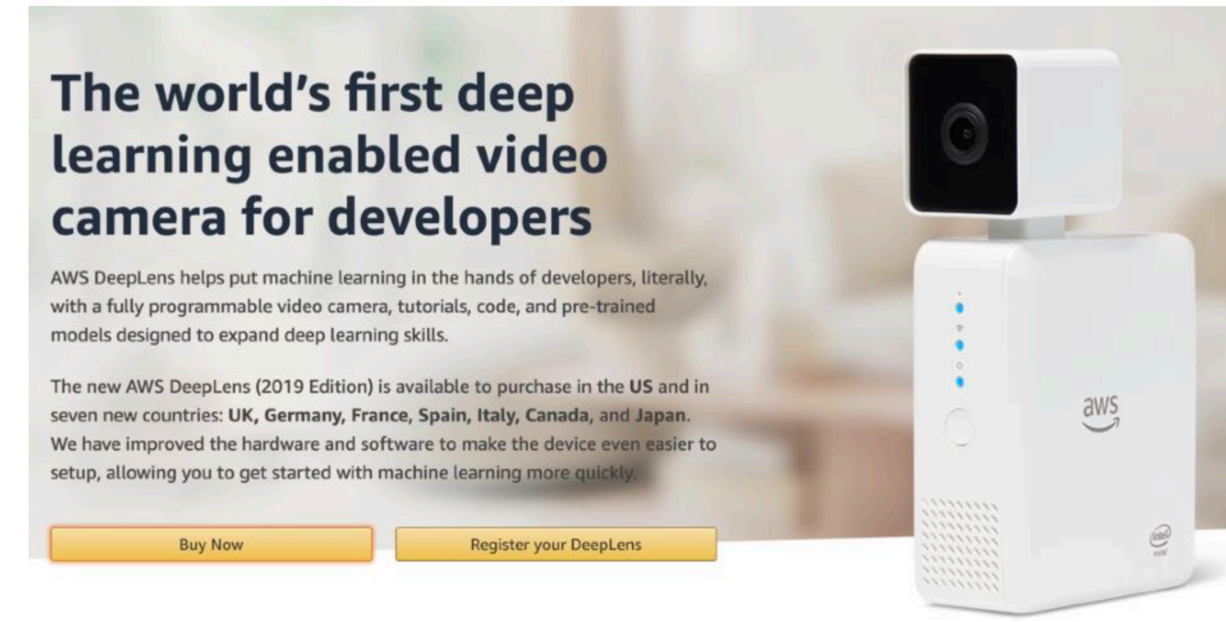
Lecture 11:

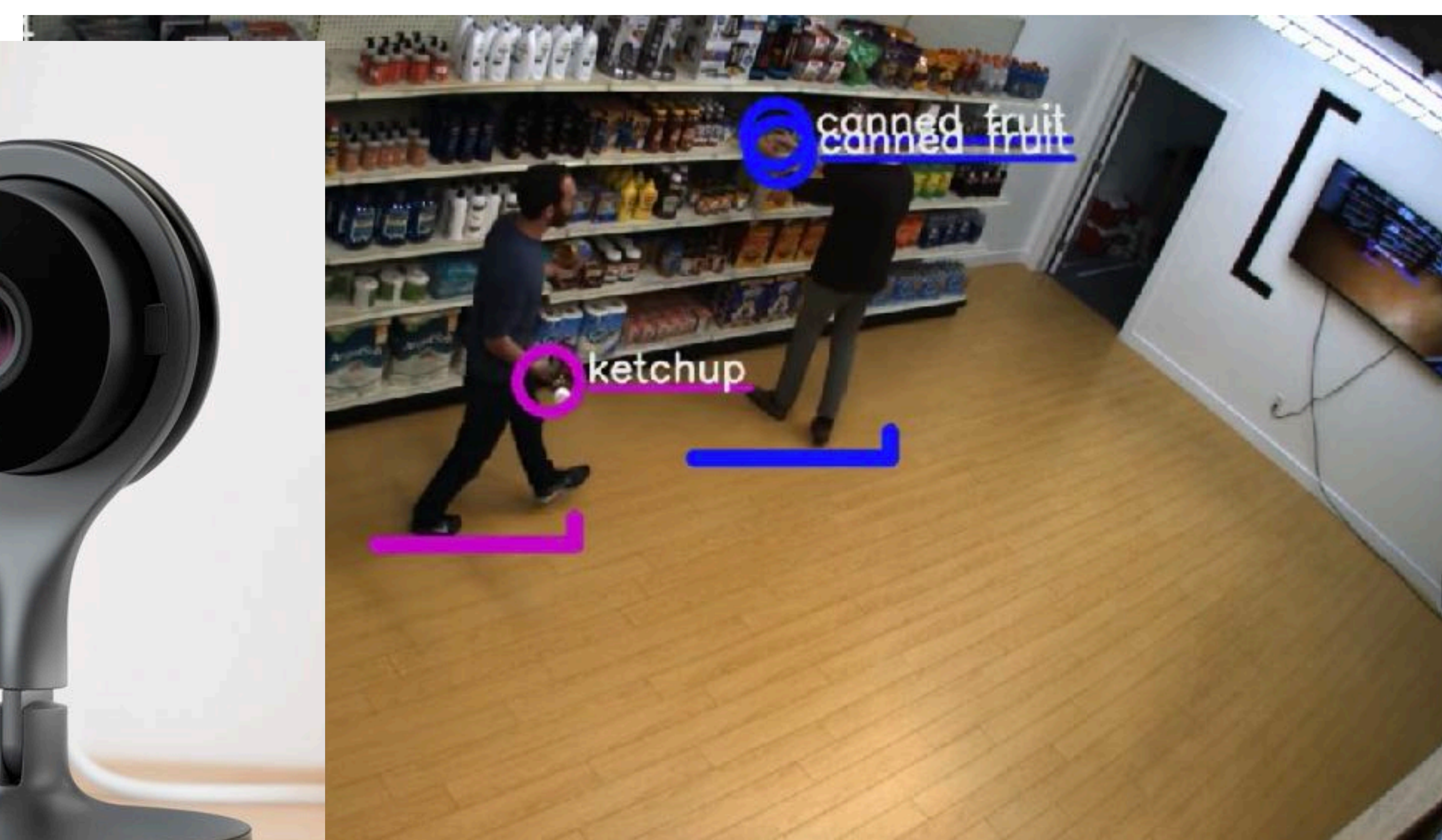
Specialization for Efficient Inference on Video

Visual Computing Systems

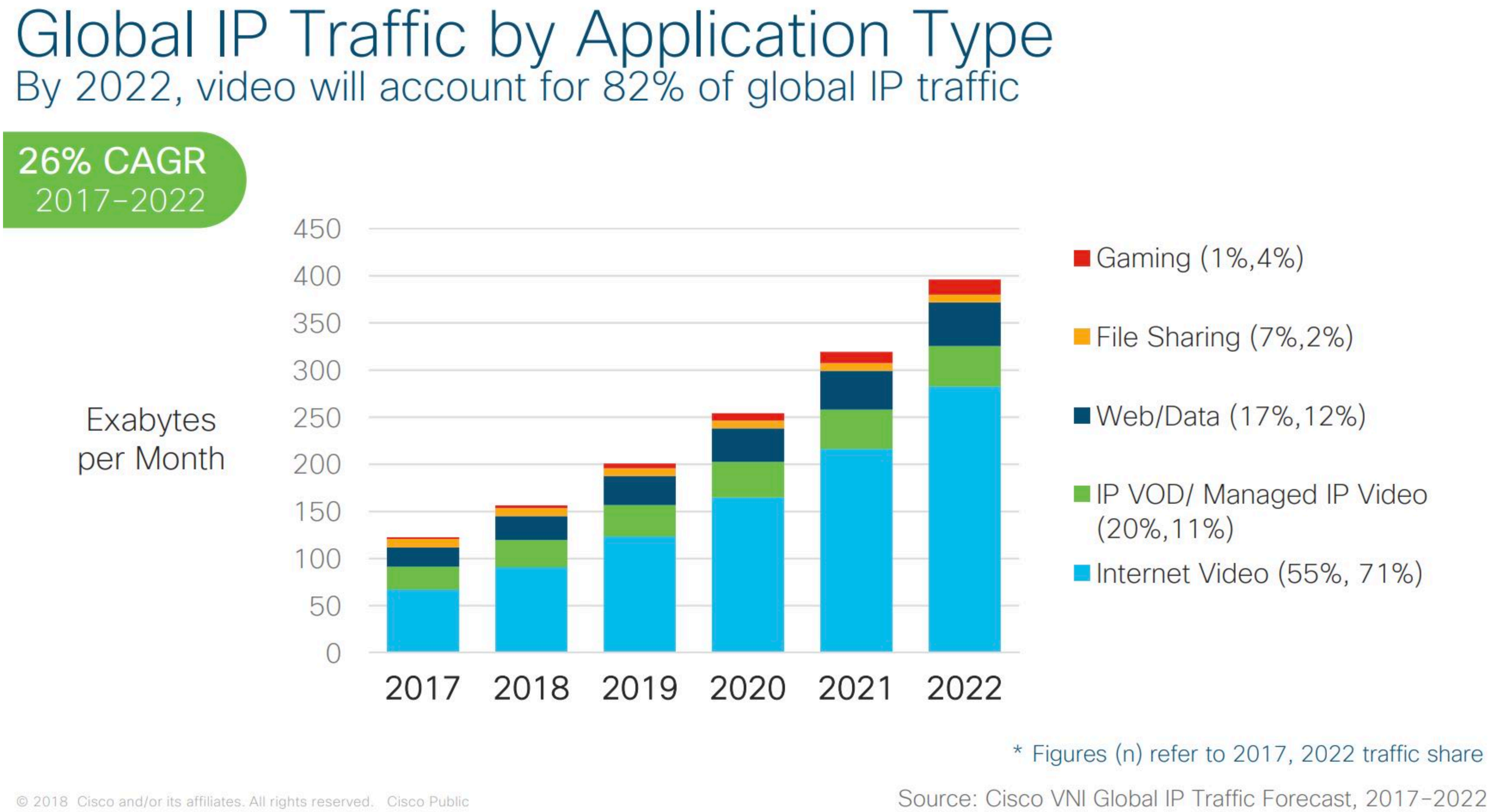
Stanford CS348K, Spring 2021

Video processing applications

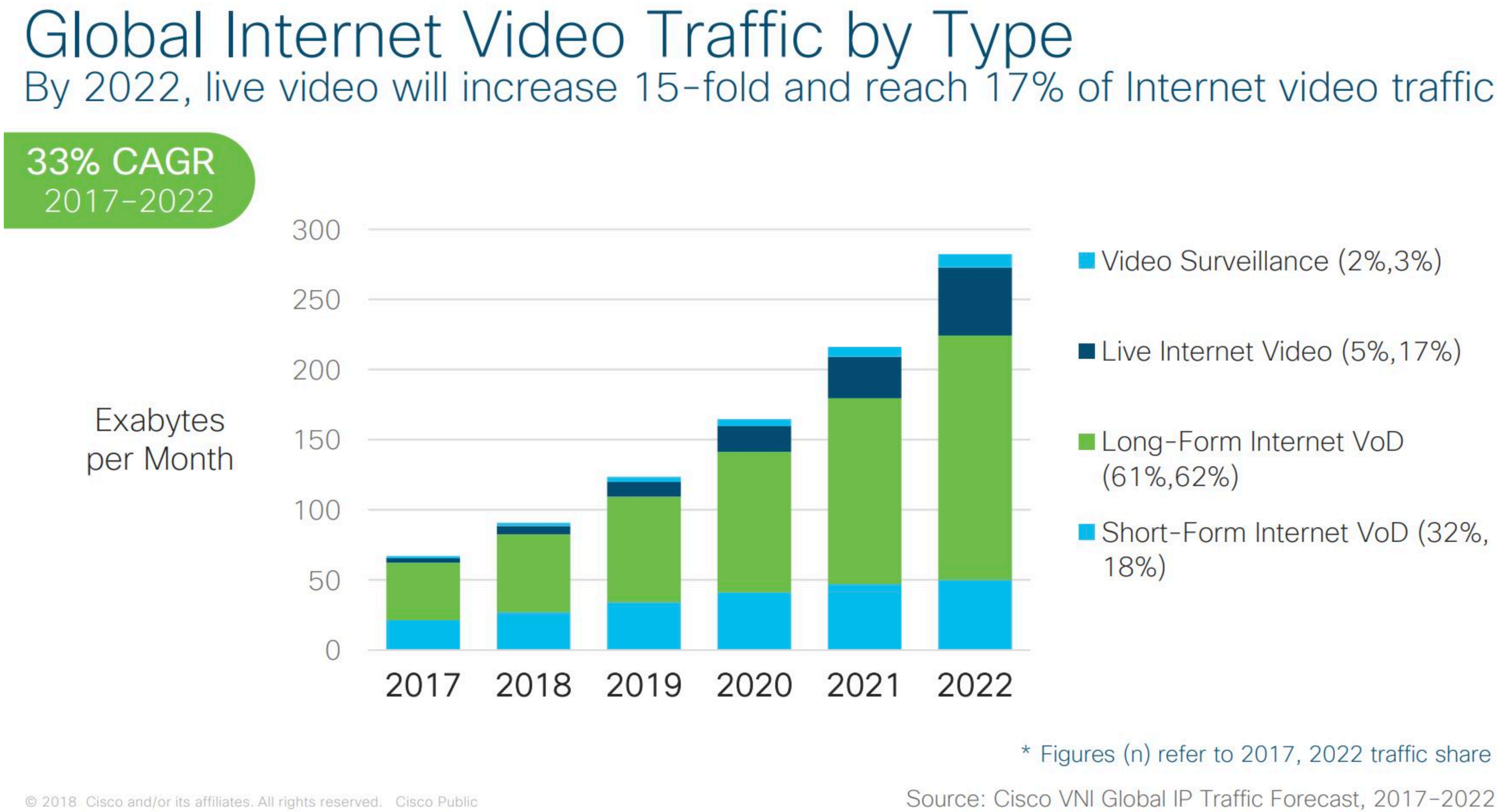




Estimate: 82% of internet traffic will be video



Basically, we're watching TV and movies…

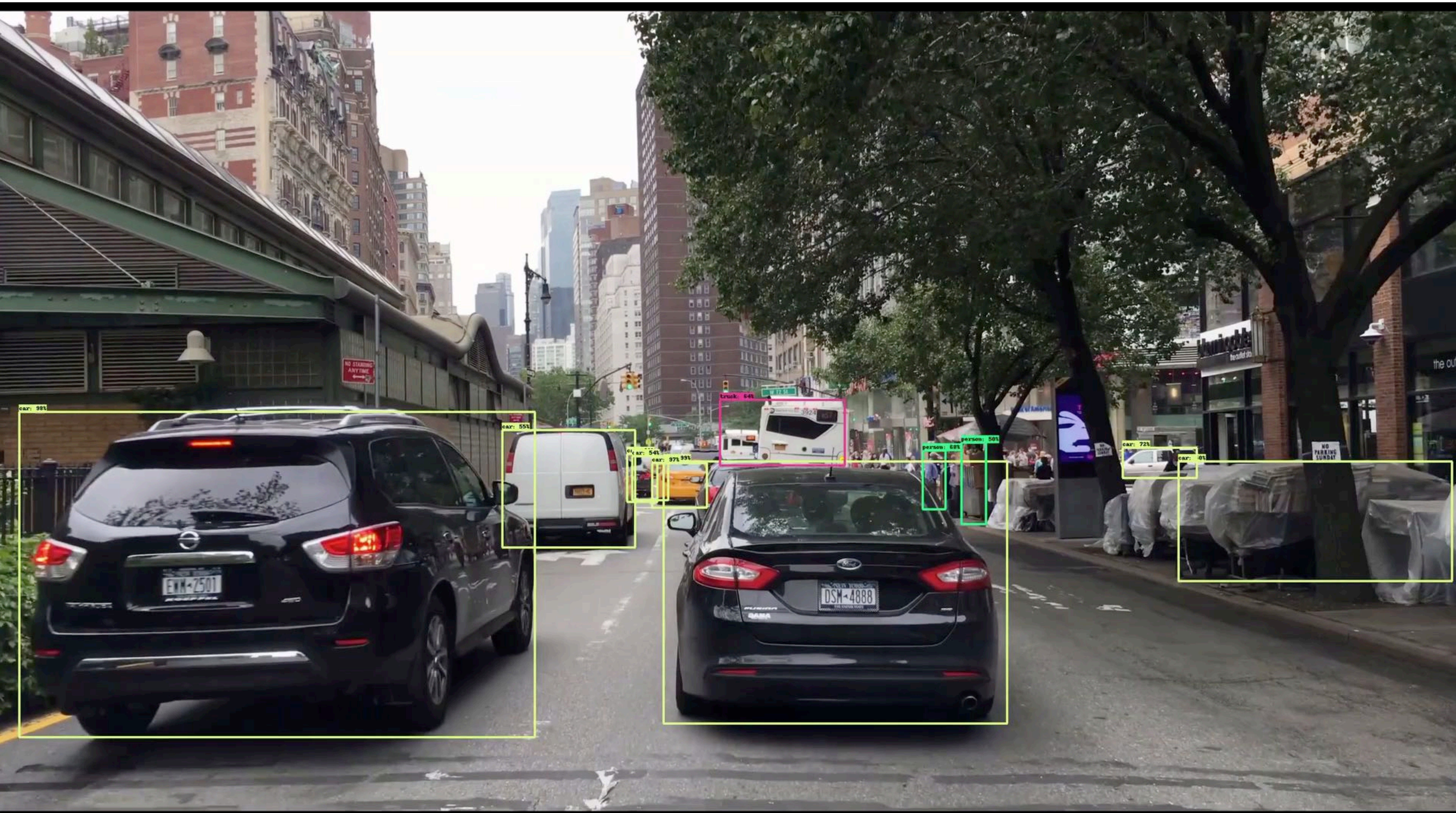


Thought experiment

Imagine we wanted to detect people/cars/bikes in a video stream

Thought experiment

Imagine we wanted to detect people/cars/bikes in a video stream



Interest in processing video efficiently

- Benefits to datacenter applications:
  - Lower cost/frame enables processing of more streams (e.g., thousands of webcams)

