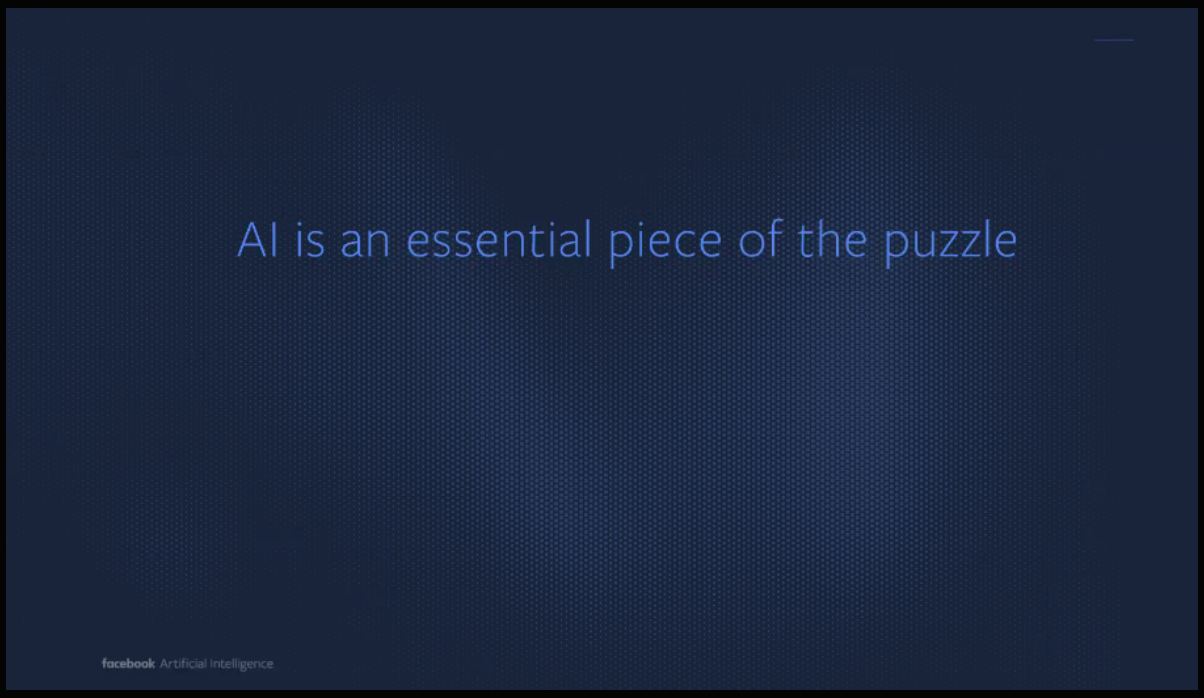
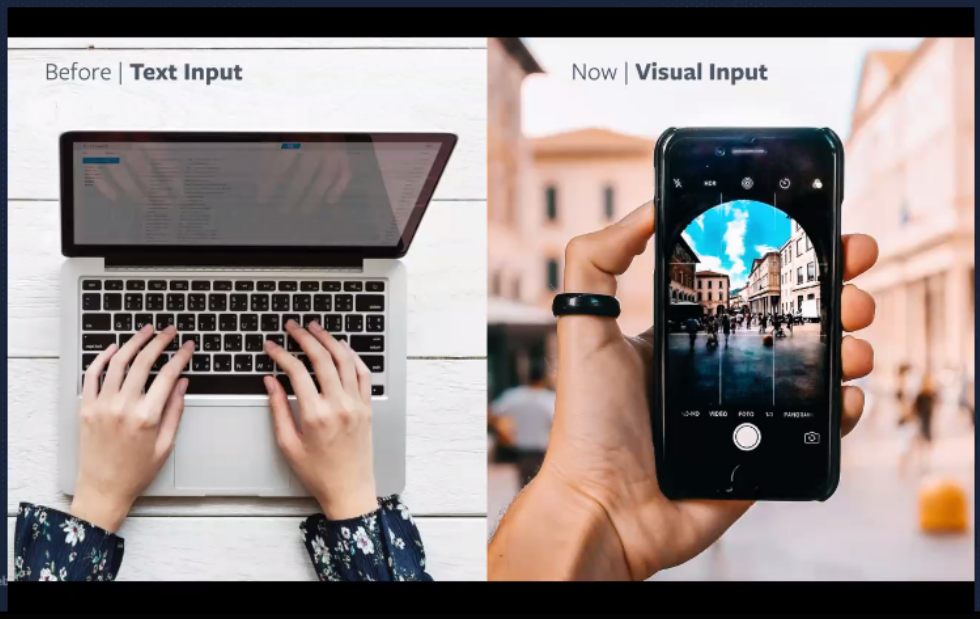
Day 18

Speaker: Manohar Paluri, Facebook AI Research

Title: Pushing the Limits of Machine Perception.



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



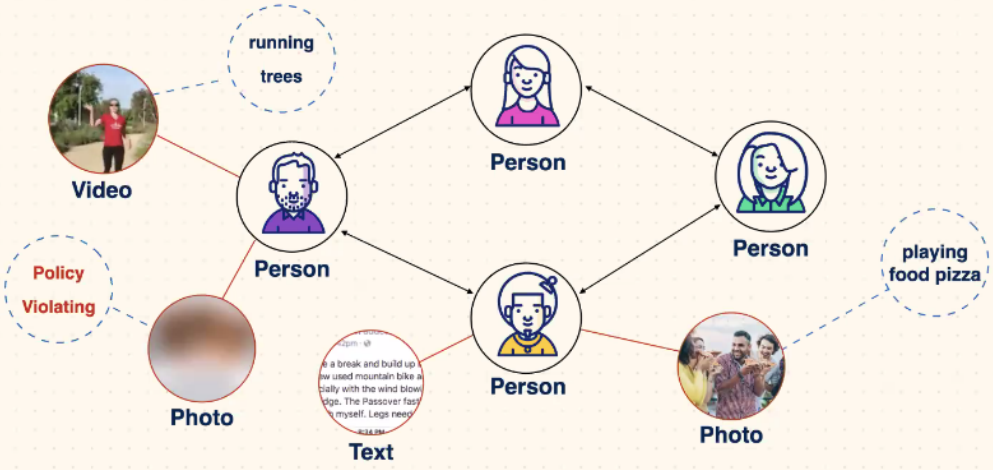
What will you learn today?

How to design a billion scale image and video recognition system?

