaster Insights

Jigar Doshi[1], Saikat Basu[2], Guan Pang[2]

CrowdAI[1], Facebook[2]

What's the Research?

- A framework for using convolutional neural networks (CNNs) on satellite imagery to identify the areas most severely affected by a disaster. This new method potentially produces **more accurate** information in **far less time** than current manual methods.

- The goal of this work is to allow rescue workers to quickly identify where aid is needed most, **without relying on manually annotated, disaster-specific data sets.**

Disaster Impact Index (DII)

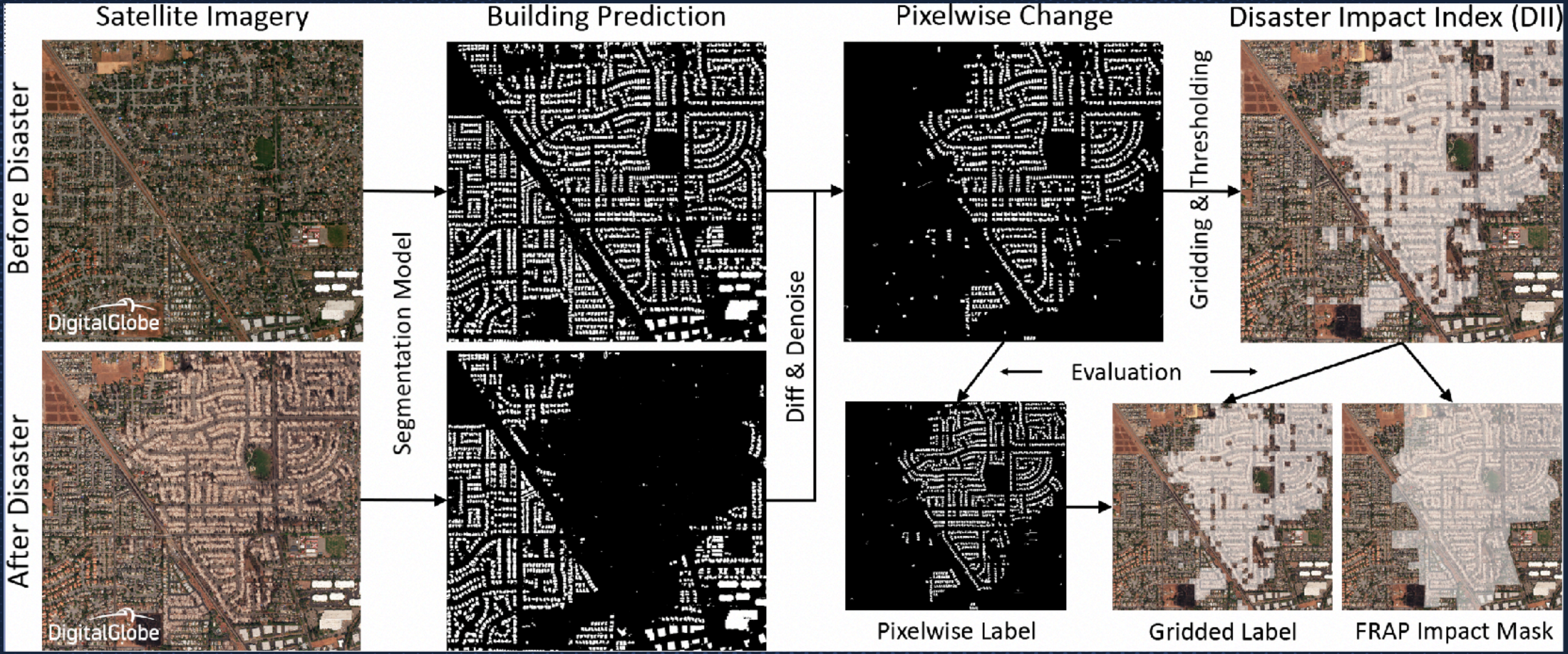
$$DII = \Delta Pred = \frac{\eta Pred_{before=1 \& Pred_{after}=0}\, grid}{\frac{1}{N_{grid}} \sum_{i=1}^{N_{grid}} \eta Pred_{before=1}\, grid_i}$$