- Benefits to edge devices:
  - Cheaper per frame costs, real-time performance on cheaper/lower energy computing hardware
  - Lower latency per frame
    - Example: automated breaking systems target ~40ms sense to brake

Trick 0: video stream subsampling

- Reduce costs by…

- Spatial downsampling:
  - Run detector on low-resolution image

- Temporal subsampling:
  - Run detector at low frame rate

Trick 1: exploit temporal coherence

Temporal differencing

- **Idea: for a new image, use labels from "empty frame" image if new image is similar to background image**

(**a**) empty frame

(**b**) frame with a car

(**c**) subtracted frames

- **Idea: use same result as previous frame if two frames are sufficiently similar**
  - **How to define sufficiently similar? (thresholds?)**
  - **Differences in feature space more robust than over pixels**

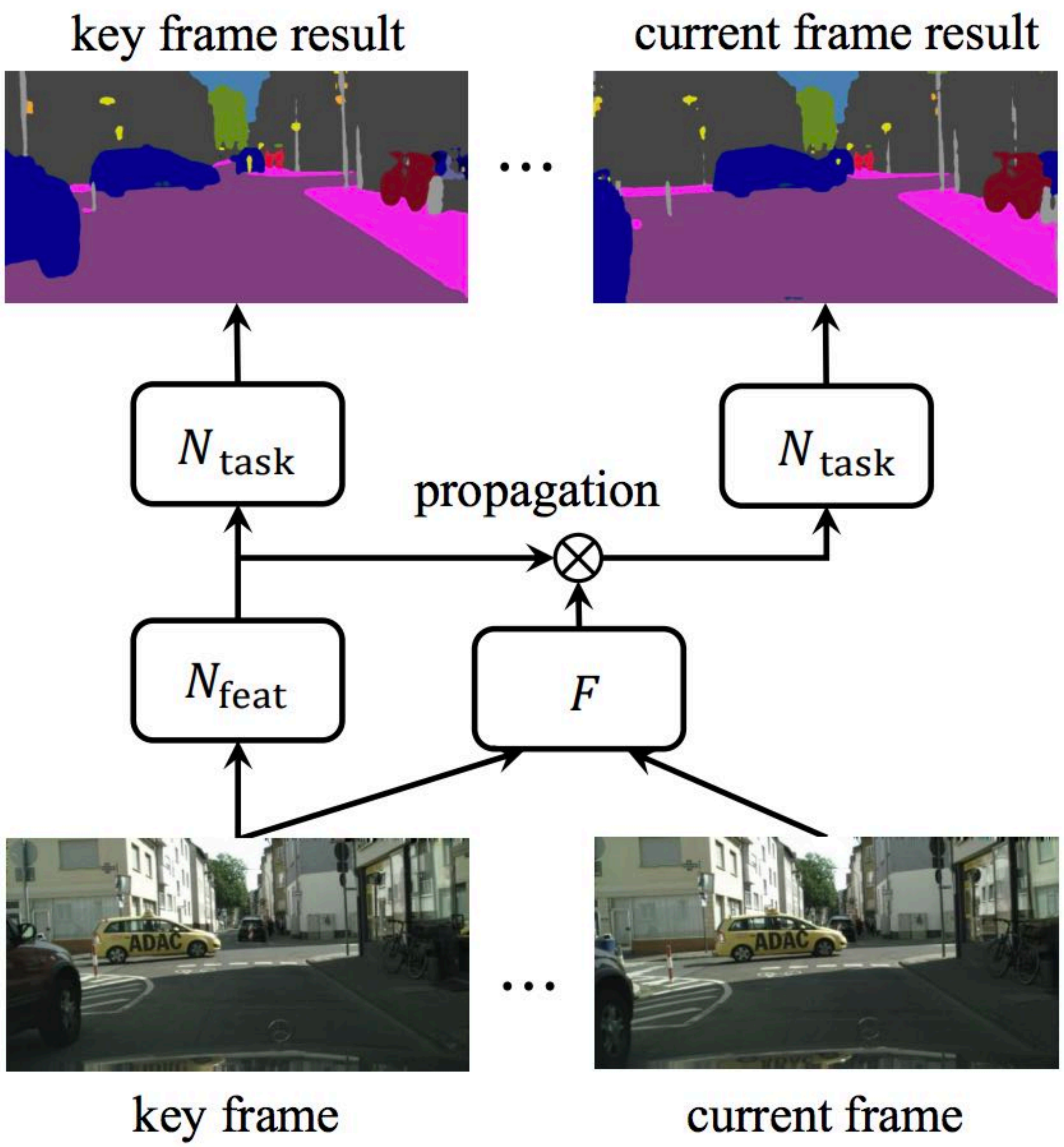
Tracking

Evaluate expensive person detector sparsely in time (e.g., every 1/2 second), then use a more efficient tracking algorithm to update annotations over sequence of frames

Tracking

Evaluate expensive detector sparsely in time (e.g., every 1/2 second), then use more efficient tracking algorithm to update annotations over sequence of frames

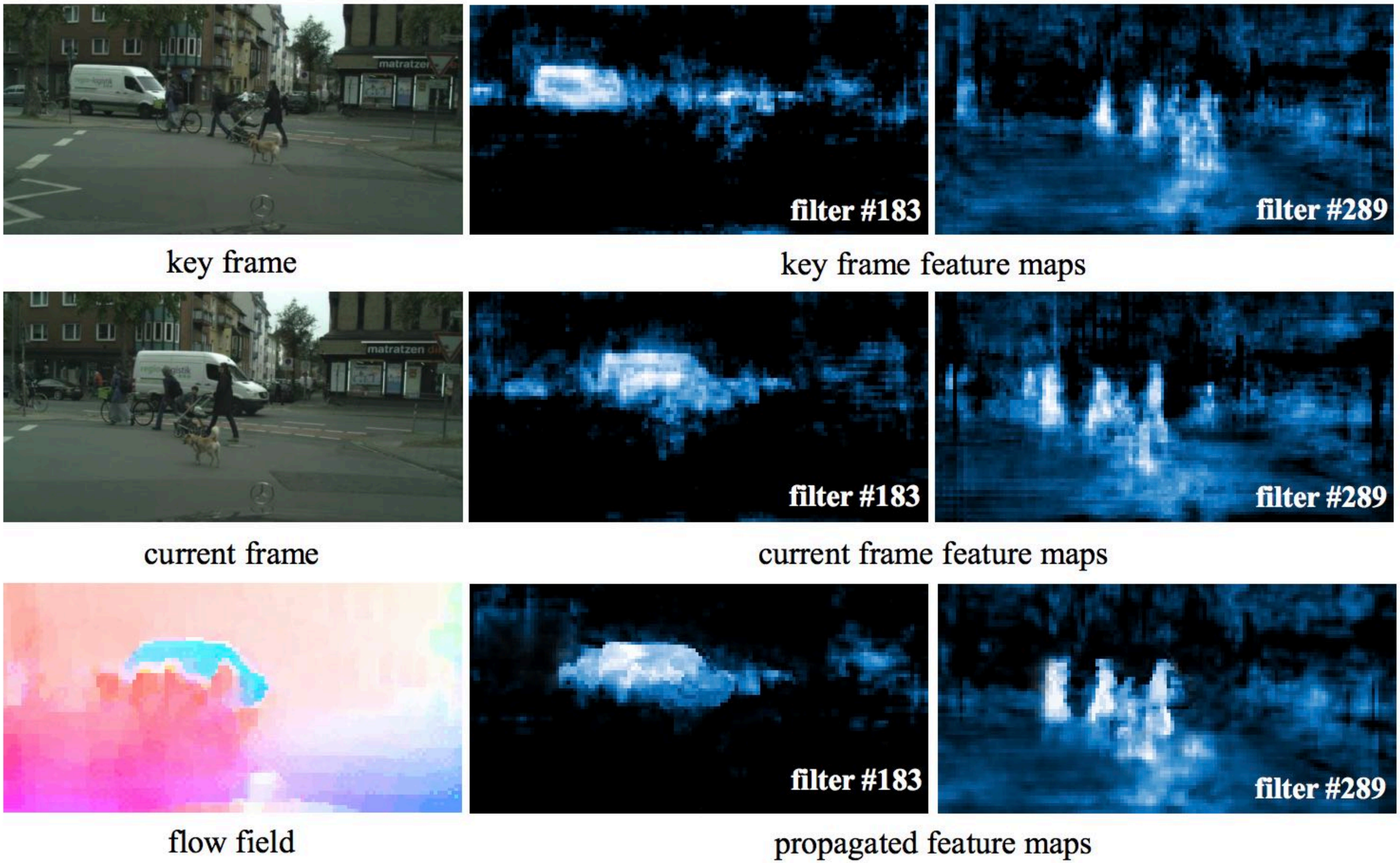
Leveraging motion in the network



Given features (or segmentation result from prior frame) use flow between prior frame and current frame to "advect" features (or segmentation) to new frame.

In other words: it's easier to produce the result for the current frame if you have the result from the prior frame

Leveraging motion in the network



In practice: despite "intellectual" appeal of advecting features, paper results show advecting segmentation is as good as advecting features.

Trick 3: specialize to content

Model specialization

- Common principle in DNN design/training is to learn most general model (via large datasets, regularization, etc.) to perform well across all instances of a task

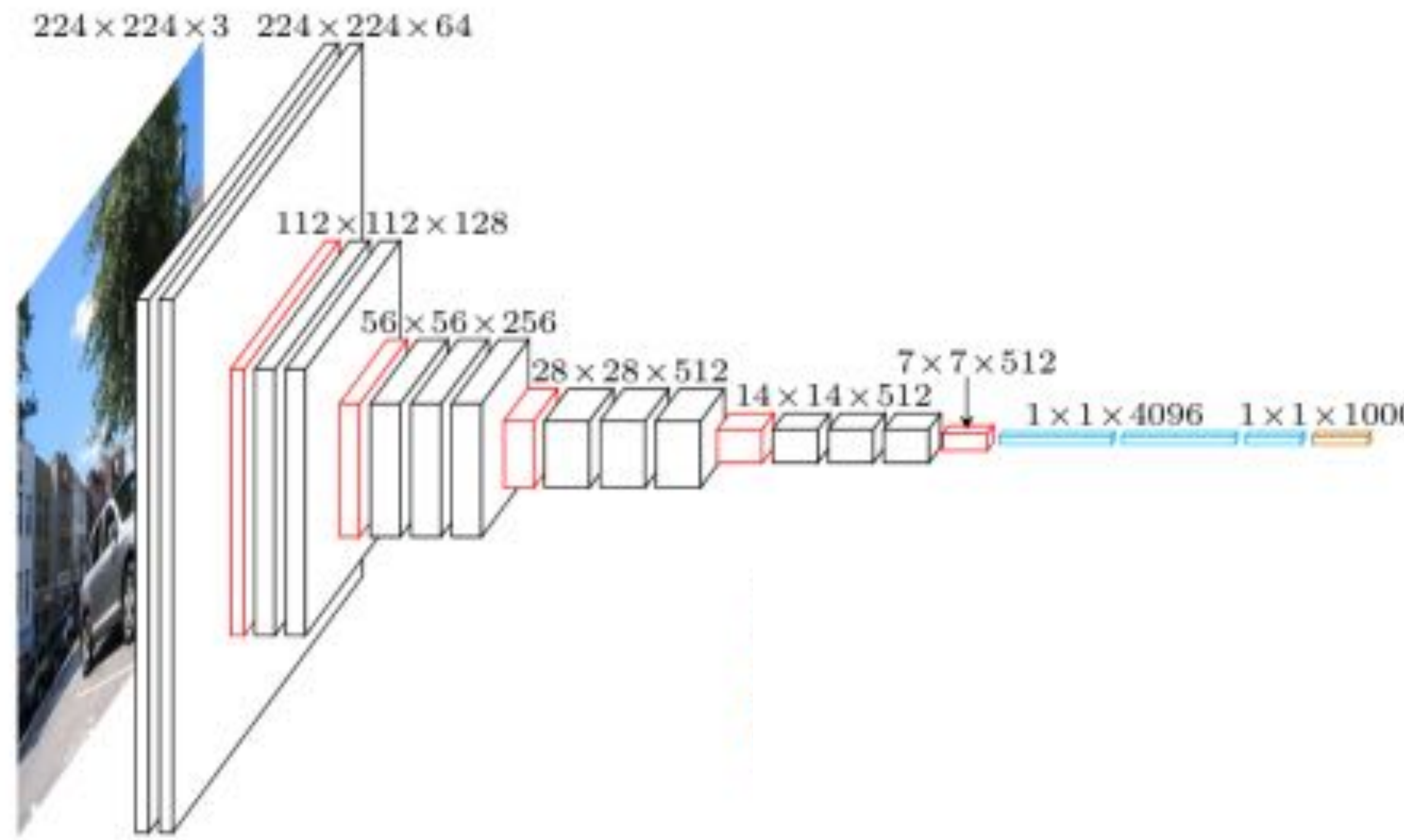
- But many cameras see a very specific distribution of images
  - Only certain types of object classes
  - Always from the same/similar viewpoint
  - Objects appear in same regions of screen

- Specialization has been a major theme in this class w.r.t. hardware design. Now we wish to specialize models to the contents of a video stream
  - "A model can be much simpler if it only needs to work for a single camera"

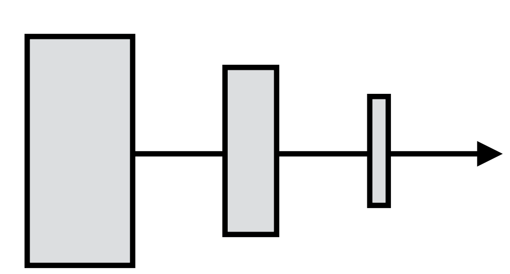
Model distillation

[Hinton 15]

- Accurate, but expensive, model: trained on full training set
  - "The teacher"



- Smaller model (cheaper) , trained to mimic the output of the teacher
  - "The student"



Noscope

[Kang 17]

- **Apply model distillation, but constrain training set to a specific video feed: Given an expensive network that performs a specified detection* task accurately on a wide range of videos, distill to low-cost model specialized to this video stream**

- **Example: binary classification (car/no car) for a single traffic camera video stream**

(**a**) empty frame (**b**) frame with a car (**c**) subtracted frames

*** Noscope actually performs a simpler classification task on a pre-cropped region of the viewport (not detection, which involves object location)**

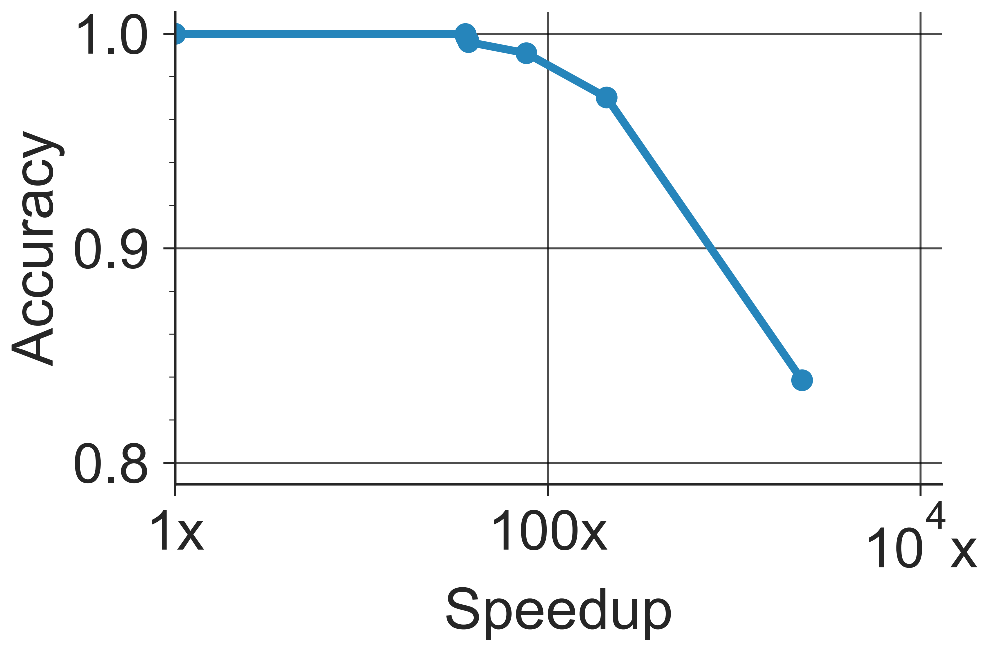
Three Noscope optimizations

- Statically specialize model to video feed
  - Teacher network: Yolo object detection network
  - Student network: compact specialized network (2-4 conv layers)
  - Low cost student "learns" to mimic the teacher