Three latest works from ICCV:

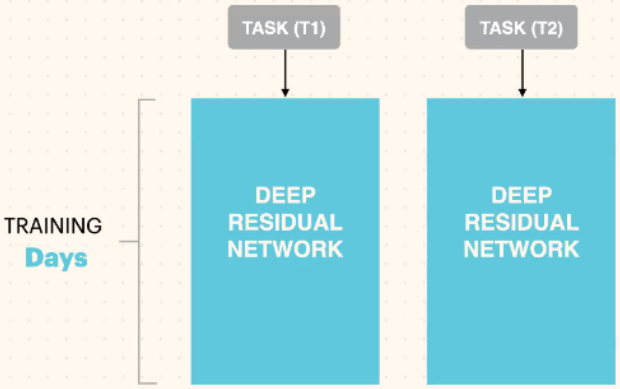
- Weakly supervised Pretraining for Detection
- Unidentified objects in video - benchmark for open world segmentation
- Generic event boundary detection - benchmark for event segmentation

Social Graph -> Semantic Graph

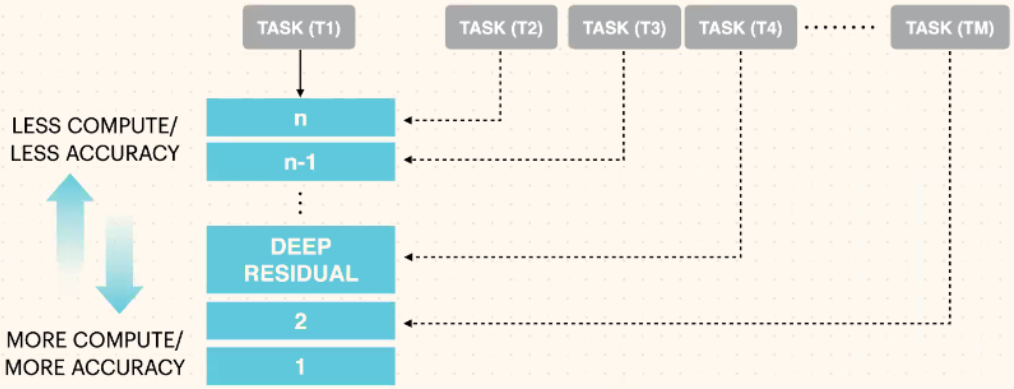


Vision Models In Production

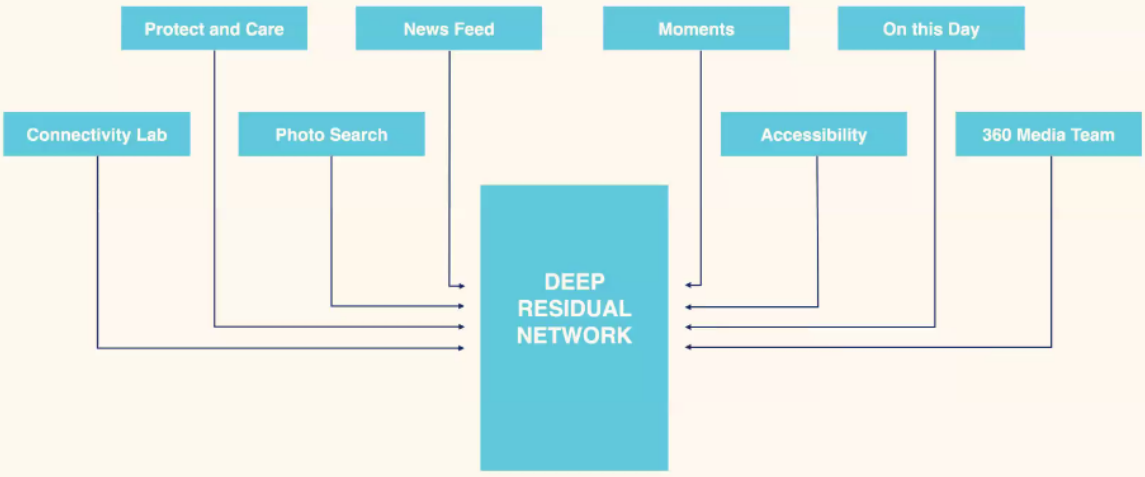
- Multiple vision tasks need to be done
- Cannot afford one separate model for each task
- Explosion in computations cost and resources



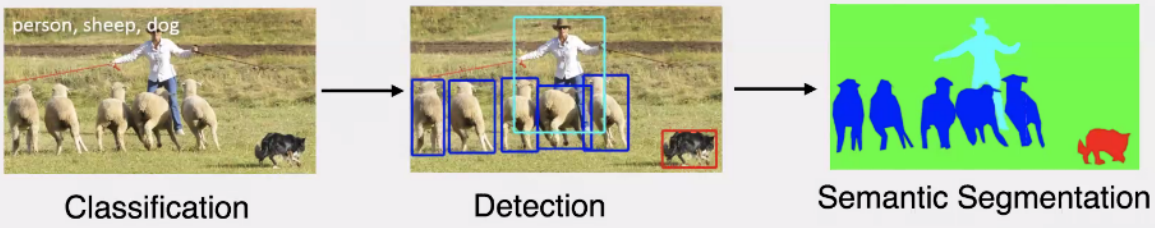
Our Vision - Towards Universal Vision Model



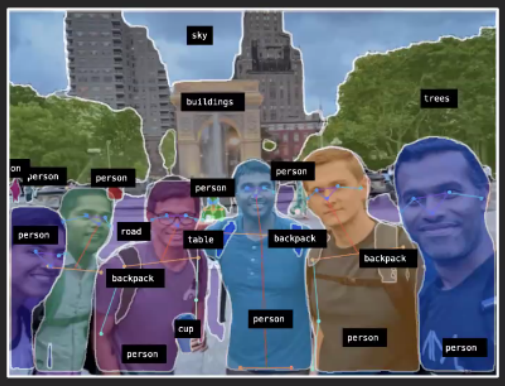
Our work allows to move the tasks towards upper layers



Progress in Image Understanding



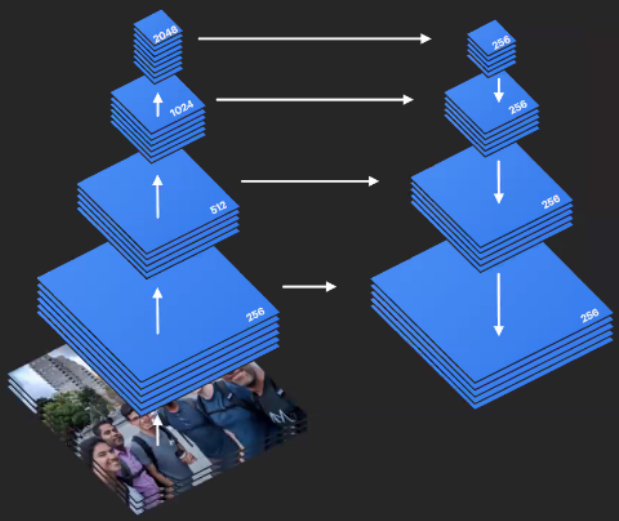
Classification → Detection → Semantic Segmentation → Instance Segmentation → Attributes → Relationships, Fine grained, Aesthetics, . .