$\eta Pred_{before=1 \& Pred_{after}=0}\, grid \rightarrow$ Pixels missing the feature post disaster

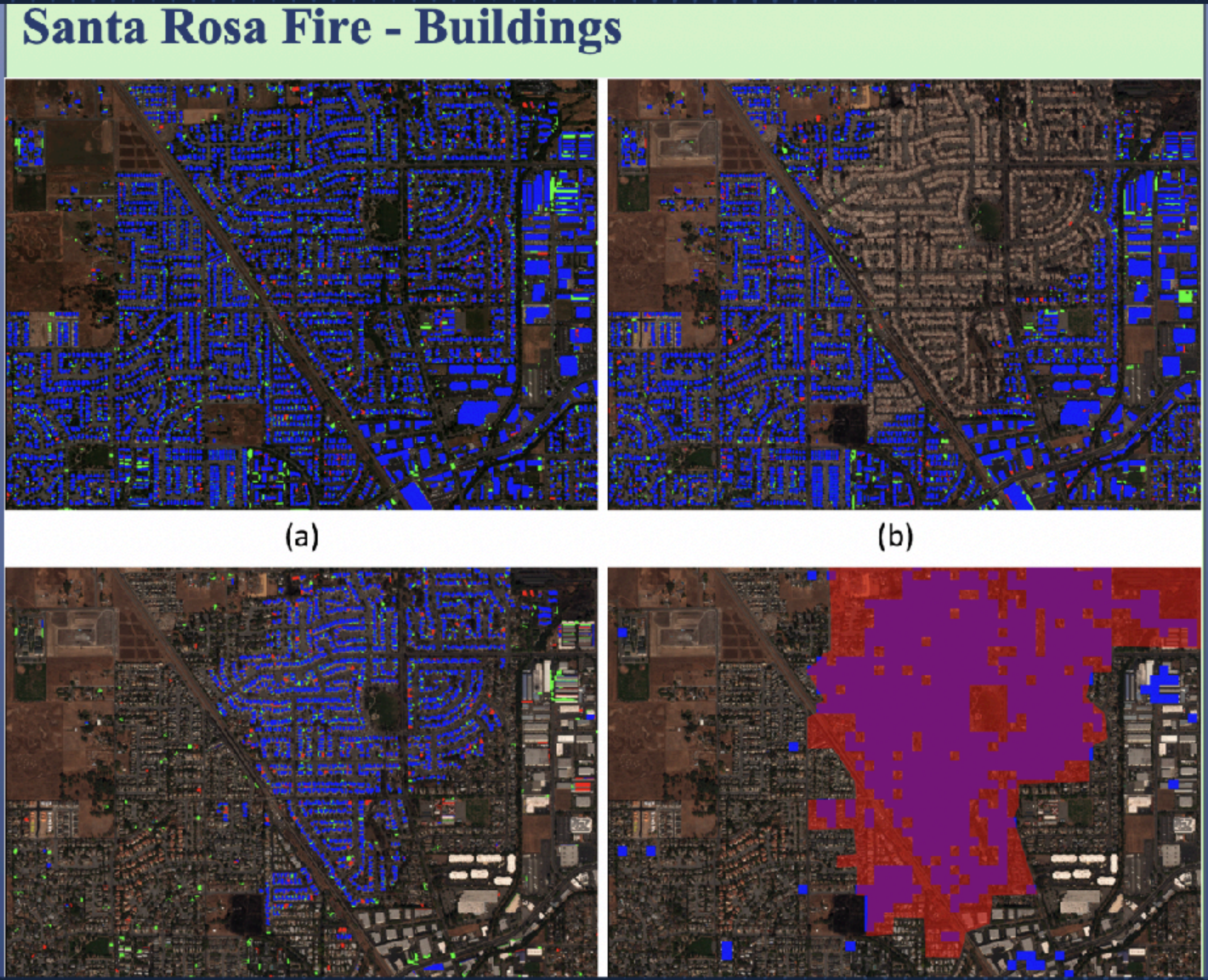
$\frac{1}{N_{grid}} \sum_{i=1}^{N_{grid}} \eta Pred_{before=1}\, grid_i \rightarrow$ num of feature pixels predicted pre-disaster