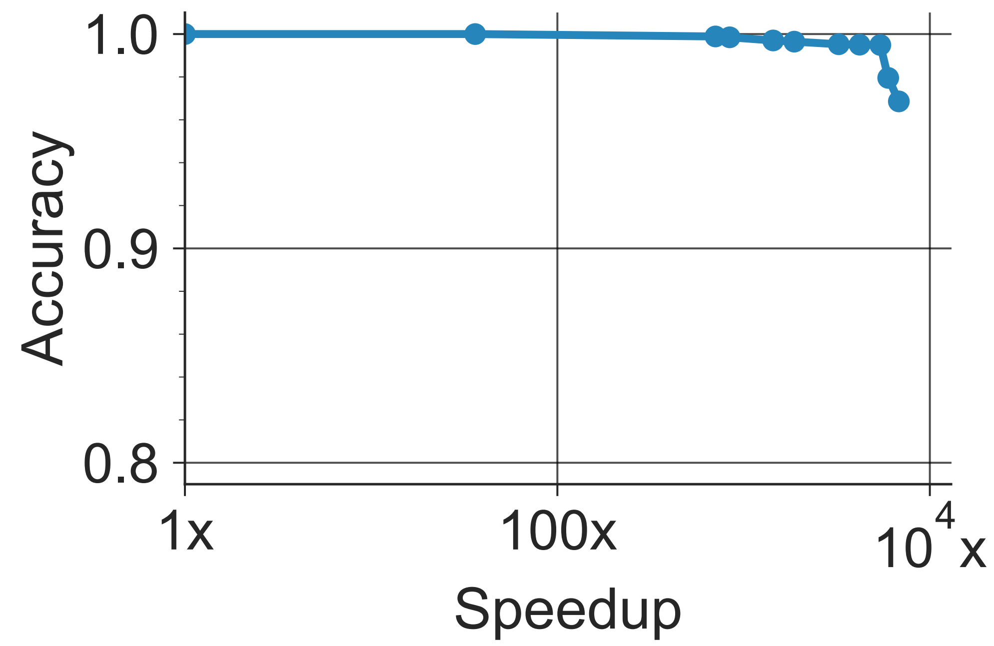
- Dynamic: utilize frame-to-frame difference detectors with learned thresholds
  - "Same as background" and "same as previous frame"
  - Learn thresholds for how often to check for differences (in frames), and what the magnitude of a meaningful difference is

- Dynamic: cascades
  - Run cheap specialized model (student) on frame first, then run teacher model if student does not make a confident prediction

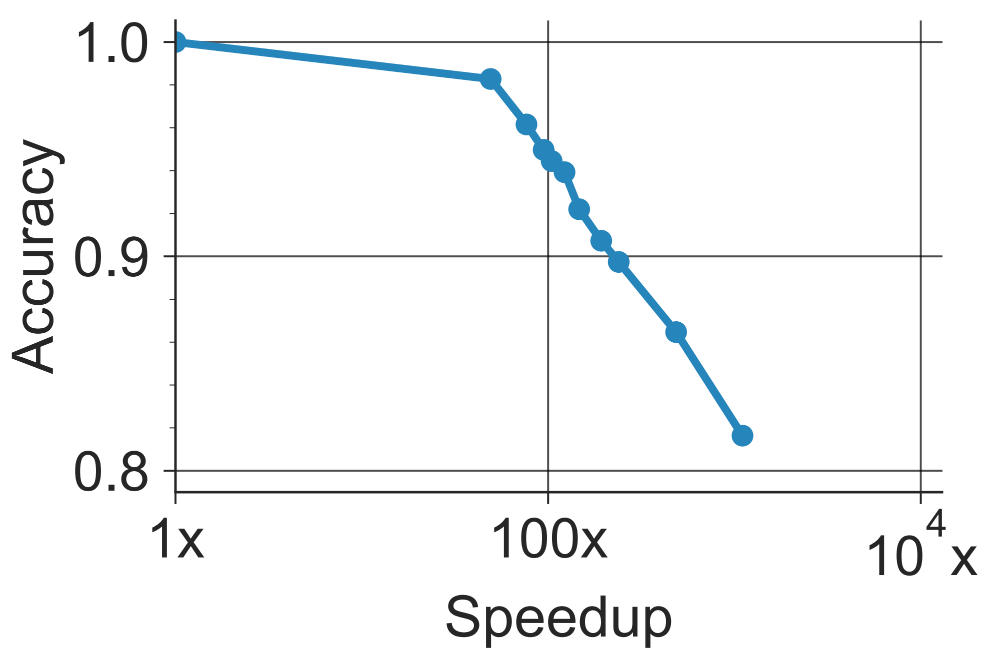
Noscope results *



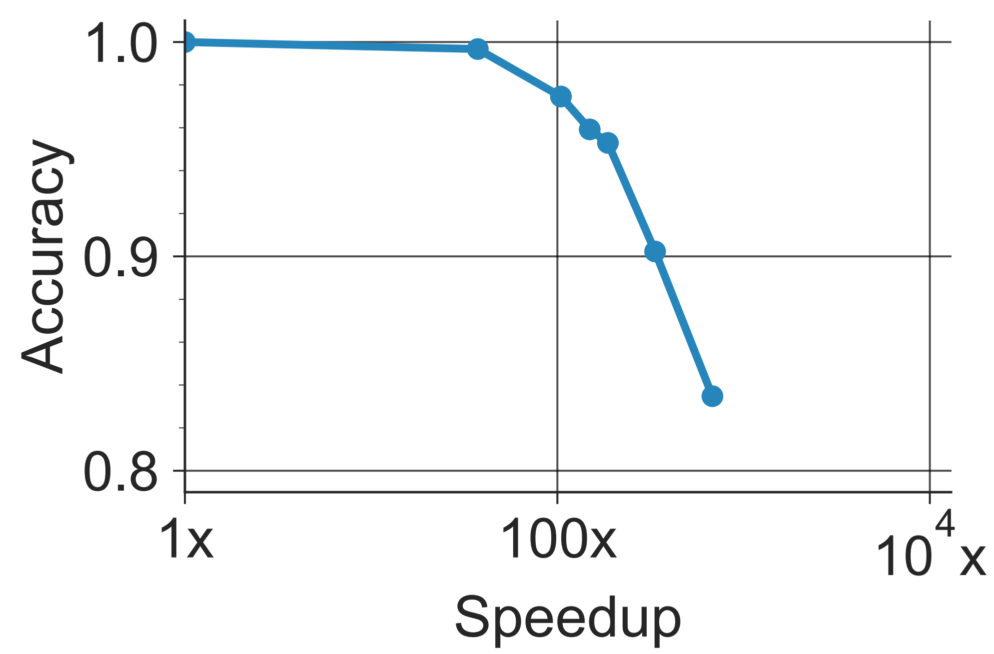
(a) taipei



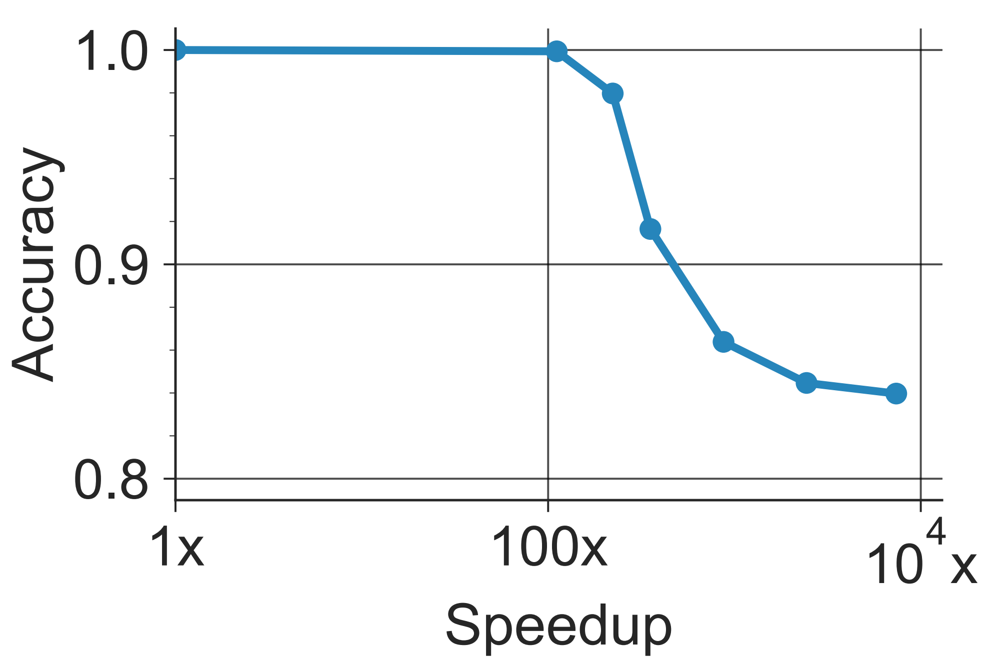
(b) coral



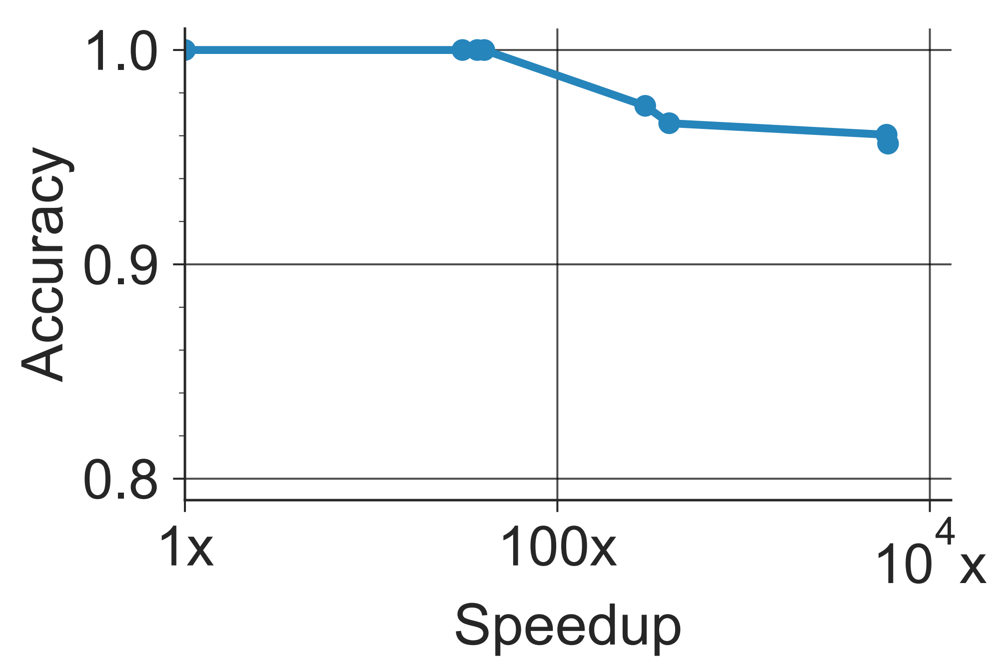
(c) amsterdam



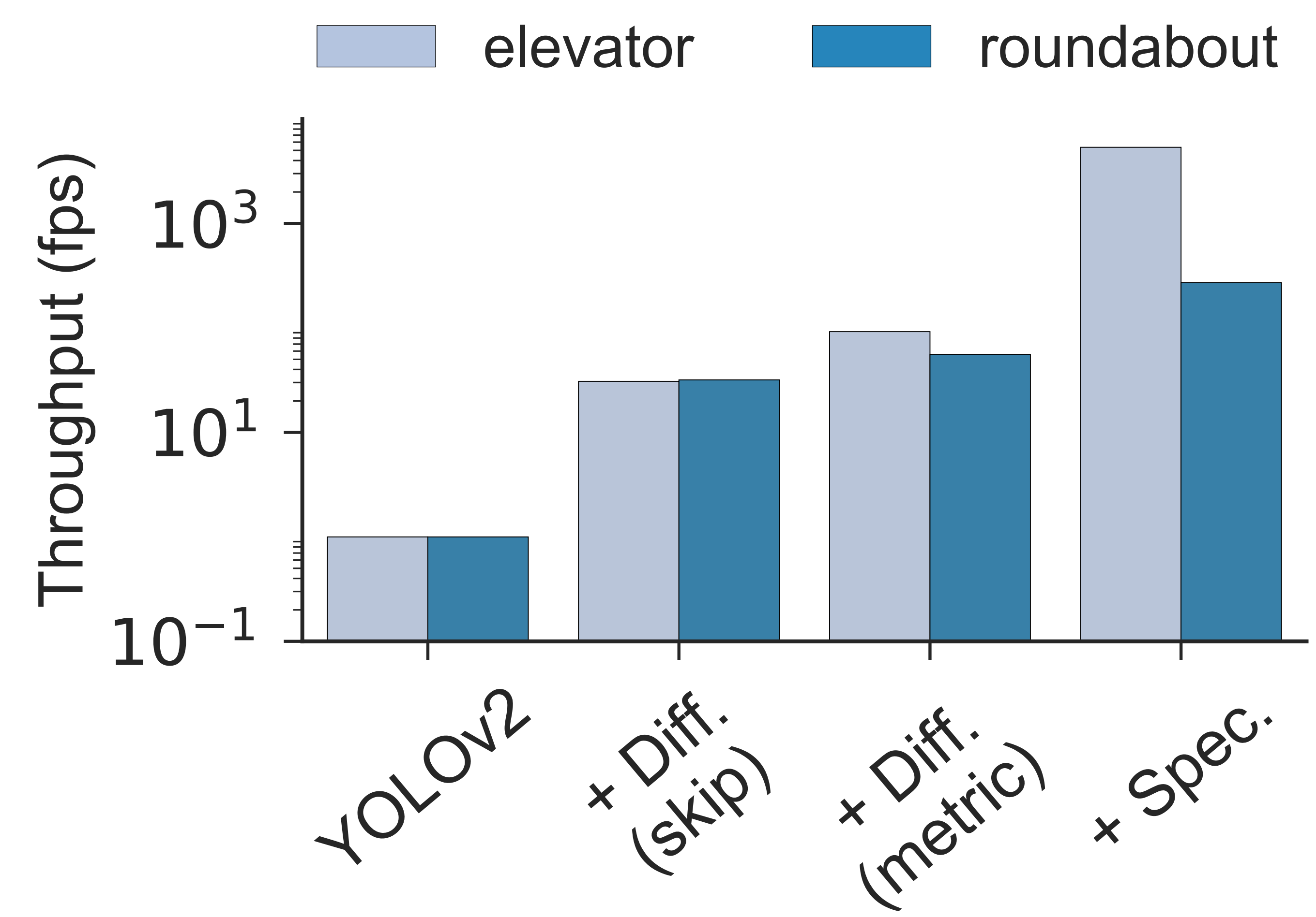
(d) night-street



(e) store



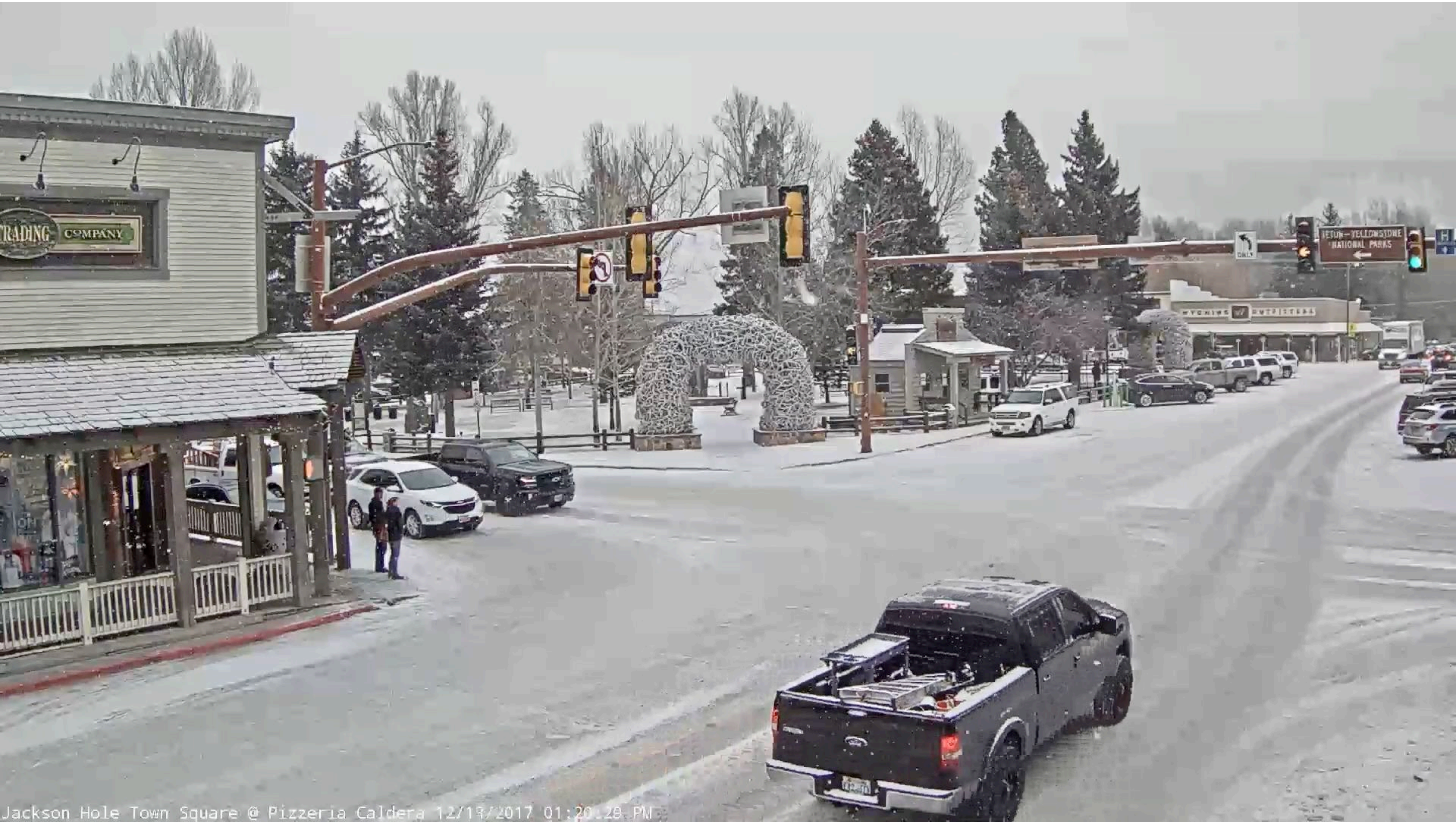
(f) elevator



Factor Analysis

* **Noscope actually performs a simpler classification task on a pre-cropped region of the viewport (not detection, which involves object location)**