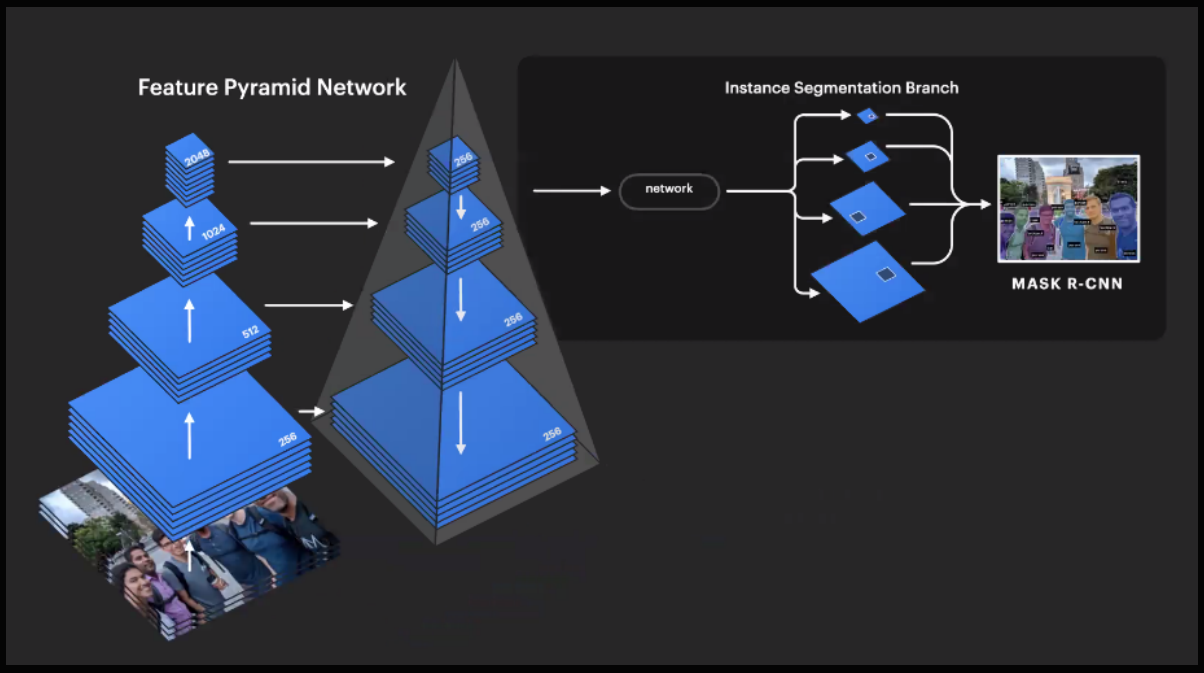


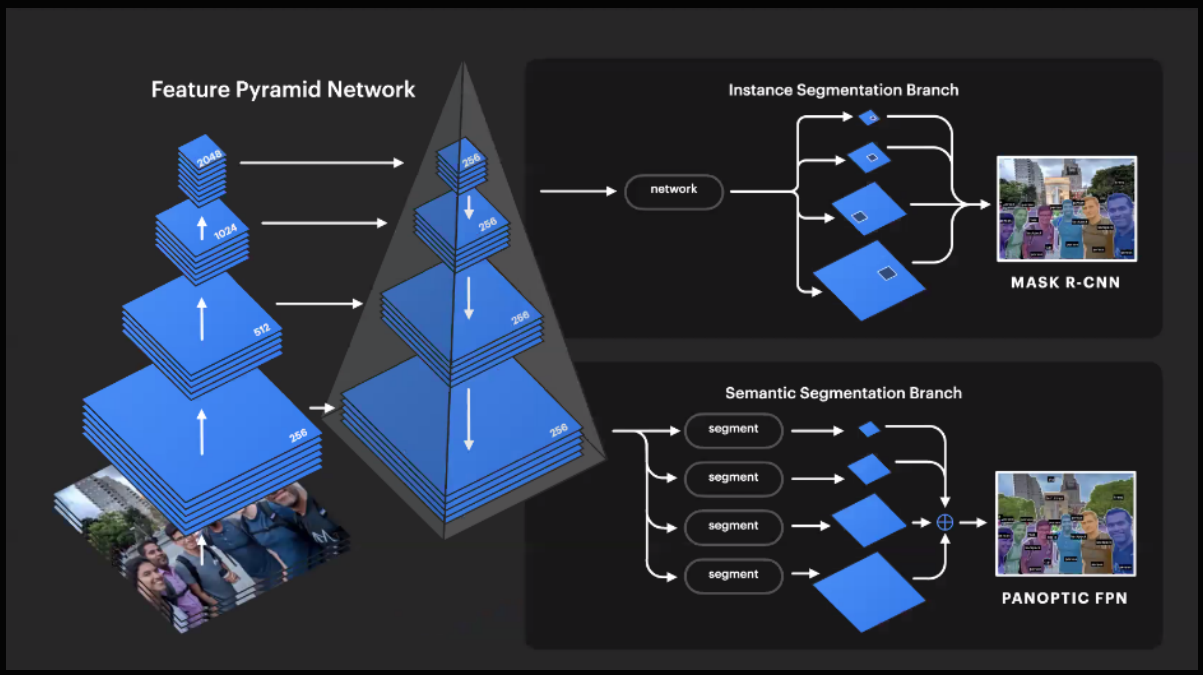
PANOPTIC FPN

2019

Feature Pyramid Network




Feature Pyramid Network
2048
1024
512
256
256
256
256
256
Instance Segmentation Branch
network
MASK R-CNN


Feature Pyramid Network
2048
1024
512
256
256
256
256
256
Instance Segmentation Branch
network
MASK R-CNN
Semantic Segmentation Branch
segment
segment
segment
segment
PANOPTIC FPN