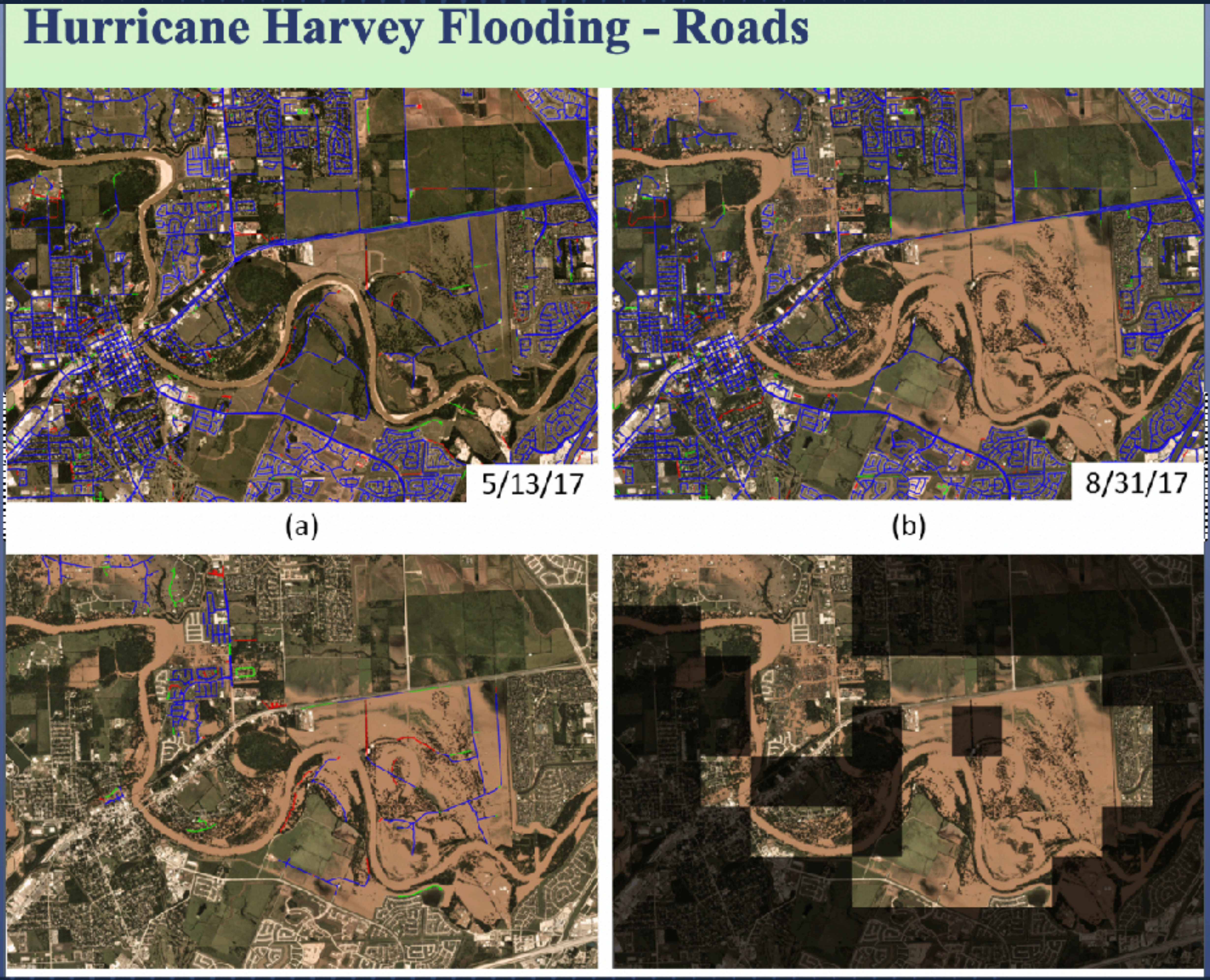
$N_{grid} \rightarrow$ total number of grids in this case 256×256