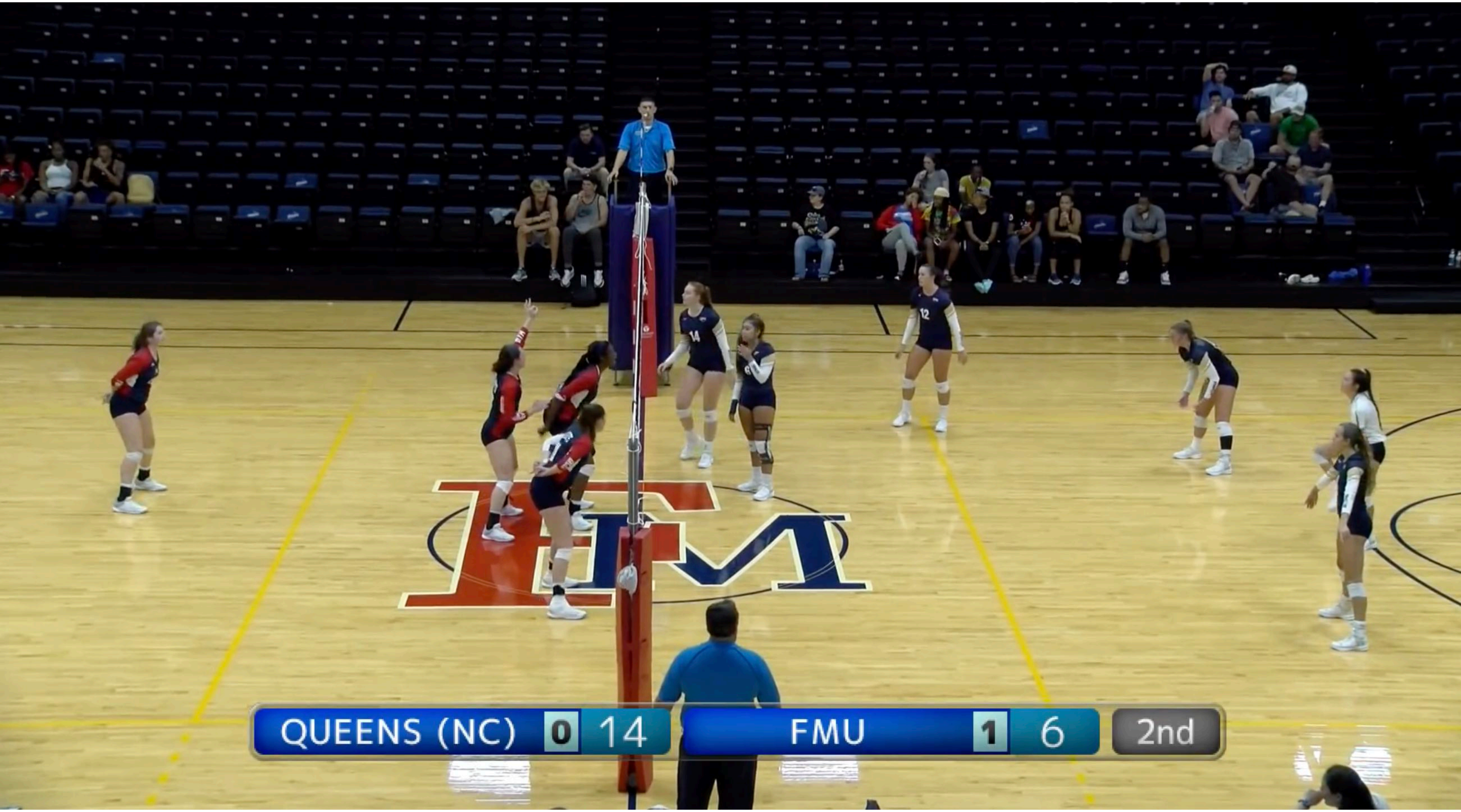
Example video



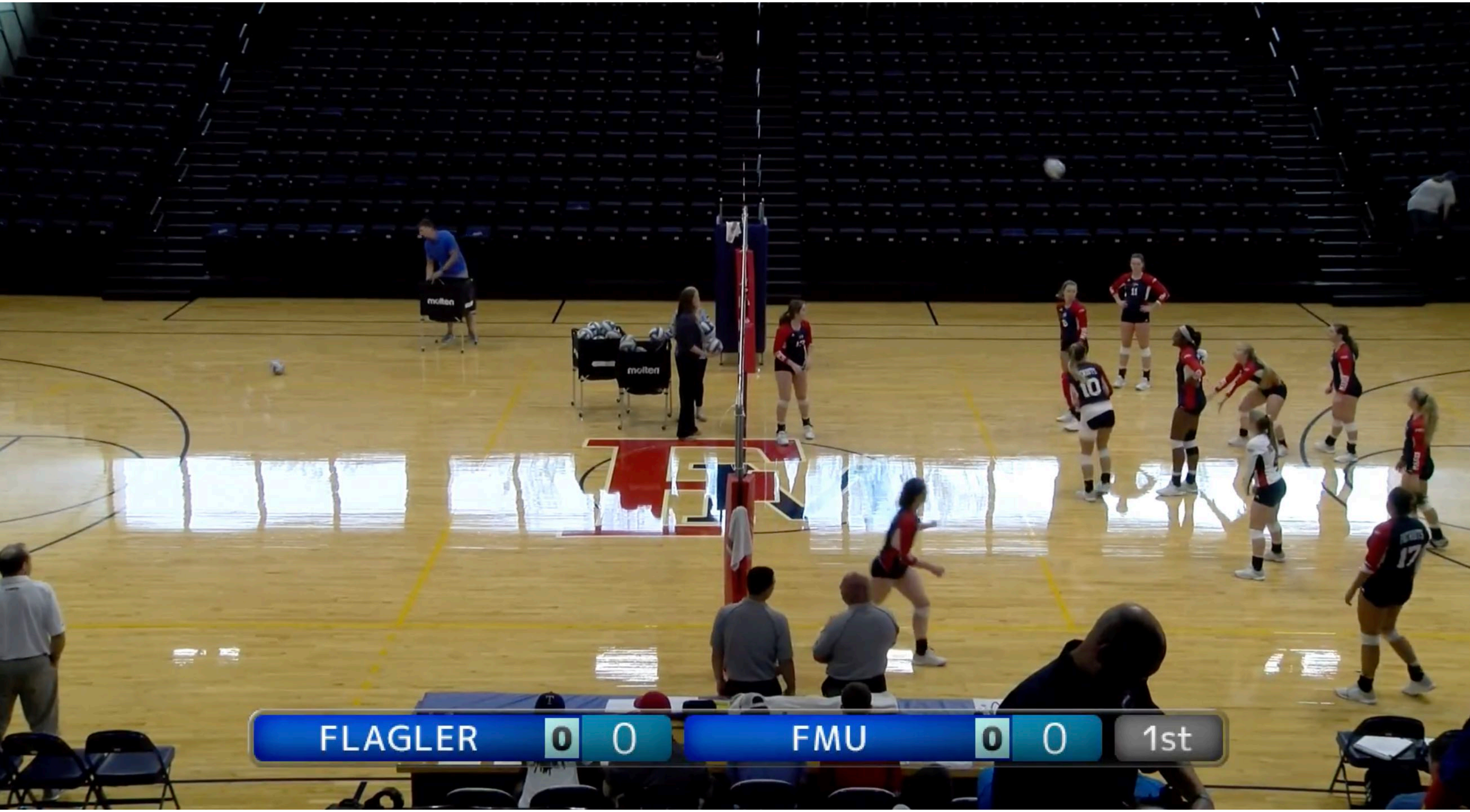
Problem: distribution shift

Weather, time-of-day, types of vehicles in view, etc…

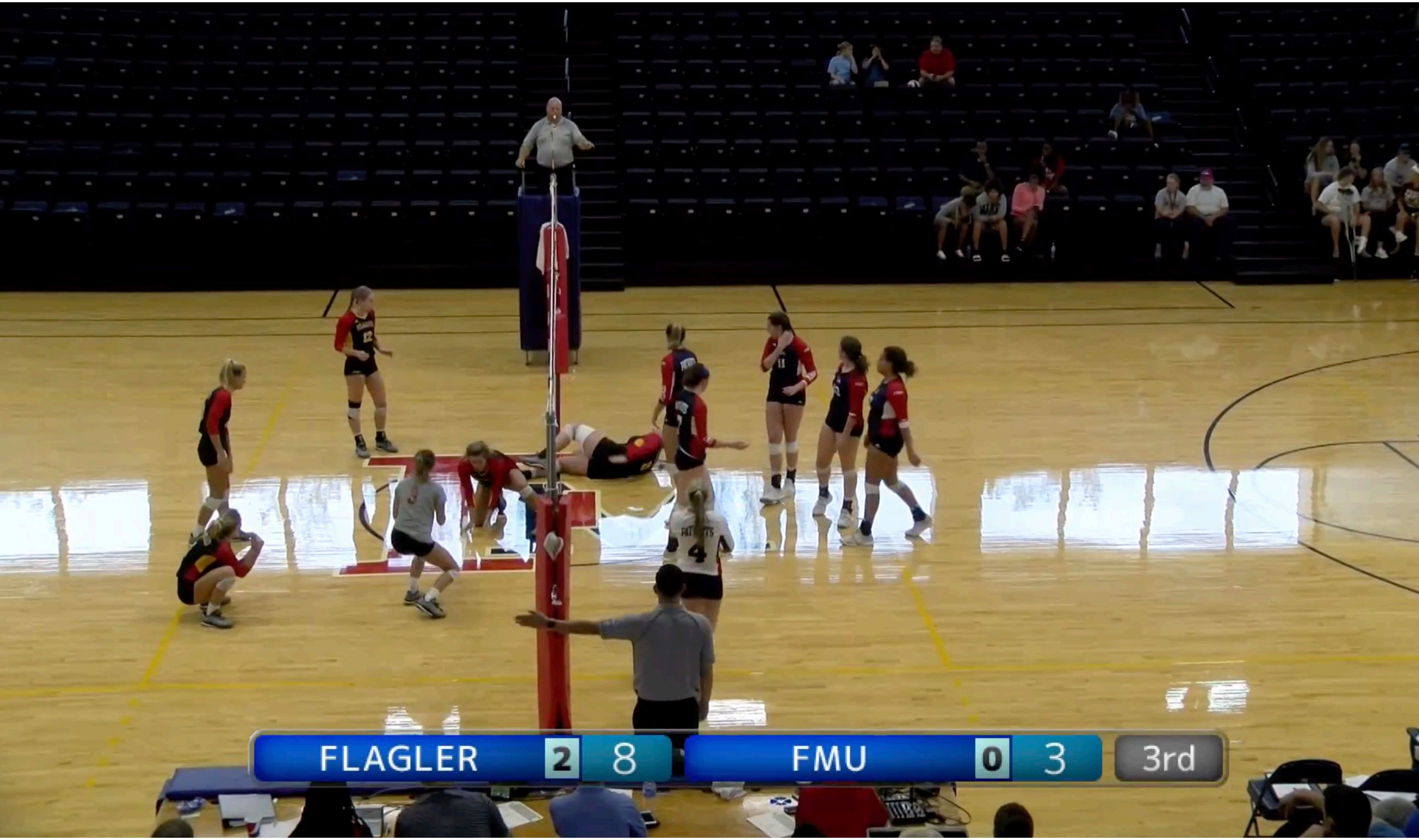
Challenge: distribution shift



Challenge: distribution shift



Challenge: distribution shift



Last example:
Specialize "up front"

Another example:
Periodically chose from a number of pre-specialized models"