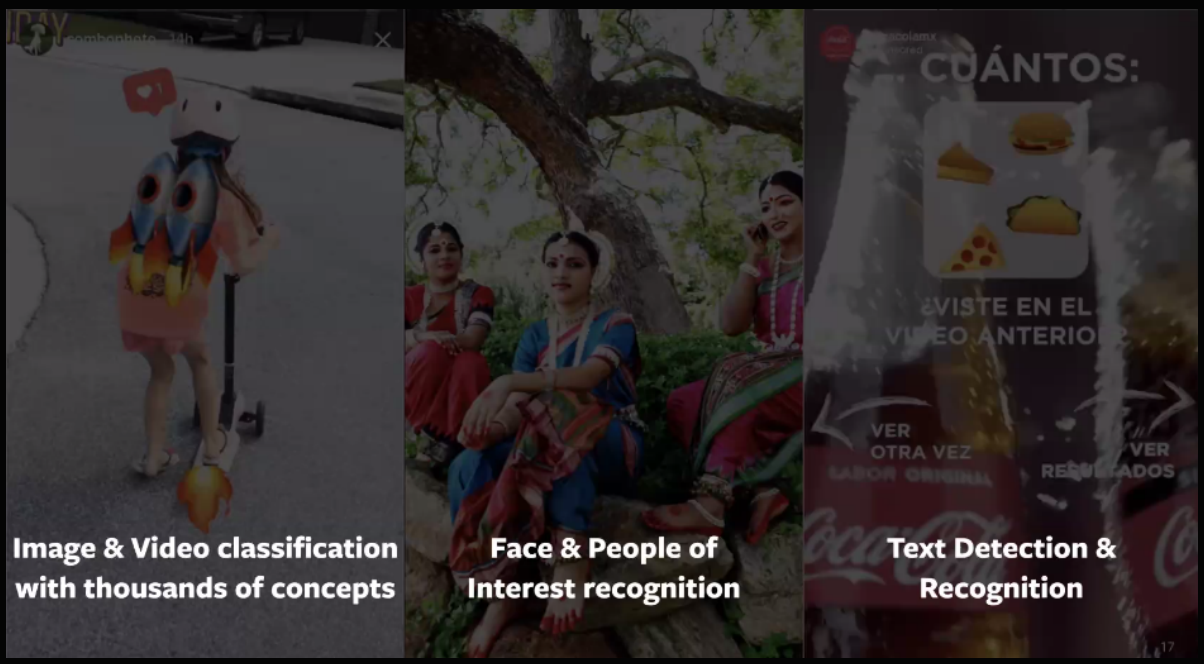
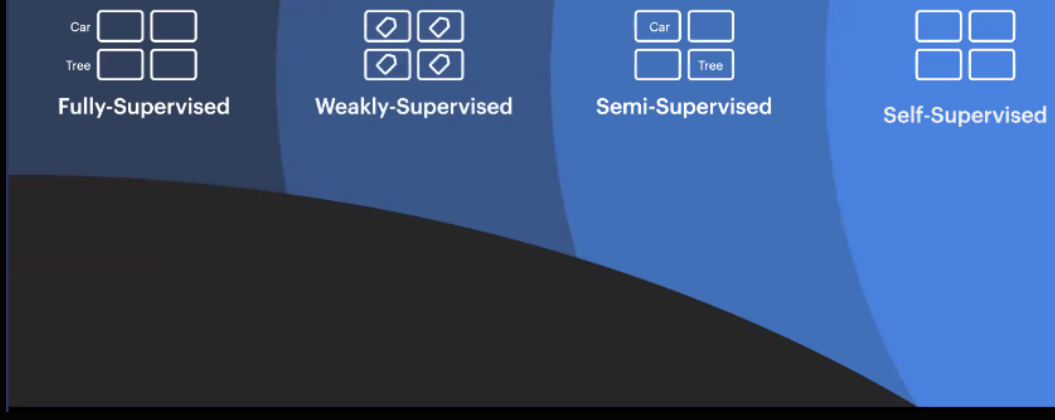


Image & Video classification with thousands of concepts
Face & People of Interest recognition
CUÁNTOS:
¿VISTE EN EL VIDEO ANTERIOR?
VER OTRA VEZ
VER RESULTADOS
Text Detection & Recognition

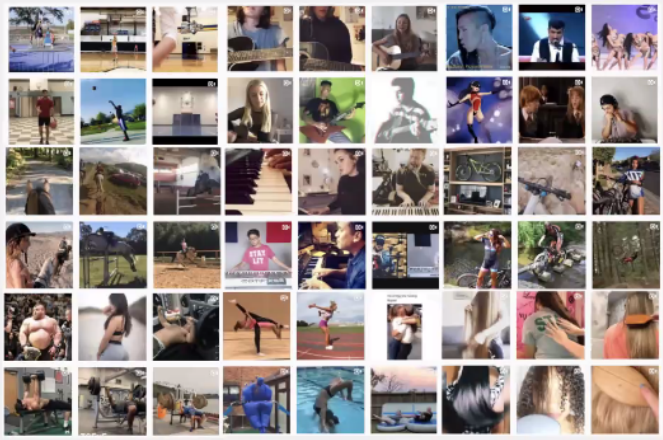
Levels of Supervision



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

CAT, DOG, WOODEN FLOOR ?

Challenges of Training at Billion Scale

NOISY DATA

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

Missing Labels

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

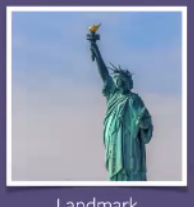
83.1%

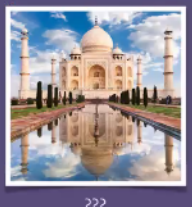
PREVIOUS SOA

Before


Food


Food


Landmark

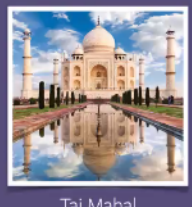

???