How it works!


Satellite Imagery
Building Prediction
Pixelwise Change
Disaster Impact Index (DII)
Before Disaster
After Disaster
DigitalGlobe
DigitalGlobe
Segmentation Model
Diff & Denoise
Gridding & Thresholding
Evaluation
Pixelwise Label
Gridded Label
FRAP Impact Mask

Helping in real world scenarios





1st CVPR Workshop on

COMPUTER VISION FOR GLOBAL CHALLENGES

16/17 June 2019

Long Beach, CA

Computer Vision for Global Challenges

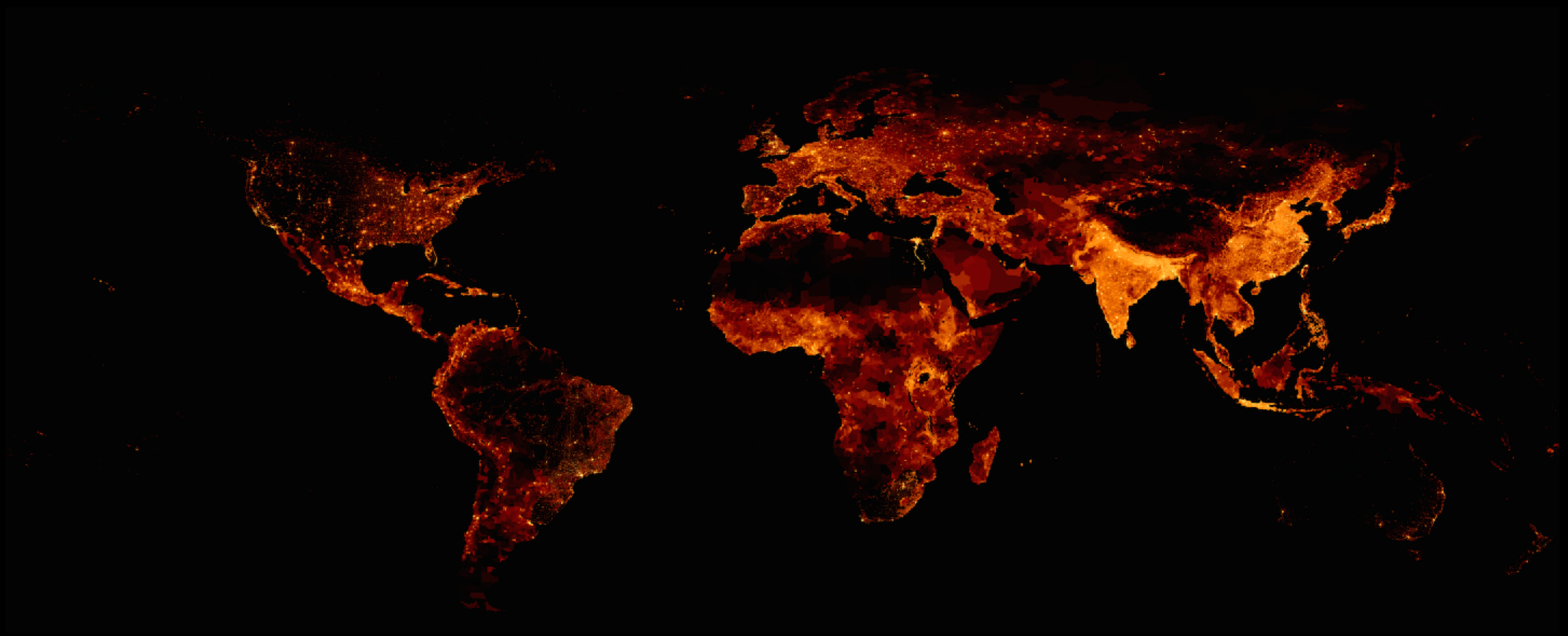
@cv4gc