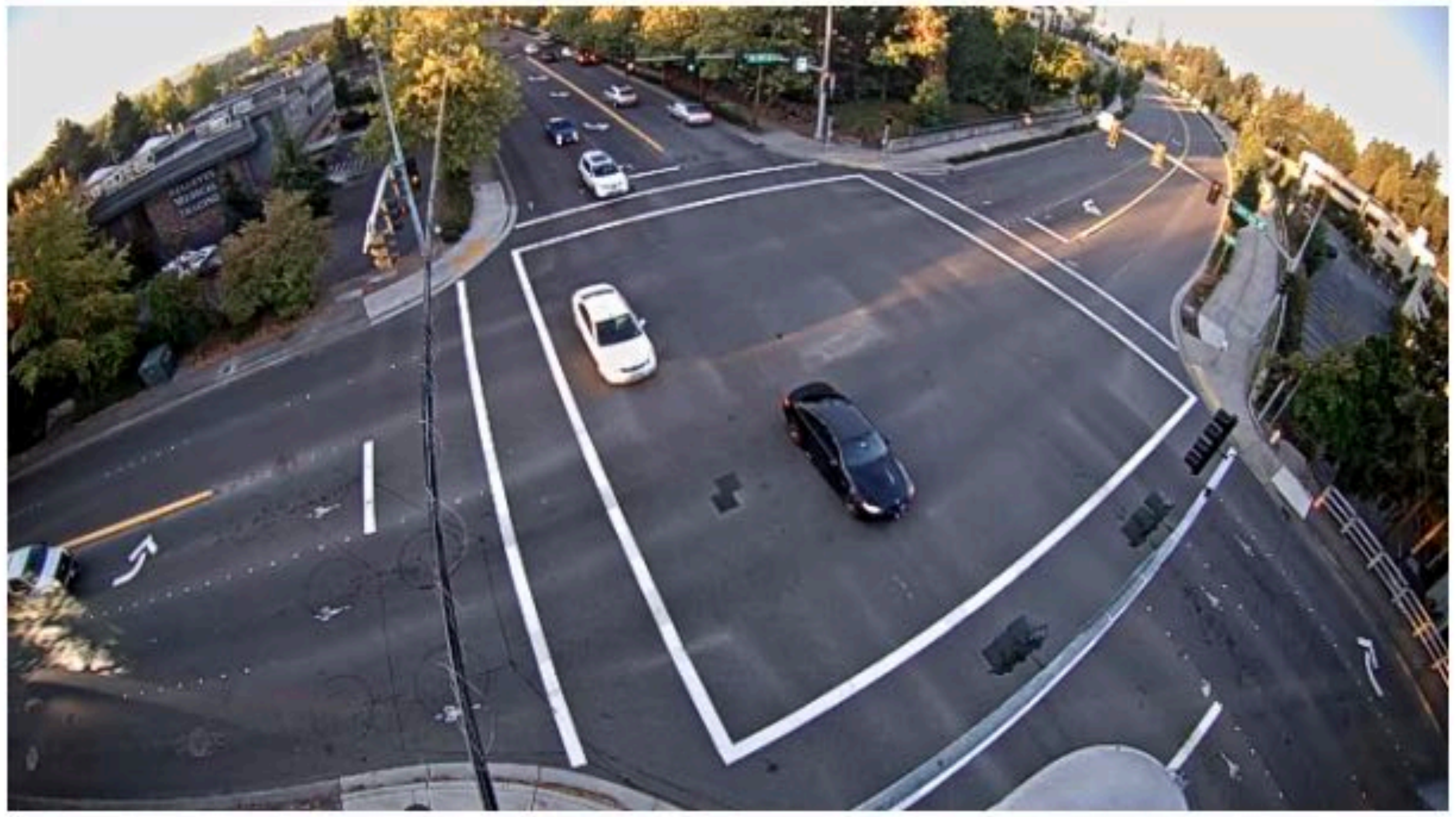
Chameleon

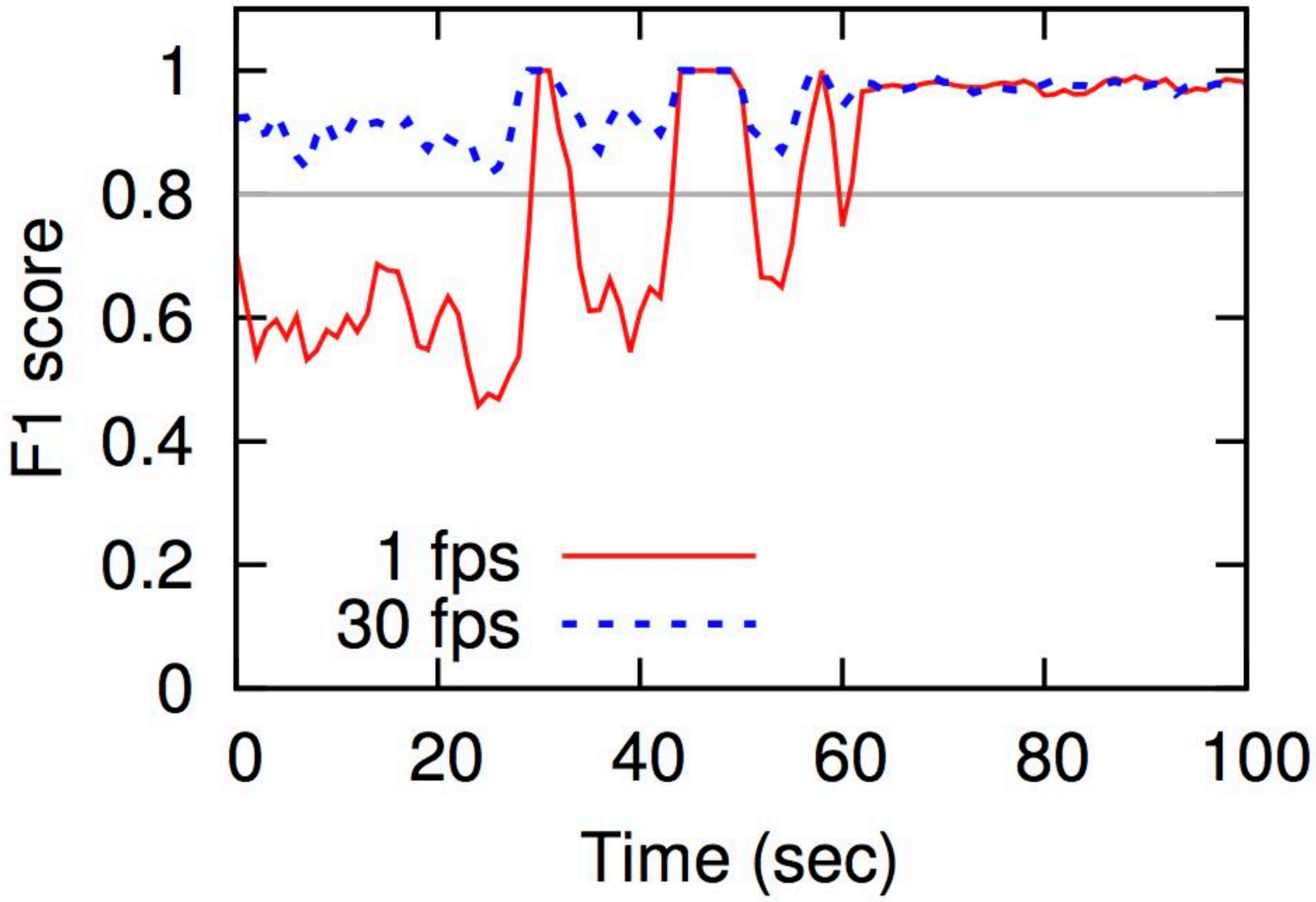
[Jiang 18]

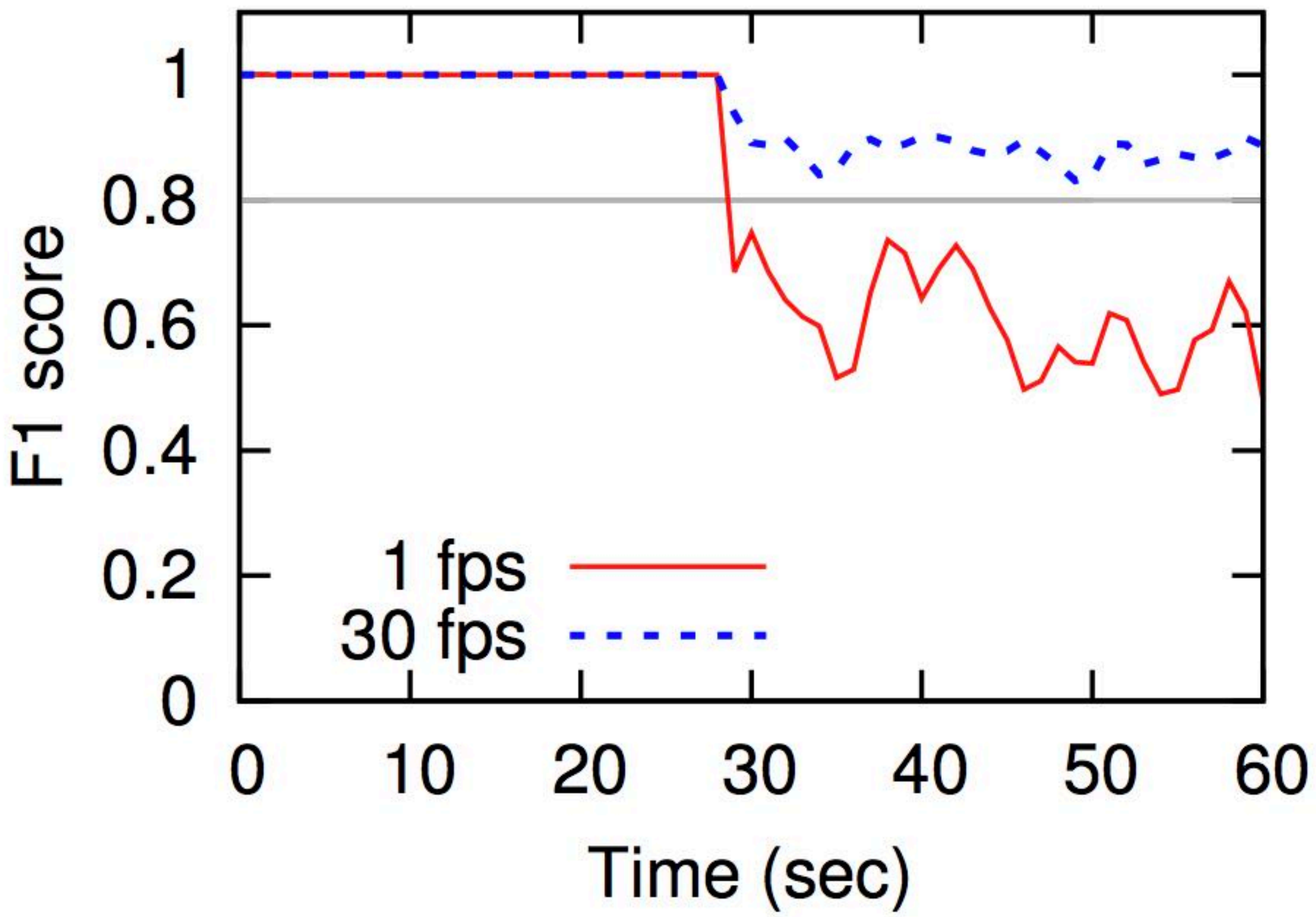
- Specialization strategy: choose among set of pre-trained models to find cheapest (sufficiently accurate) model for the job
  - "Knobs" to configure:
    - Input image resolution
    - Input image frame rate
    - DNN to use (Resnet101, Resnet50, Inception, MobileNet, etc.)
    - Thresholds on frame-to-frame difference detectors, etc.

Simple example

Appropriate frame-rate sampling depends on whether or not cars are moving

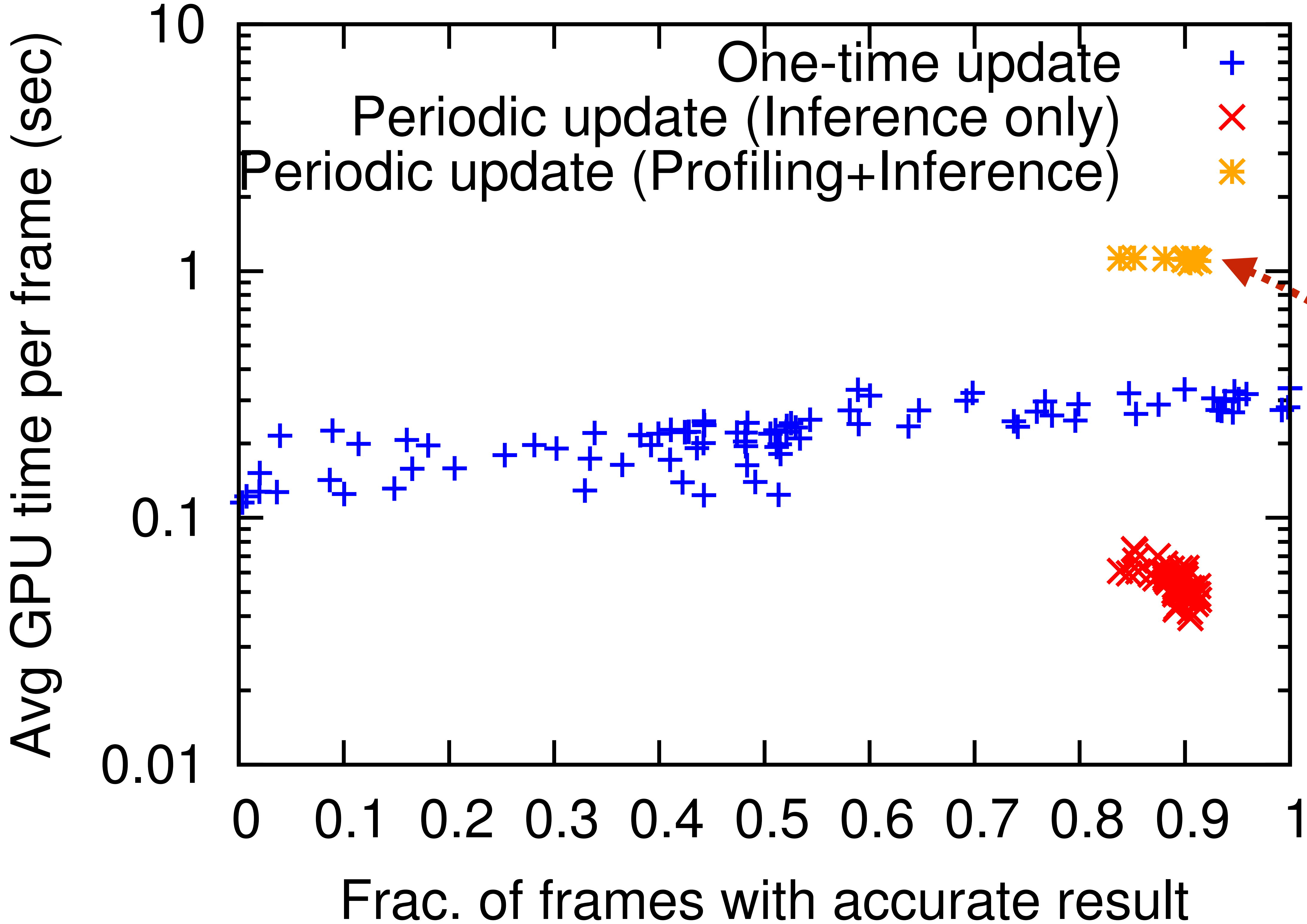






Challenge of distribution shift

- If distribution in video stream is non-stationary, cheap model determined via up-front profiling looses accuracy as contents of video change
  - Implication: choice of specialized model needs to be periodically changed



Results from object detection task on traffic camera video

Periodic update = every 4 seconds

Challenge: cost of profiling to adaptively determine which model to run eliminates potential benefits of model specialization

Reducing the cost of profiling

- The cost of profiling is running the candidate models at points in search space (profiling different values for all knobs)

- Idea 1: the set of most-likely-to-be-good models changes slowly over time

- Idea 2: visually similar streams are likely to have similar set of most-likely-to-be-good candidate models

Employing idea 1

- Assume model can change every video "segment" (e.g., 4 seconds)
- Profile all C model configurations for time segment 1
  - Retain top-K configurations
- Profile only top-K configurations in future segments
- Reset after N segments

Let S be number of segments before reset (~4)

Let K be size of candidate set ($K << C$)

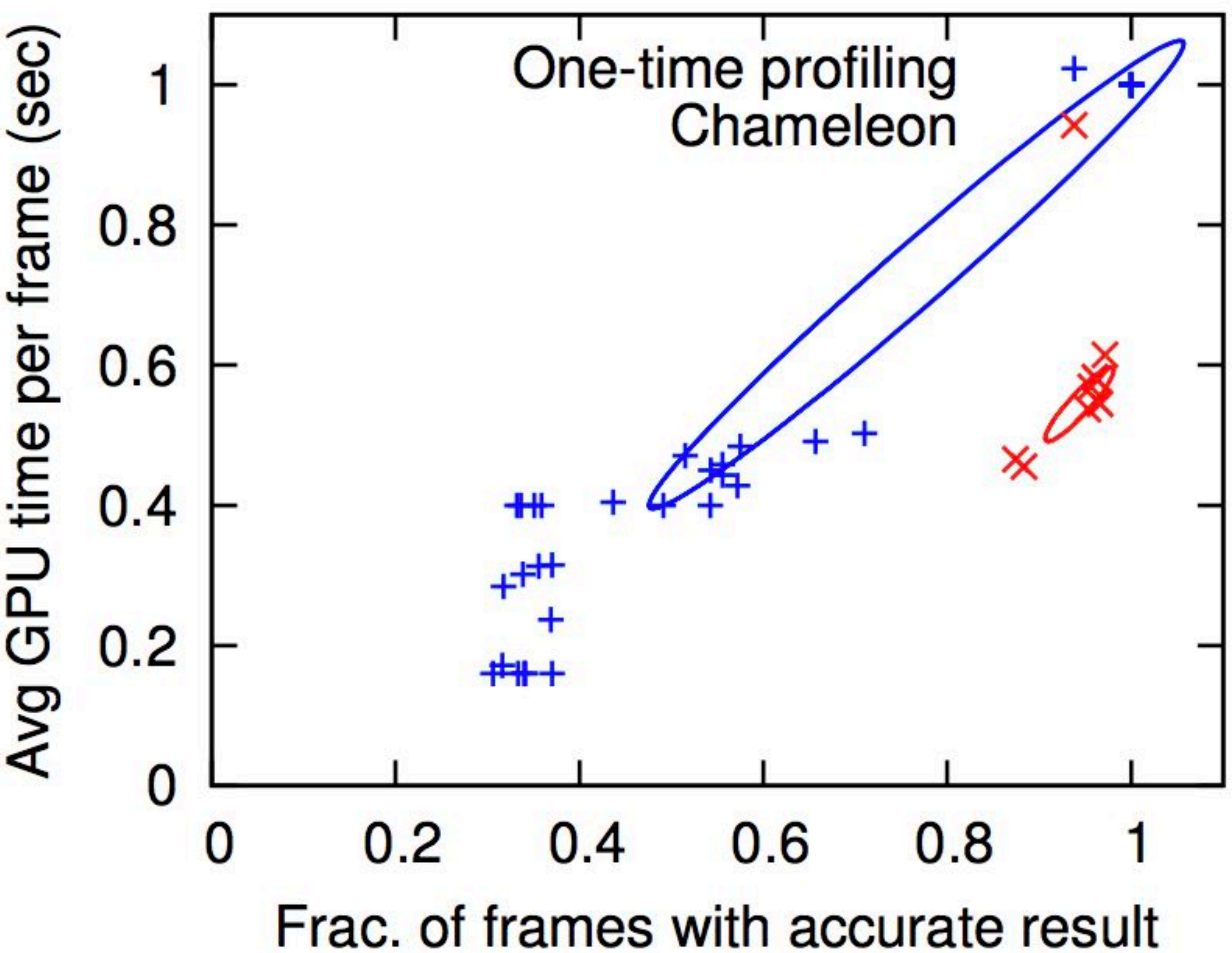
profiling cost $= C + (N-1) \times K << C \times N$