After


Cupcake


Apple pie


Statue of Liberty


Taj Mahal

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()
```

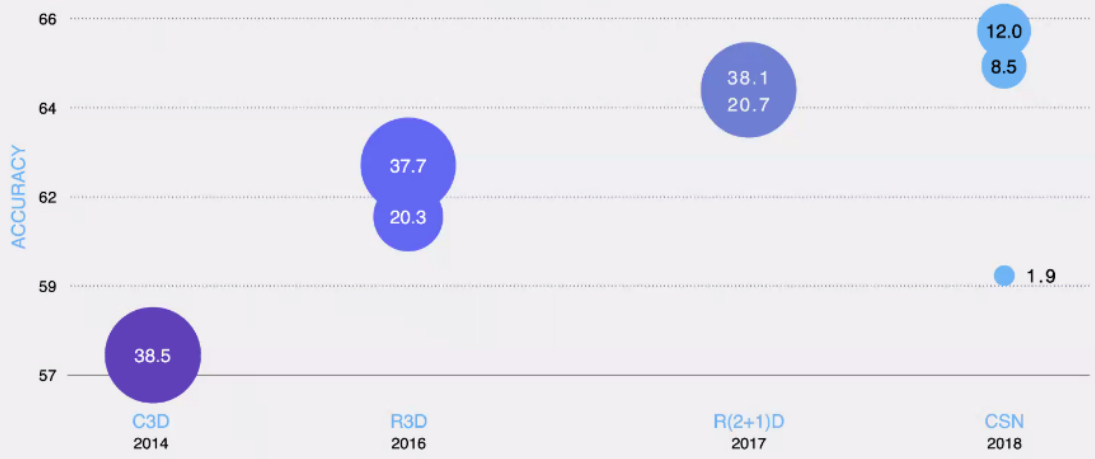
How about Videos?

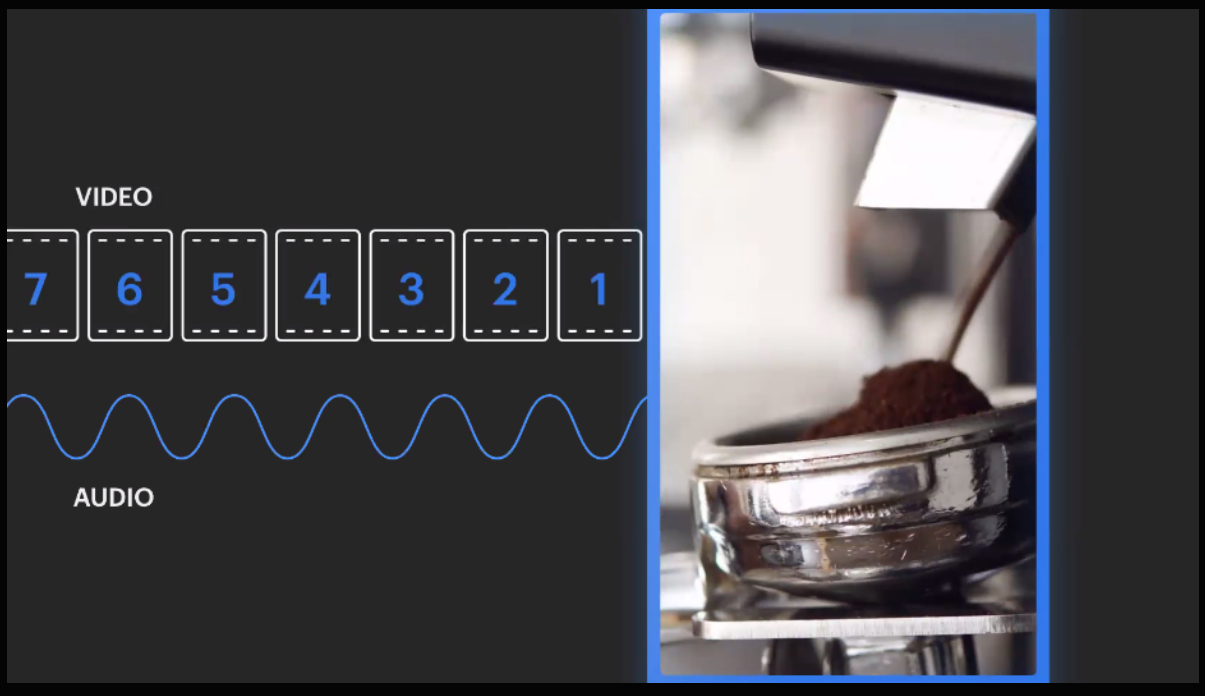
ENRICH OUR USER'S EXPERIENCE

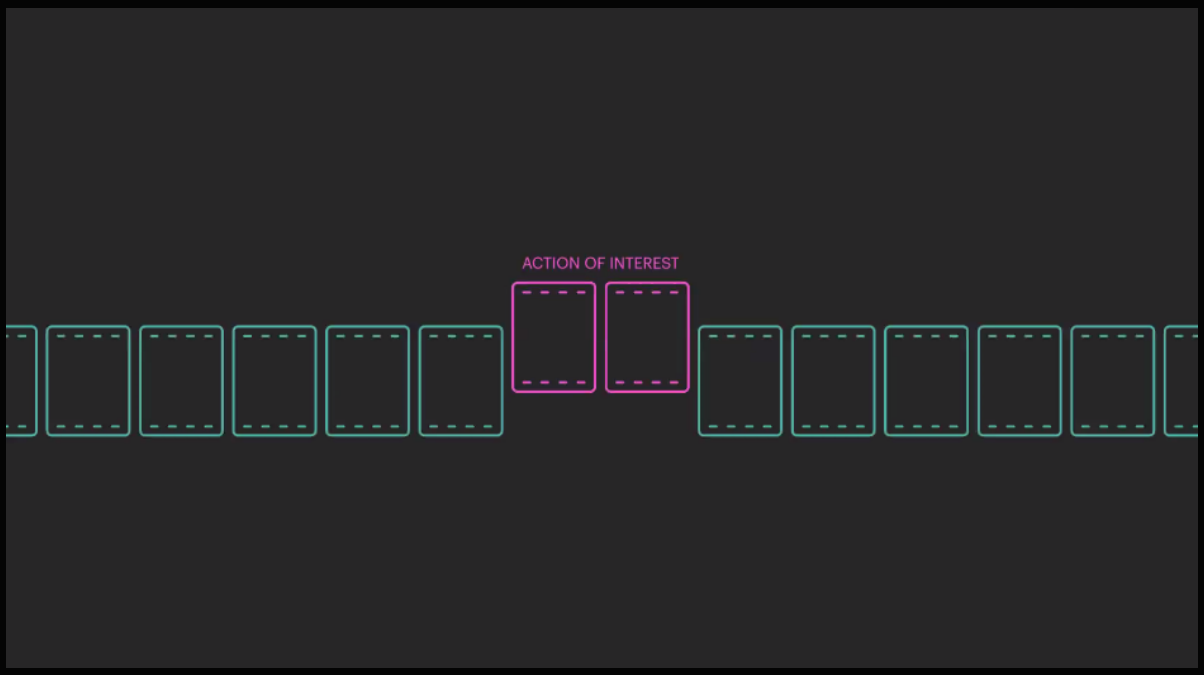
Did We Miss a Great Moment?

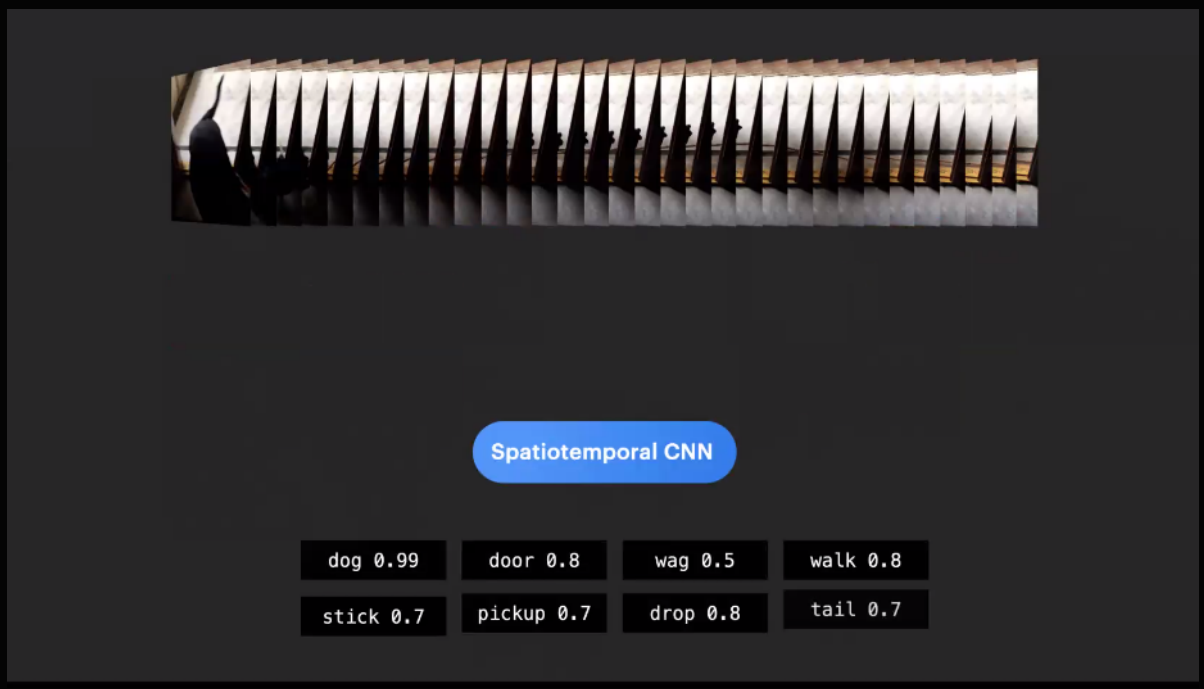
Spatiotemporal Visual Modeling

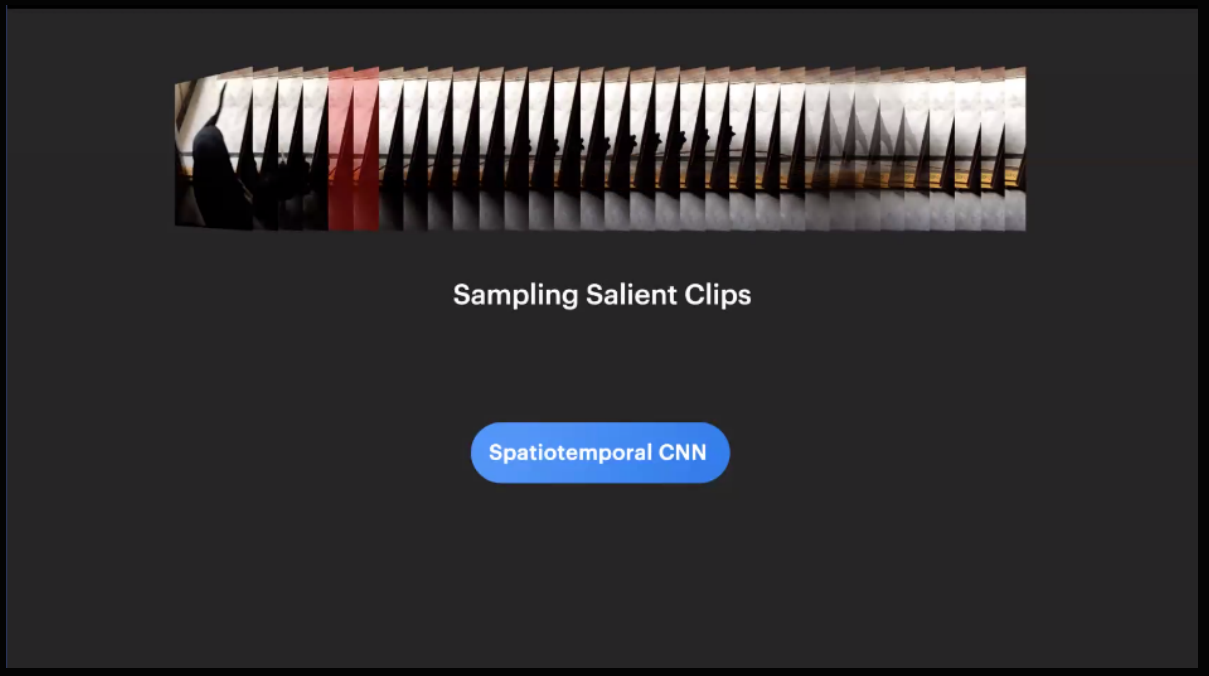
SPACE AND TIME HAVE DIFFERENT STATISTICS

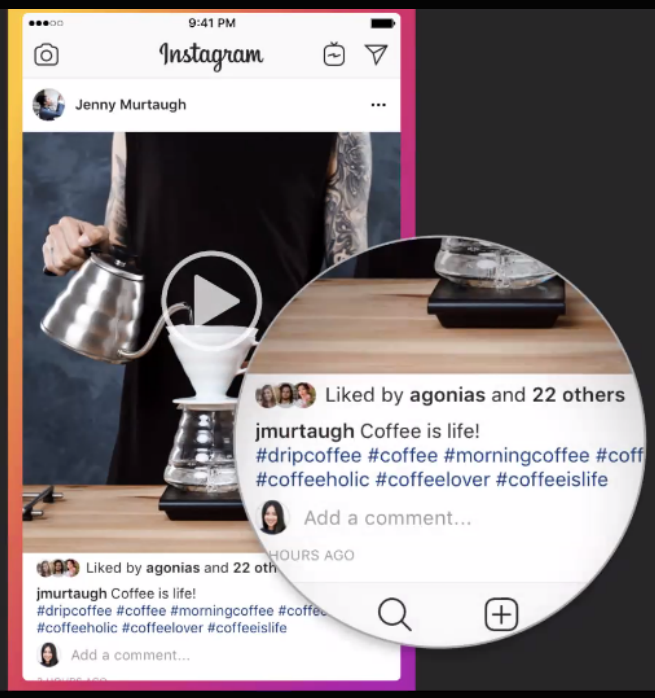





ACTION OF INTEREST


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


Sampling Salient Clips
Spatiotemporal CNN



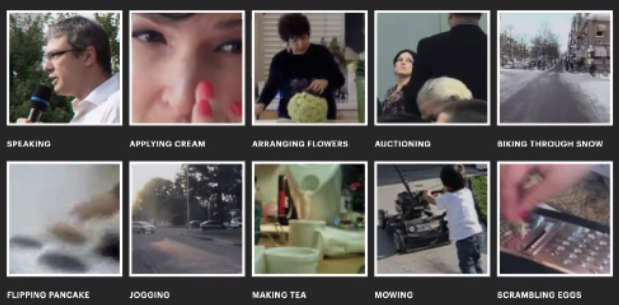
Leveraging Rich Metadata

HASHTAGS

65M

PUBLIC VIDEOS

State of the Art Results



KINETICS: 300K VIDEOS, 400 ACTIONS

Metric: Top-1 Accuracy

77.7%

PREVIOUS SOA

+5.1%

82.8%

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

Emerging Properties in Self-Supervised Vision Transformers

Mathilde Caron, Hugo Touvron, Ishan Misra,
Herve Jegou, Julien Marial, Piotr Bojanowski, Armand Joulin

https://github.com/facebookresearch/dino

ICCV 2021

Barlow Twins: Self-supervised Learning via Redundancy Reduction

Jure Zbontar*, Li Jing*, Ishan Misra, Yann LeCun, Stéphane Deny

https://github.com/facebookresearch/barlowtwins

ICML 2021

PreDet: Large-scale weakly supervised pre-training for detection

Vignesh Ramanathan, Rui Wang, Dhruv Mahajan

Problem setup & Goal

Goal: Pre-train with 50M+ weakly labelled Instagram images (hashtags only) to obtain a good detection-specific trunk. Fine-tuning the trunk for detection should result in **mAP gain for datasets like COCO**.

(Extend benefits of large-scale pre-training from classification* to detection)

Problem setup:

Pre-training input: IG images only with hashtags (pre-filtered to 1k LVIS classes)

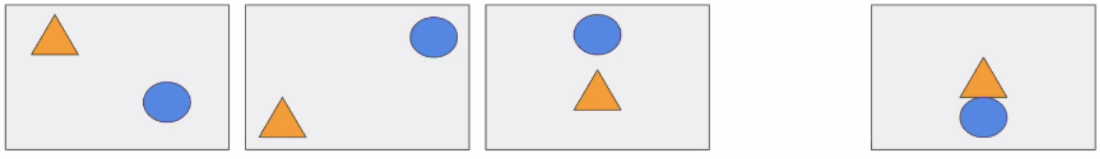
Fine-tuning input: Fully-labelled detection dataset

Mahajan et. al. "Exploring the limits of large-scale weakly supervised pre-training", ECCV'18

Some intuition and analysis