Do you have an idea for a computer vision task that would impact the lives of many? Have you identified the limitations of a vision technique because of the geographical bias of the data you are using it on? Is there an application of computer vision that would be helpful to your community? Are you looking for potential vision expert partners or feedback on your idea?

Apply for the Call of Challenges, and come and experience the premier computer vision conference, and participate in an active discussion with the top vision researchers!

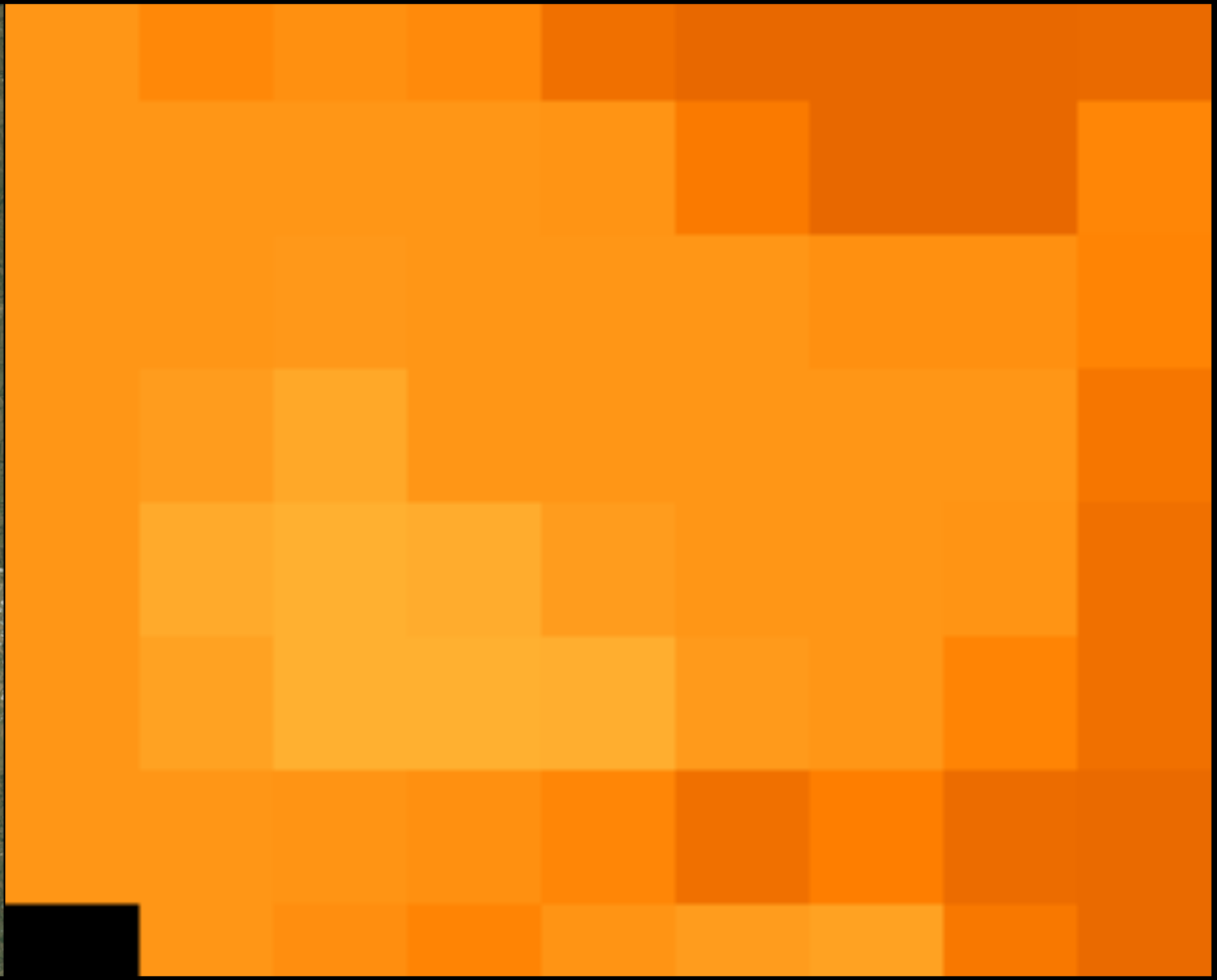
More info at: http://www.cv4gc.org/#challenges