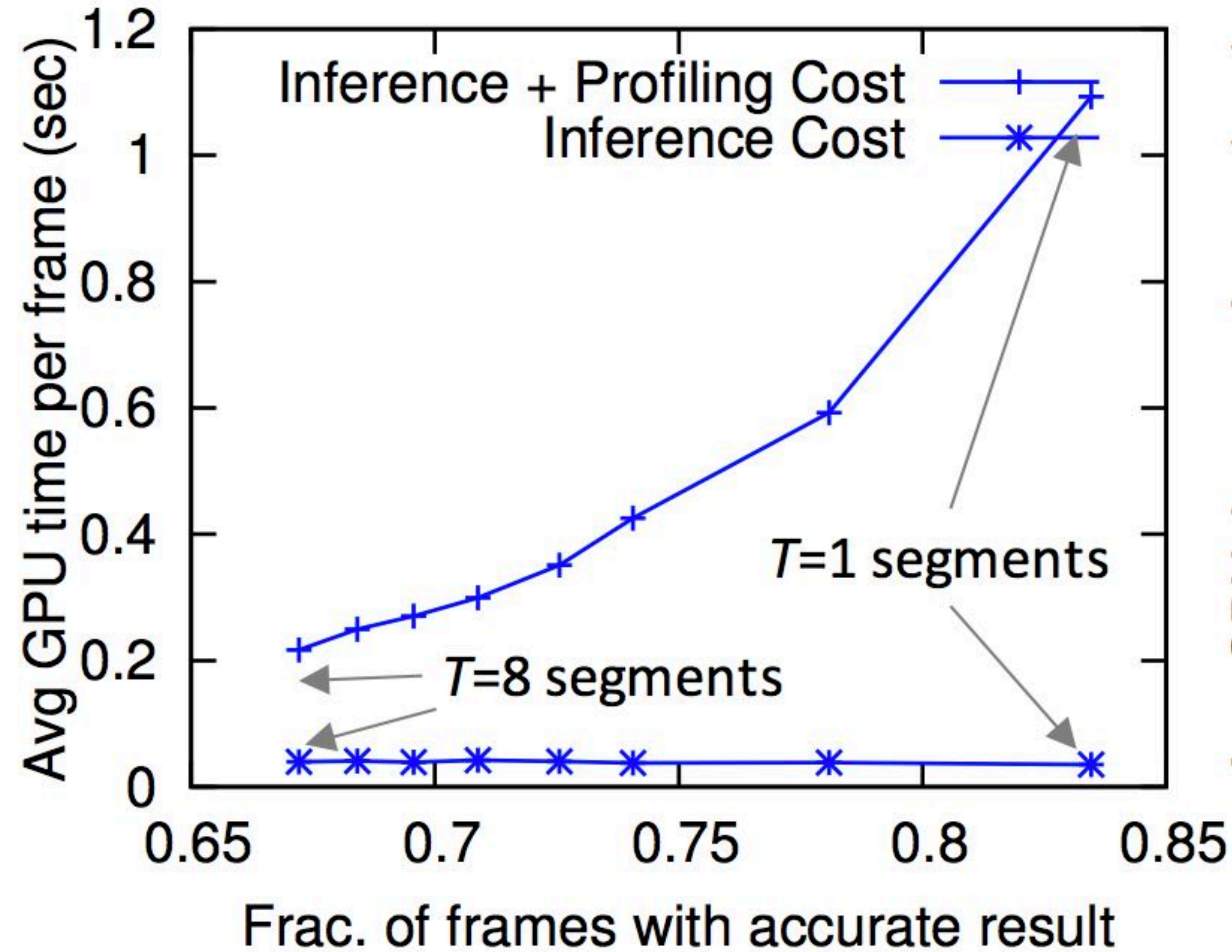
Assumption: bad model configurations tend to remain bad for longer periods of time

Employing idea 2

- Say there are many video cameras throughout a city
- Cluster video streams by visual similarity
- Only one camera per cluster needs to perform full profiling of the C configurations to identify top-K candidate set
  - Other cameras just perform top-K profiling

Intelligent profiling makes adaptive specialization profitable





Across dataset of multiple street light cameras,
when keeping accuracy similar, adaptive profiling over the 150 second test video yields 2-3X speedup compared to profiling once up front

But really the problem with profiling once is that accuracy is highly variable (see accuracy variance of blue crosses)

Specialize “up front”

Periodically chose from a number of pre-specialized models”

Re-specialize the model on the fly.

Problem: distribution shift

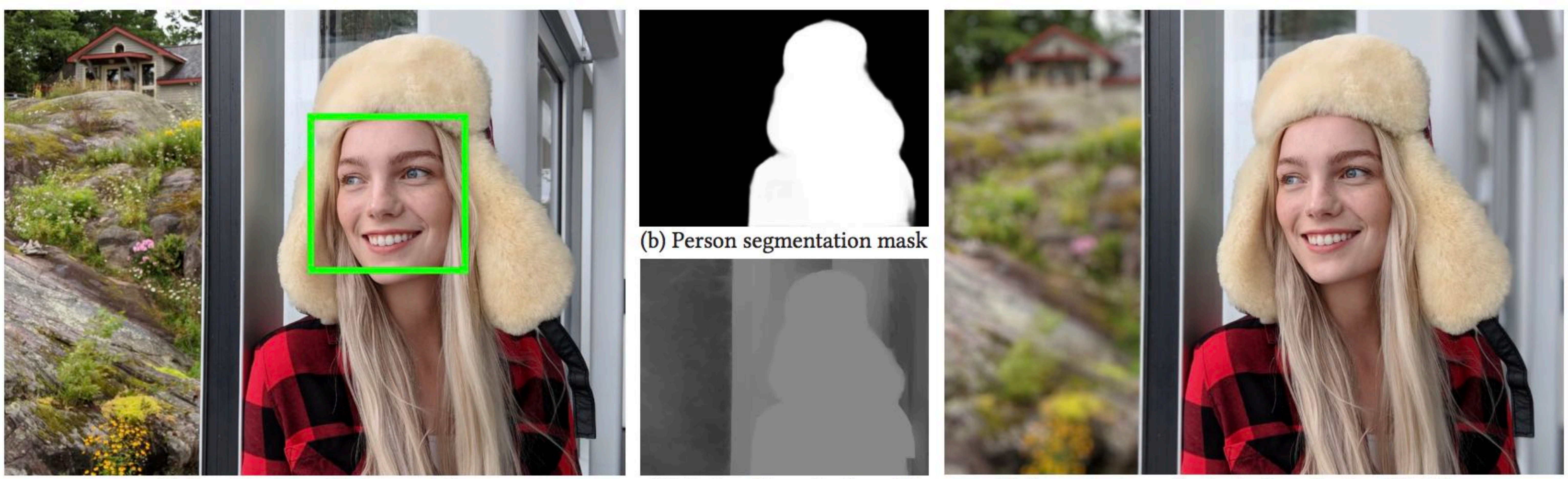
Weather, time-of-day, types of vehicles in view, etc…

Name of the game: good training data

“We cannot stress strongly enough the importance of good training data for this segmentation task: choosing a wide enough variety of poses, discarding poor training images, cleaning up inaccurate [ground truth] polygon masks, etc. With each improvement we made over a 9-month period in our training data, we observed the quality of our defocused portraits to improve commensurately.”

Synthetic Depth-of-Field with a Single-Camera Mobile Phone

NEAL WADHWA, RAHUL GARG, DAVID E. JACOBS, BRYAN E. FELDMAN, NORI KANAZAWA, ROBERT CARROLL, YAIR MOVSHOVITZ-ATTIAS, JONATHAN T. BARRON, YAEL PRITCH, and MARC LEVOY,
Google Research

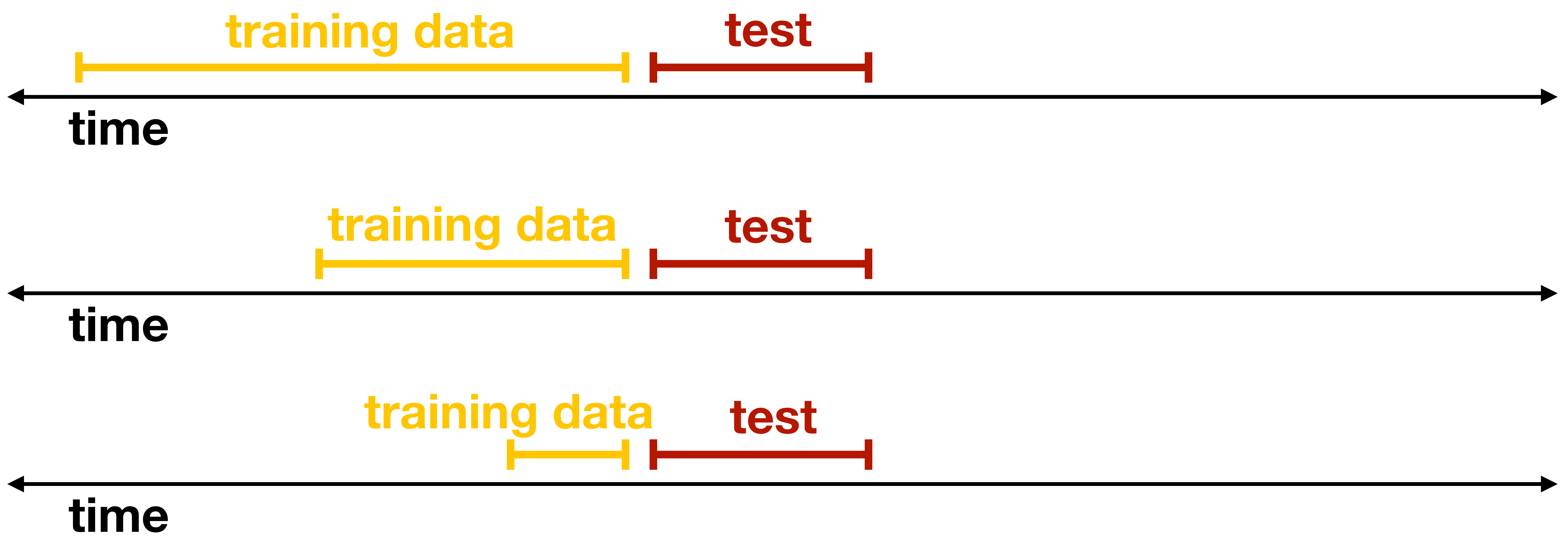


(b) Person segmentation mask

(a) Input image with detected face (c) Mask + disparity from DP (d) Our output synthetic shallow depth-of-field image

Experiment

- Plop camera down in a new environment
- We want a specialized (tiny) model for processing the stream from this camera
- How much data is needed to train the model?



Continuous model adaptation

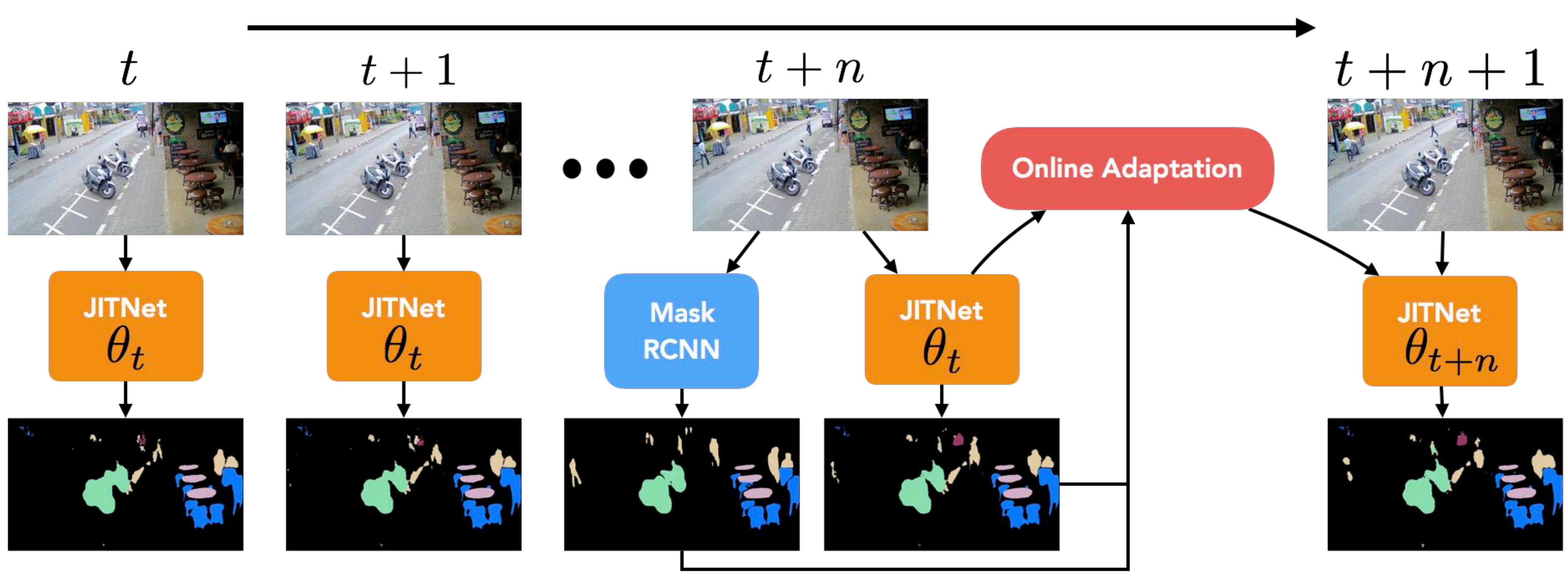
- **Tiny, efficient models can retain high accuracy for complex tasks in challenging environments *if they are continuously specialized to the contents of video streams***

- **A.k.a. Don't worry about carefully sampling everything you might see to create a good training set, just make sure you can adapt quickly online when you see it**

Example task: semantic segmentation

JITNet ("Just-in-time net")

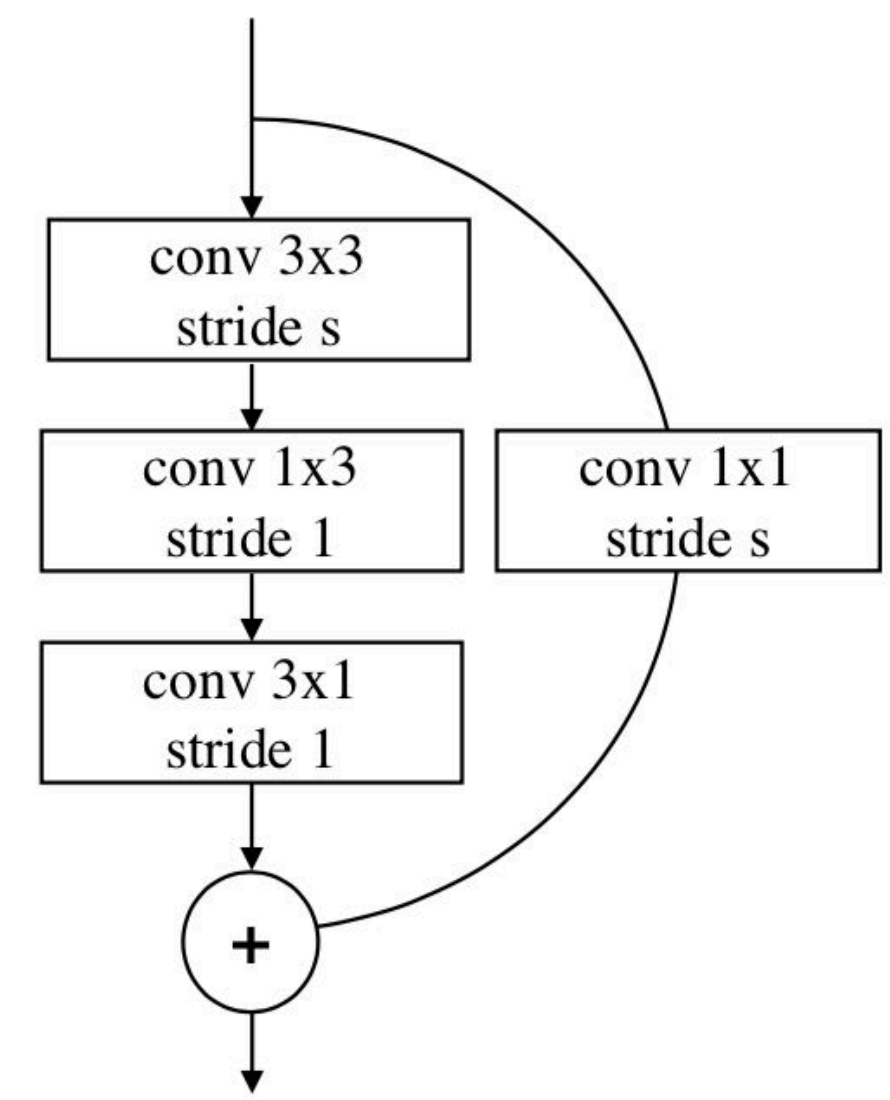
- Step 1: design a compact DNN that can evaluated and trained quickly
  - Our model: > 90x less flops for inference than Mask R-CNN
- Step 2: continuously retrain model as necessary as video stream evolves
  - Continuously train tiny "student" model to mimic output of expensive "teacher" model



JITNet model architecture

- Standard encoder-decoder with skips
- Each block is ResNet inspired (internal skips)

Input Size	Operation	s	r	c
1280 x 720	conv 3x3	2		8
640 x 360	conv 3x3	2		8
320 x 180	enc_block 1	2		64
160 x 90	enc_block 2	2		64
80 x 45	enc_block 3	2		128
40 x 23	dec_block 3	1	2	64
80 x 45	dec_block 2	1	2	32
160 x 90	dec_block 1	1	4	32
640 x 360	conv 3x3	1		32
640 x 360	conv 3x3	1	2	32
1280 x 720	conv 1x1	1		32



- 15.2B FLOPs for inference on 720p video
- Trains at high learning rates (0.01) and high momentum (0.9)

Online model distillation: results

Mask R-CNN
300ms/frame

Online JITNet
(~20x faster including training)