Why large-scale classification pre-training* does not improve detection?

Possible reason: Too much translation invariance!



classification pre-training collapses all variations to the same representation!

Mahajan et. al. "Exploring the limits of large-scale weakly supervised pre-training", ECCV'18

Motivation for model

- Train with **traditional classification loss**
- However, **reduce translation invariance** through an additional loss
- Make the model explicitly distinguish different boxes in an image and ensure they have different representations

PreDet model

- Detection model has 3 components
  - Classification
  - Coarse localization of objects
  - Fine-grained bbox regression
- Train each component with a separate loss
  - **Classification**: supervised loss with hashtags same as URU
  - **Coarse localization**: Self-supervised loss (predict overlap b/w bboxes as an example)
  - **Bbox regression**: Self-supervised loss (predict regression between two randomly sampled bboxes in an image)

PreDet motivation

Classification: supervised loss with hashtags same as URU

dog, donut

1. URU-like softmax loss milti-label setup
2. For FPN models, we have one classification loss at each fpn layer and sum them together

PreDet motivation

Coarse localization: Self-supervised loss

Contrastive loss for overlapping boxes

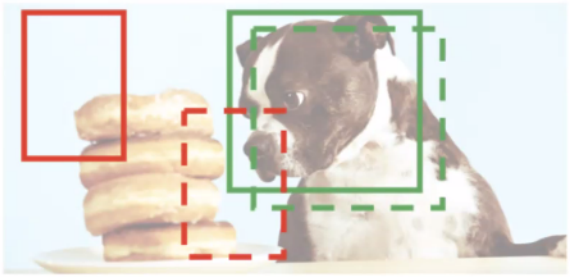
query box

positive key box (overlap with query > 0.7)

PreDet motivation - bbox classification

Coarse localization: Self-supervised loss

Contrastive loss for overlapping boxes



query box

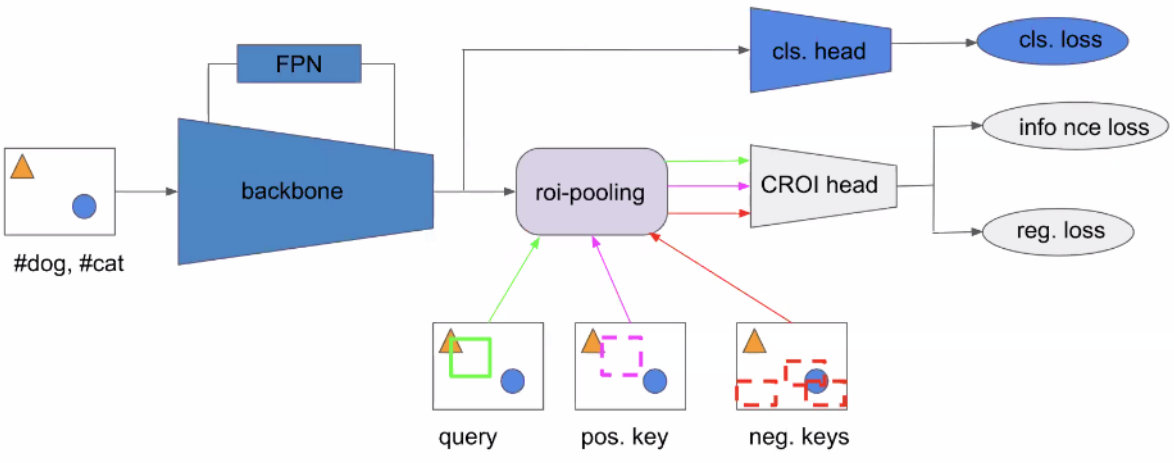
positive key box (overlap with query > 0.7)

hard-negative key box (overlap with query < 0.3, > 0.0)

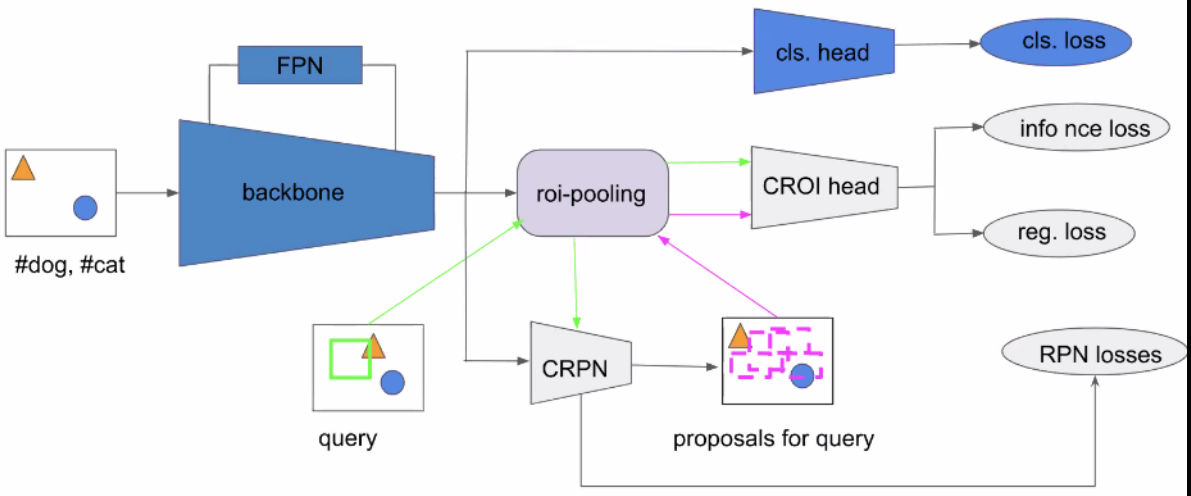
random negative key box

dist(query, pos. key) < dist(query, neg. key)
Use InfoNCE loss similar to MOCO

PreDet with contrastive roi-head (CROI)

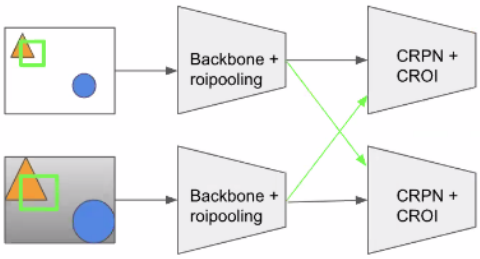


PreDet with CROI + contrastive RPN (CRPN)



PreDet with CROI + CRPN trained with 2-views

- Self-supervised loss is “too easy” for the model. Losses converge to a low value