Population Density Estimation

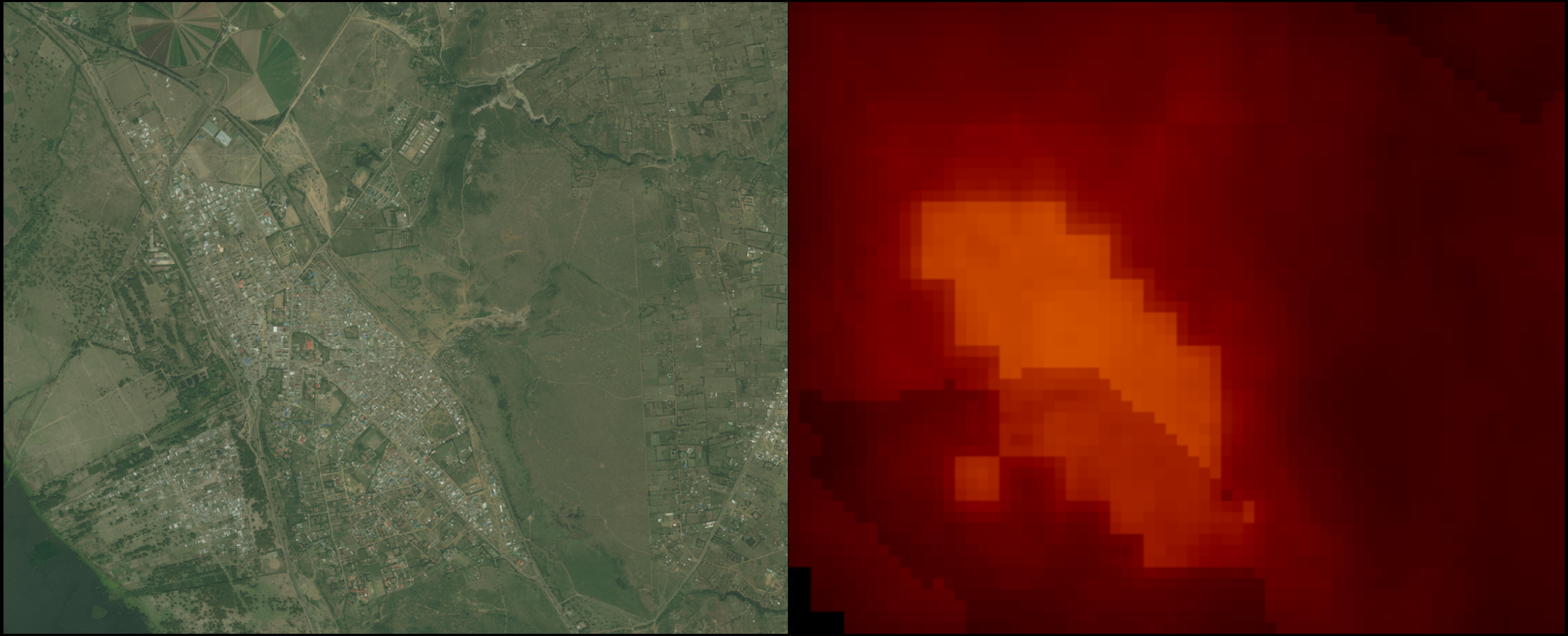


Naivasha in Kenya

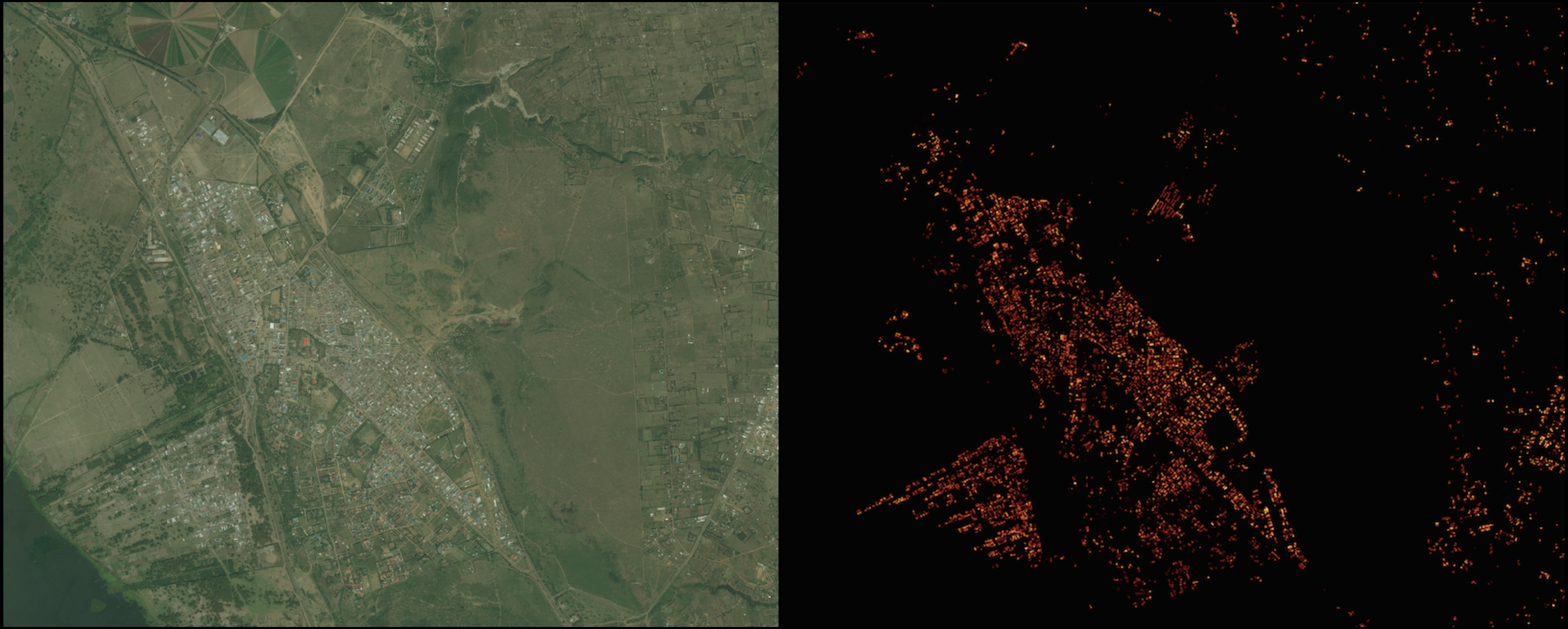
GPWv4 from CIESIN at Columbia University



WorldPop



Facebook



What will you learn today?

How do you design a image and video recognition system for billion scale?

Can you remove the requirement of annotation to learn best representations?

Can we understand video faster than understanding individual frames?

How does pushing state of the art in CV make a meaningful difference to everyone in the world?

Thank You!