Online model distillation: results

Mask R-CNN
300ms/frame

Online JITNet
(~7.3x faster including training)

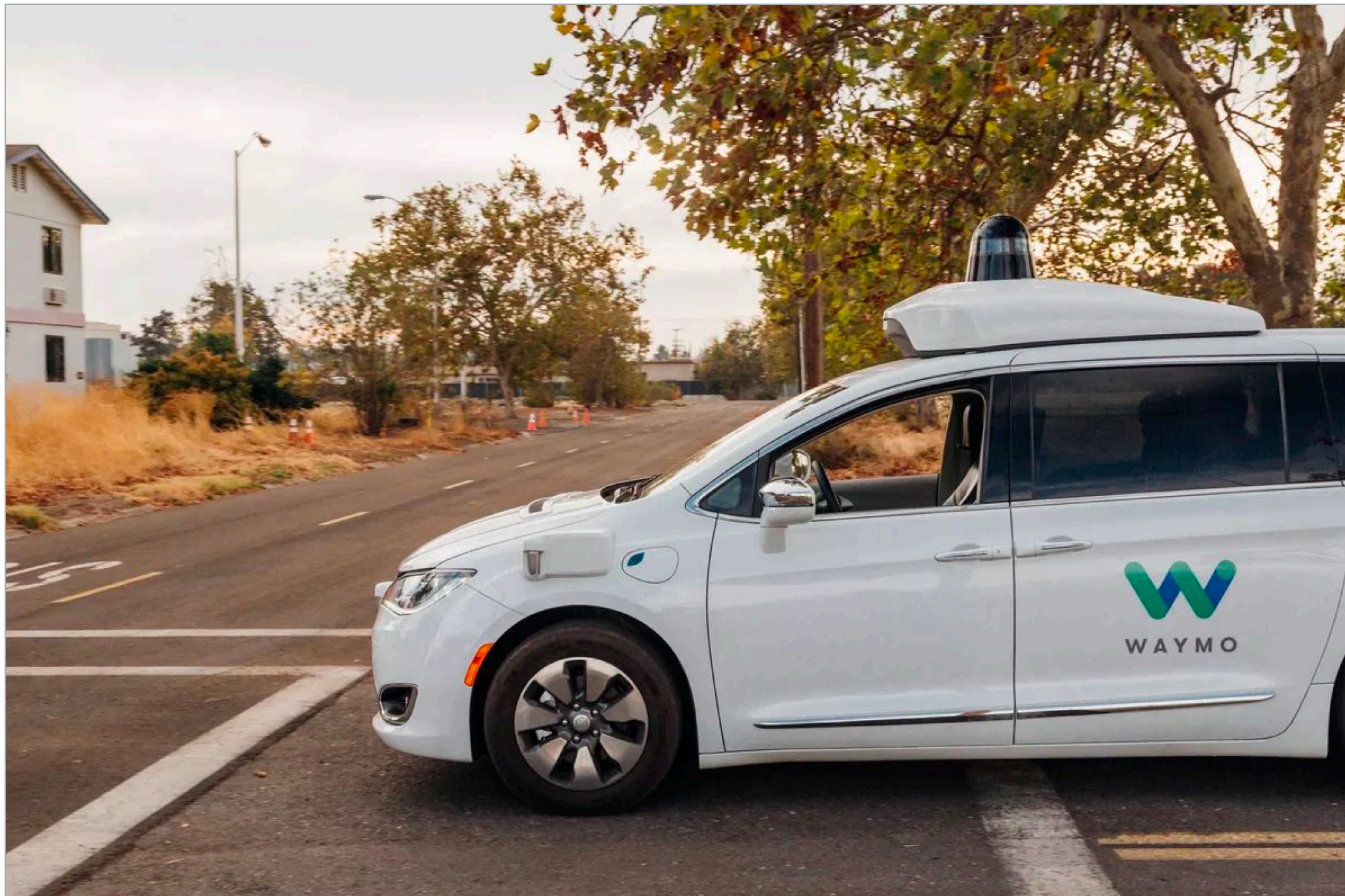
Online model distillation: results

Mask R-CNN
300ms/frame

Online JITNet
(~9x faster including training)

Discussion:
When should the cameras always be on?

Analyzing images for robot navigation



Analyzing images for urban efficiency

"Managing urban areas has become one of the most important development challenges of the 21st century. Our success or failure in building sustainable cities will be a major factor in the success of the post-2015 UN development agenda."

- UN Dept. of Economic and Social Affairs

Analyzing egocentric images to augment humans

Gwangjiang Market (Seoul)


DISCOVER CARD
TDK
Twizzlers
UBER
ADVANCED TECHNOLOGIES CENTER
amazon go

Some recent examples

Comprehensive capture of athlete performance

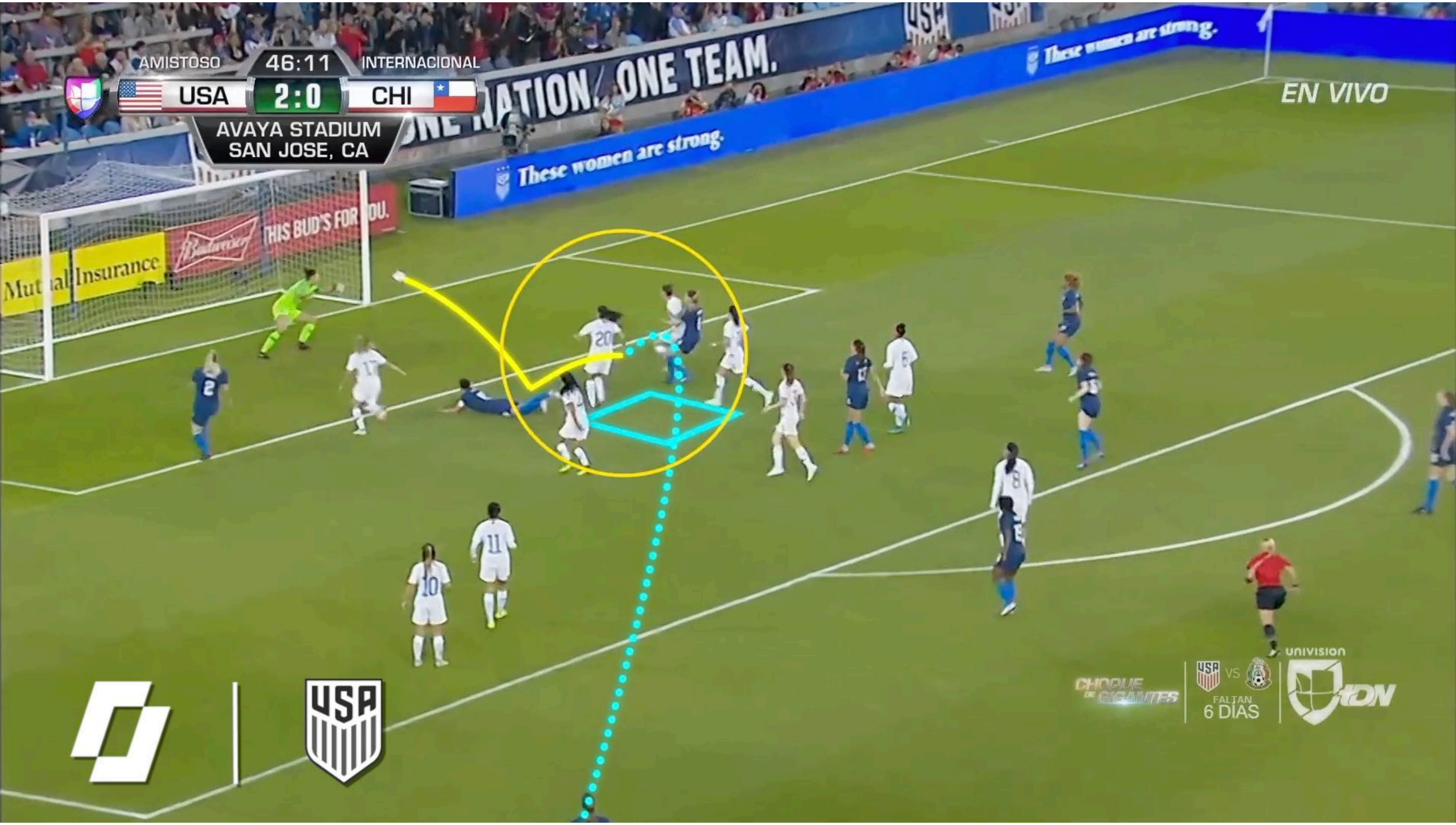


Image credit: Second Spectrum

Some recent examples

Comprehensive capture of athlete performance

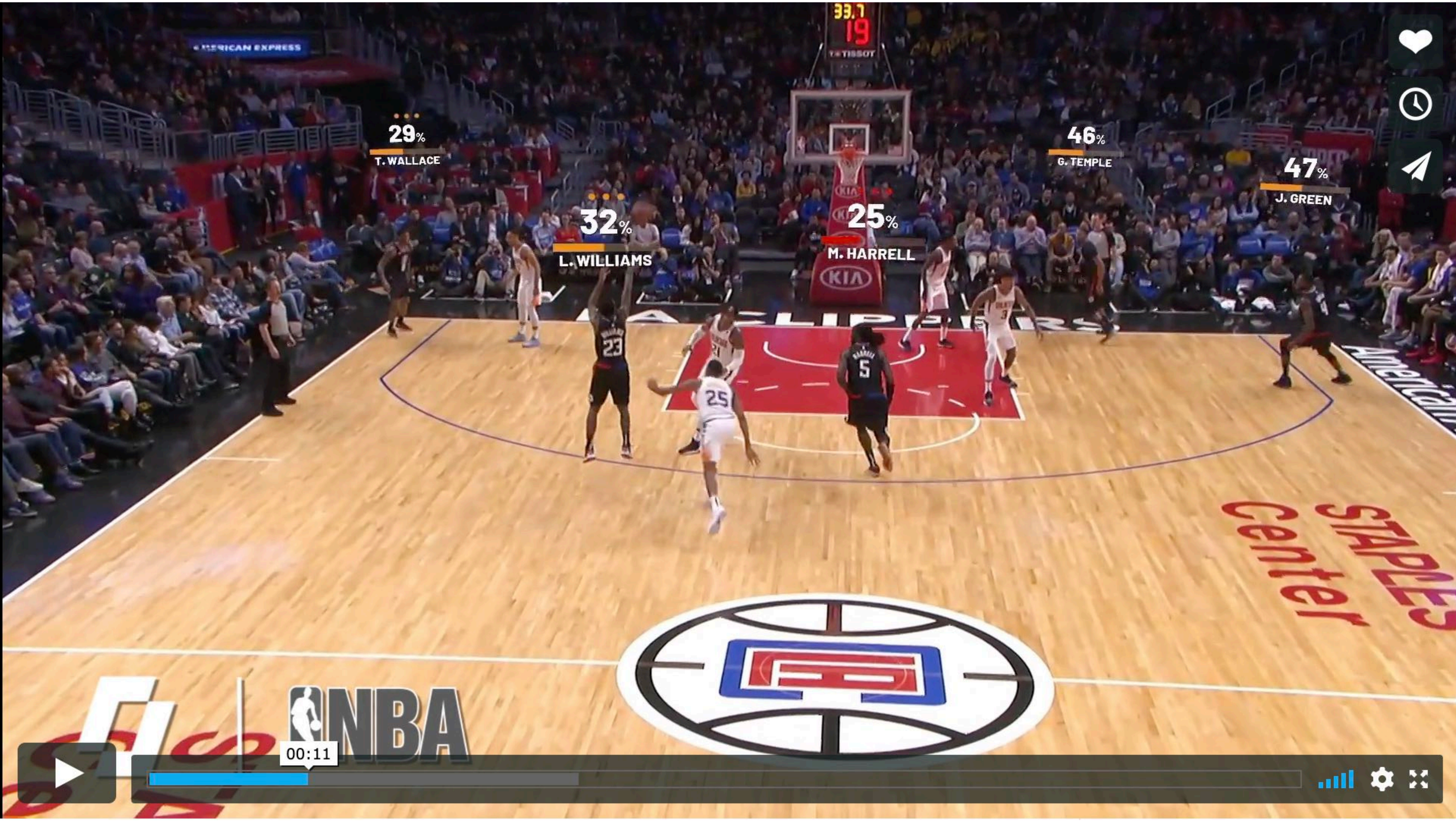
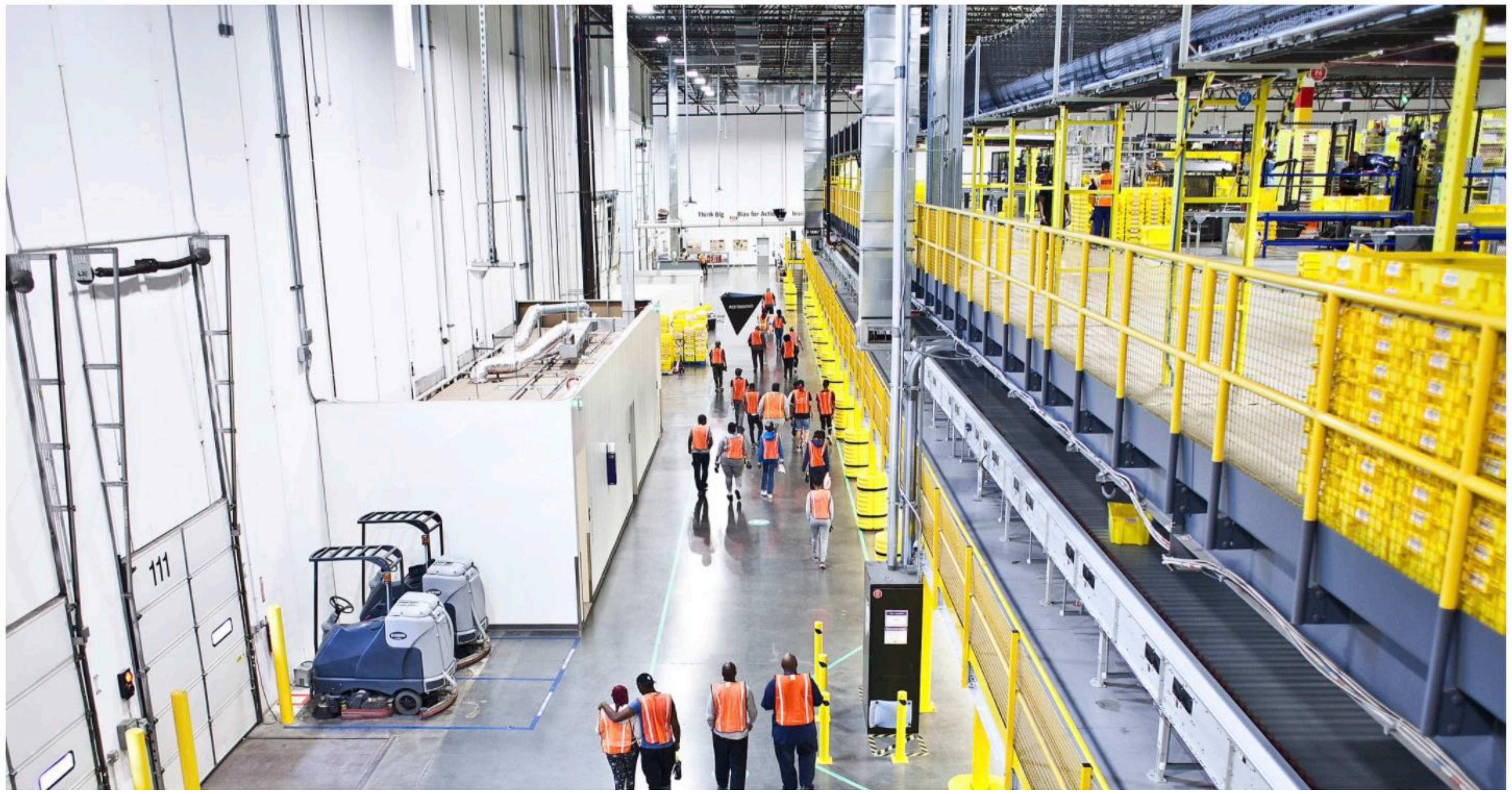


Image credit: Second Spectrum

Comprehensive capture of worker performance?

If Workers Slack Off, the Wristband Will Know. (And Amazon Has a Patent for It.)