- Make task harder, inspiration from MOCO
  - ROI-pooled features for query comes from a different view of the same image
  - Every batch has two-views of same image. Query features are obtained by pooling from the features of a different view. (Much harder task)



Results (coco-full) - Resnext101-32x8d

pre-training	Mask R-CNN					RetinaNet		
	1× sched.		best sched.			1× sched.	best sched.	
	AP^{box}	AP^{mask}	sched.	AP^{box}	AP^{mask}	AP^{box}	sched.	AP^{box}
from scratch	33.9	31.0	9×	45.8	40.7	27.6	9×	40.7
cls-ImageNet	43.8	39.0	6×	44.9	39.9	41.4	1×	41.4
cls-IG50M	44.4	39.4	3×	44.6	39.5	41.8	1×	41.8
InfoMin [54]	44.8	40.2	3×	45.6	40.5	43.0	1×	43.0
SEER-IG1B [13]	44.3	39.9	3×	45.1	40.1	40.3	6×	41.7
PreDet-ImageNet	45.8	40.8	1×	45.8	40.8	43.1	1×	43.1
PreDet-IG50M	**47.1**	**41.7**	1×	**47.1**	**41.7**	**45.1**	1×	**45.1**

Our model needs significantly fewer (> 3x faster) fine-tuning iterations and leads to higher detection mAP!

Pre-training impact for smaller fine-tuning datasets

Dataset size	From scratch	cls-ImageNet	PreDet-IG50M
1k	4.2	11.3	16.1
5k	12.5	22.3	26.7
10k	25.3	26.5	30.3
35k	37.8	37.6	40.1
118k	45.8	44.9	47.1

Training on smaller subsets of COCO shows bigger benefits from PreDet!

Unidentified Video Objects:

A Benchmark for Dense, Open-World Segmentation

Weiyao Wang

Matt Feiszli

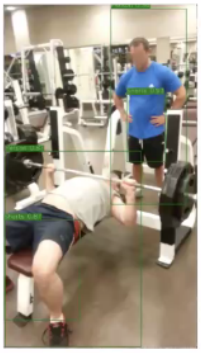
Heng Wang

Du Tran

Motivations

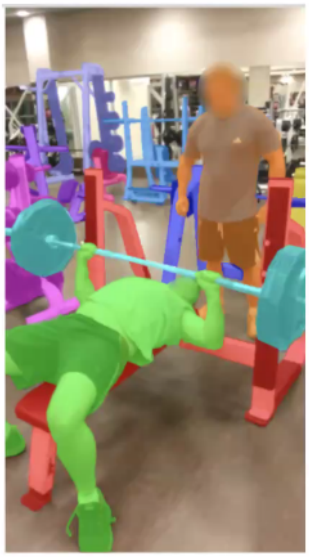
SoTA detectors work well under **closed-world** assumption:	• Detecting **in-taxonomy** objects.
Real-world is **open**:	• Real-world contains objects that are outside taxonomy.
Humans can handle open-world	• Detect object **without** knowing object's concept

Mask-RCNN prediction
(80 COCO categories)



Google Cloud API
(550+ object categories)

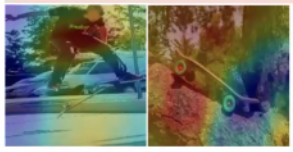
How can we get here?



Open-world is difficult

Modern top-down approach (e.g., R-CNN):

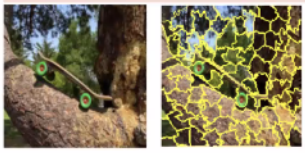
- SoTA performance
- Couple recognition and detection
- Rely on strong contextual cue



Detecting Skate-board relies on recognizing context of person [1]

Classical bottom-up approach (e.g., super-pixels/voxels)

- Agnostic to training data
- Inferior results
- Have no notion of semantics



Super-voxel methods fail to predict whole segments and contain no notion of semantics

[1] Singh el al., Don't Judge an Object by Its Context: Learning to Overcome Contextual Bias,

Why open-world?

Ability to handle novel (unseen) objects

* Key to embodied AI (e.g., robotics, autonomous driving)

Enable holistic understanding of images / videos

* Allow object-centric reasoning
* Enable long-term video modeling

UVO: benchmark to study open-world

Open-world requires:

- **Exhaustively** annotate **all** objects present in a video
- **Taxonomy-free** annotation

UVO for open-world:

- Dense: 1200 videos annotate at 30fps
- Sparse: 9k videos annotate at 1fps
- On average **19** unique object instances

UVO statistics

Dataset	Videos	Frames	Taxonomy	Objects per video/frame	Annotation fps	Total mask annotations
DAVIS	150	11k	"salient"	2.99	24	~32k
YouTube-VOS	4453	120k	94 classes	1.64	6	~196k
YouTube-VIS	2883	78k	40 classes	1.68	6	~131k
UVO-Dense	1200	108k	open-world	**16.7**	30	878k
UVO-Sparse	9000	27k	open-world	**13.4**	1	361k
*COCO17**	NA	164k	80 classes	7.3	NA	~1197k

Recap

- Most current SoTA methods focus on closed-world setting.
- Open-world segmentation is essential to enable object-centric reasoning and long-term video understanding.

Participate in our **ICCV 21 Challenge**:
https://sites.google.com/view/unidentified-video-object/workshop-and-challenge

For more info:
https://arxiv.org/abs/2104.04691

Generic Event Boundary Detection (GEBD):

A Benchmark for Event Segmentation

Mike Shou, Stan Lei, Weiyao Wang, Deepti Ghadiyaram, Matt Feiszli

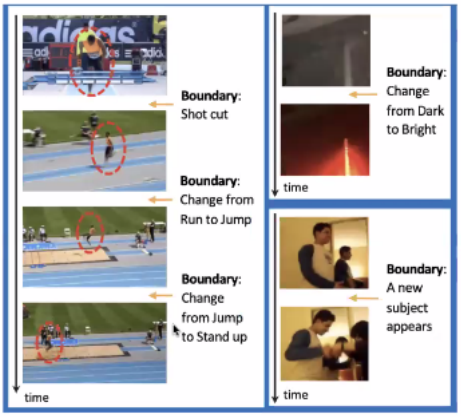


Figure 1. Examples of generic event boundaries: 1) A long jump video segmented at the boundaries of shot cut, change from Run to Jump, change from Jump to Stand up (the dominating subject highlighted with red circle). 2) Boundary due to the color or brightness changes. 3) Boundary due to a new subject appears.