By Ceylan Yeginsu

Feb. 1, 2018

118

Leer en español

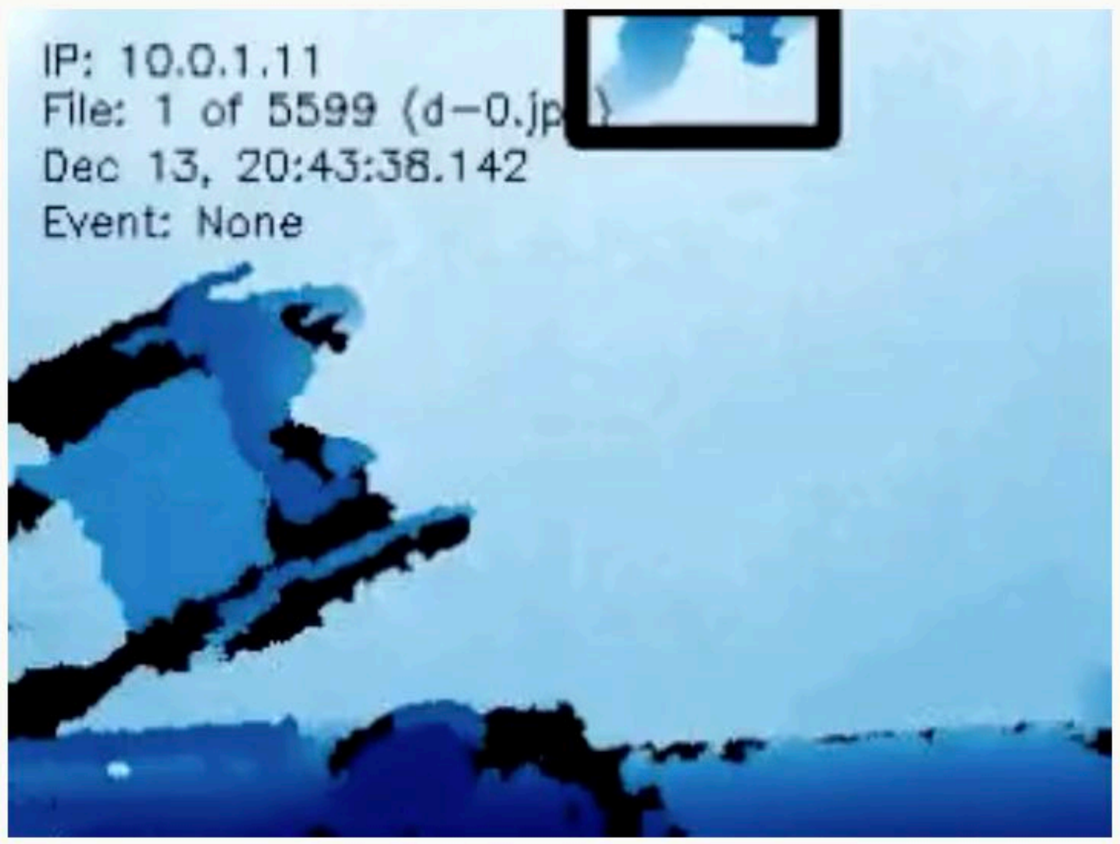
LONDON — What if your employer made you wear a wristband that tracked your every move, and that even nudged you via vibrations when it judged that you were doing something wrong?

What if your supervisor could identify every time you paused to scratch or fidget, and for how long you took a bathroom break?

Some recent examples

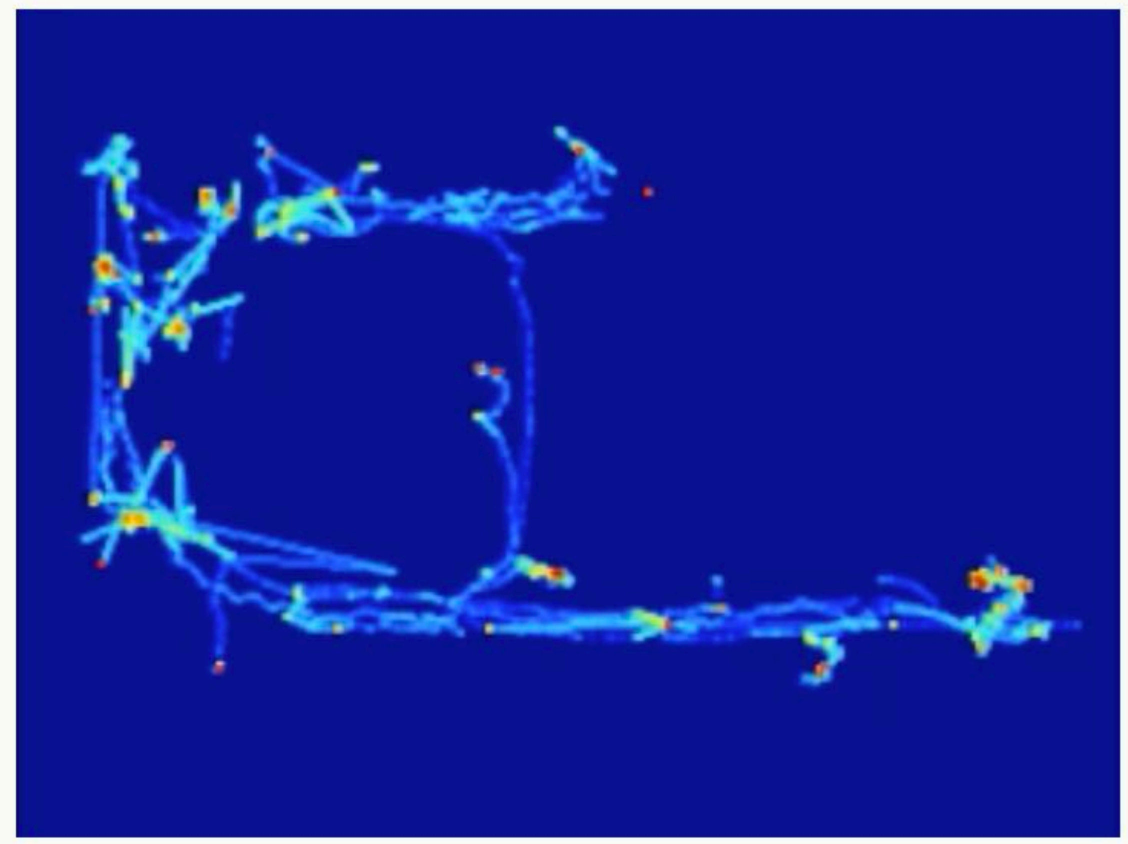
Surveillance of hospital workers (hand washing)

Dispenser Usage Detection



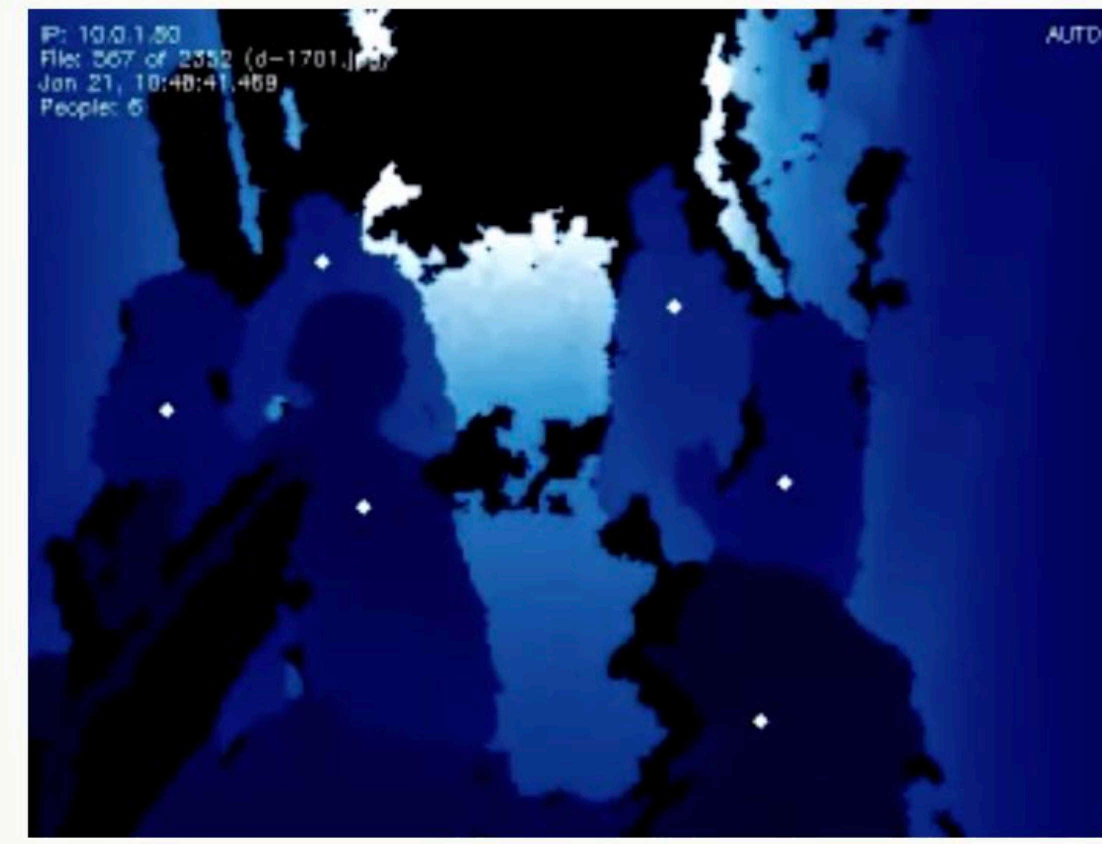
With the help of artificial neural networks, our method uses deep learning to automatically detect usage of an alcohol-based sanitizer dispenser from challenging ceiling-mounted top views.

Physical Space Analytics



Intuitive, qualitative results analyze human movement patterns and conduct spatial analytics which convey our method's interpretability. Red regions denote high traffic areas while blue denotes low traffic regions.

Privacy Safe Assessment



To comply with privacy regulations, we use de-identified depth images instead of color photos to track and analyze hand hygiene compliance. Our method can track multiple clinicians throughout a hospital ward.

Towards Vision-Based Smart Hospitals: A System for Tracking and Monitoring Hand Hygiene Compliance
Haque et al. 2017

Surveillance for contact tracing

MARKETS BUSINESS INVESTING TECH POLITICS CNBC TV

Use of surveillance to fight coronavirus raises concerns about government power after pandemic ends

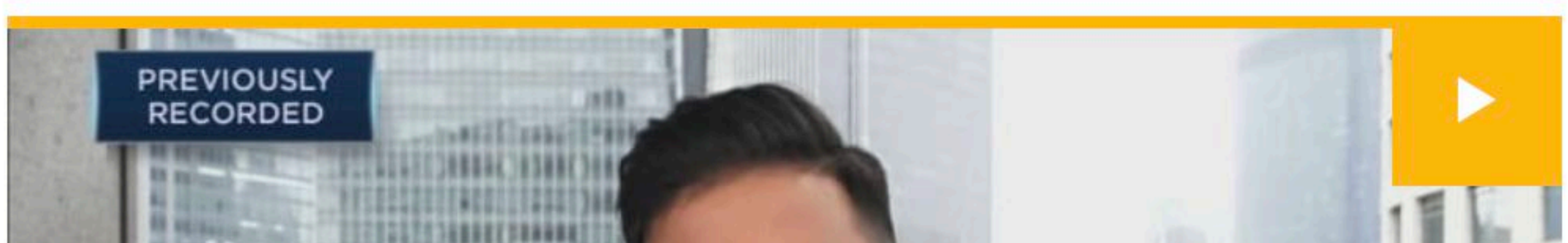
PUBLISHED THU, MAR 26 2020·7:58 PM EDT | UPDATED MON, MAR 30 2020·12:17 PM EDT

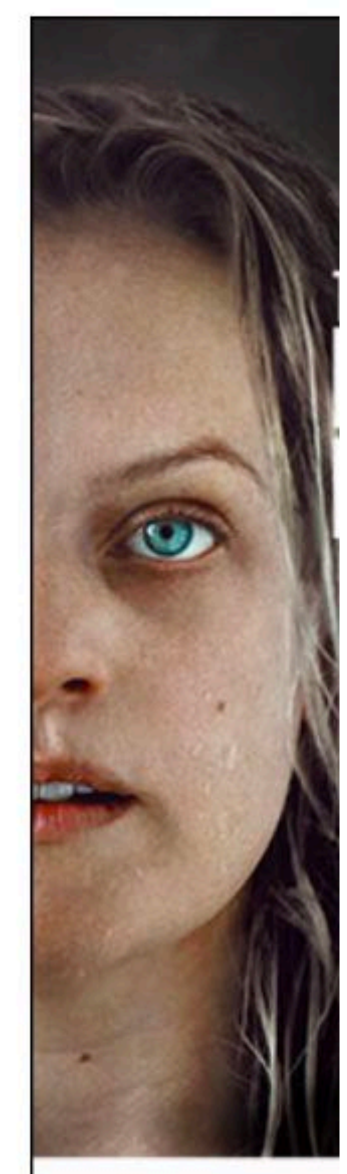
Arjun Kharpal

SHARE

KEY POINTS

- China mobilized its mass surveillance tools, from drones to CCTV cameras, to monitor quarantined people and track the spread of the coronavirus.
- Other nations like Israel, Singapore and South Korea are also using a combination of location data, video camera footage and credit card information, to track COVID-19 in their countries.
- But privacy experts raised concerns about how governments were using the data, how it was being stored and the potential for authorities to maintain heightened levels of surveillance — even after the coronavirus pandemic is over.





Privacy and ethics in a world with always-on video

Amazon's Rekognition messes up, matches 28 lawmakers to mugshots

ACLU: "And running the entire test cost us $12.33—less than a large pizza."

CYRUS FARIVAR - 7/26/2018, 5:00 AM

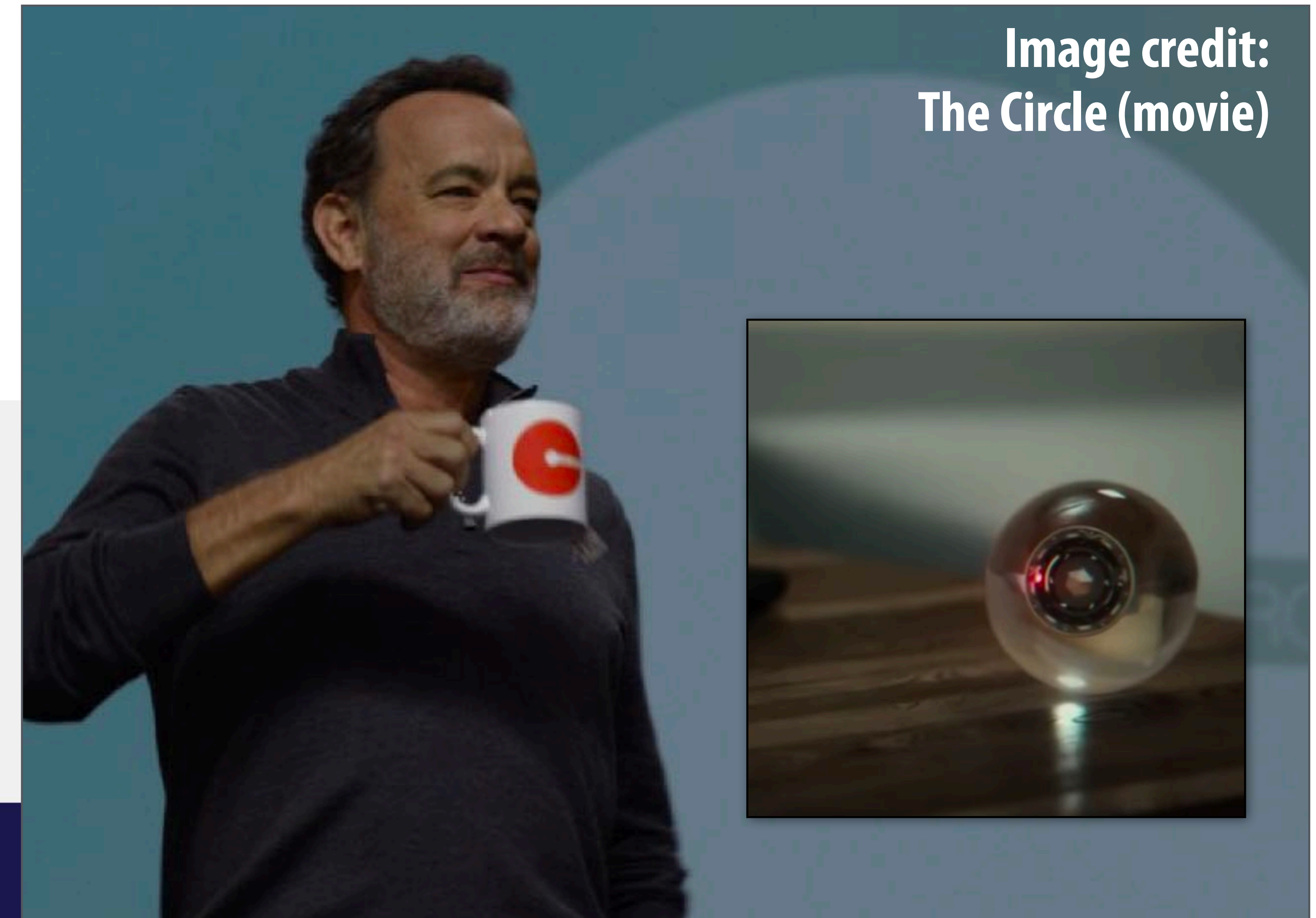


Image credit:
The Circle (movie)

Discussion:
What are your standards for when observational technology is reasonable to be deployed?

What safeguards (both technical and non-technical) should be put in place to protect privacy?

Summary

- An increasing number of cameras across the world will be capturing near continuous video

- Many applications will seek to extract value from these data streams
  - Implications for efficiency of cities (transportation, infrastructure monitoring), brick-and-mortar commerce, security, health-care, robotics, human-robot interactions, autonomous vehicles

- Need significant efficiency gains to process this worldwide visual signal
  - We've already talked about hardware specialization
  - Today's focus: specialization of model to video stream or scene context