Google Research

Directions in Machine Learning

Michele Covell
Google Research

with many thanks to Rahul Sunkthankar and many Google Machine Perception Teams for the slide materials

Google | YouTube

Directions in ML

Need to understand **intent** and **importance** in audio and video

- Major progress on: detection, categorization, embeddings, 3D modeling, people-centric information, action/interaction recognition
- missing piece: need to determine **what to leave out**
    - want a **synopsis** for **authored** (and situated) media
    - want less constrained interactions for **live** situations

Generating new **creative** content can help highlight shortcomings

(as well as providing useful content)

Understanding: Starting from...

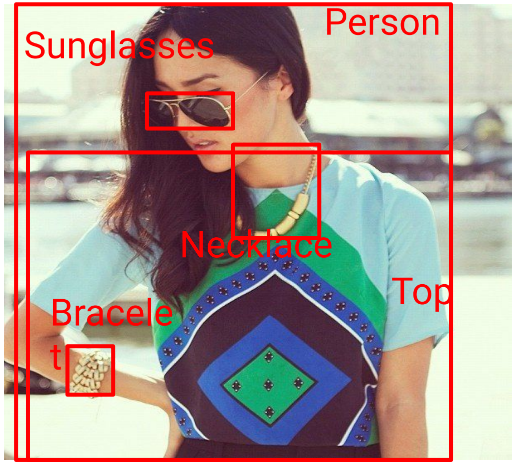
Classification

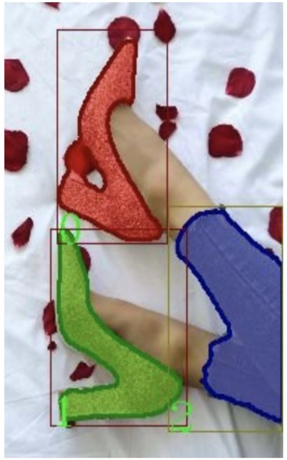
Images ⇒ Labels

Cake (0.93669)
Dessert (0.91911)
Birthday (0.89697)
Child (0.89183)
Fondant (0.88340)
Birthday cake (0.87525)

Detection

Images ⇒ Labelled Boxes or Regions





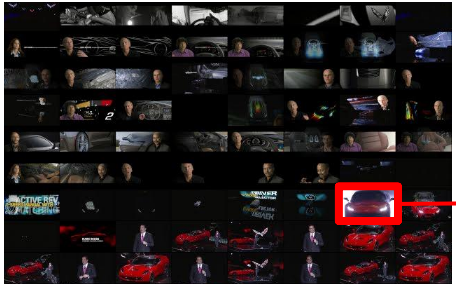
Embedding

Images ⇒ Features

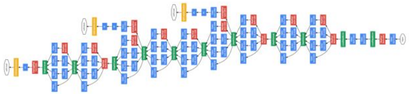
Similar Images and Labels

Deep-learned visual features

[Natsev, et al.]



Videos with noisy labels



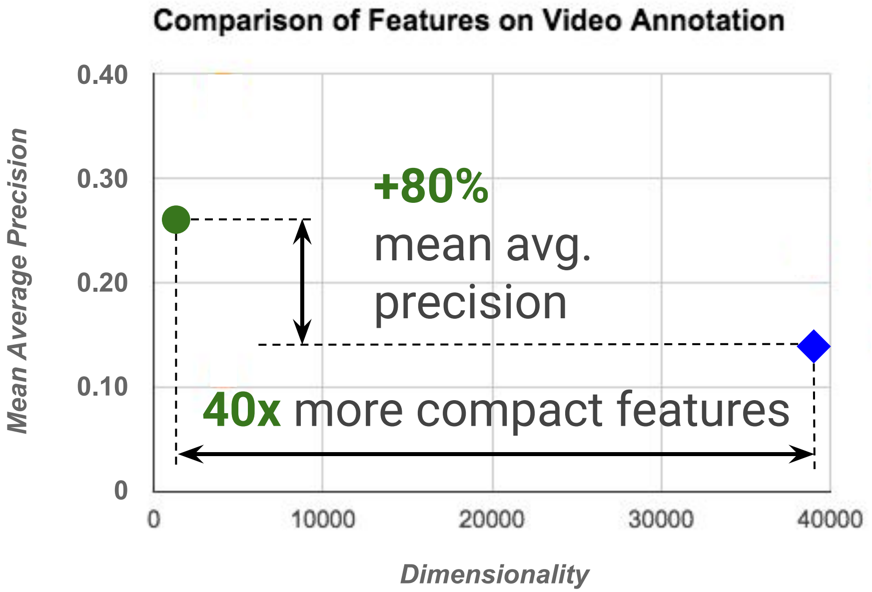
Inception model trained on noisy data (images)

Frame-level

Bottleneck embedding layer (1000-d)

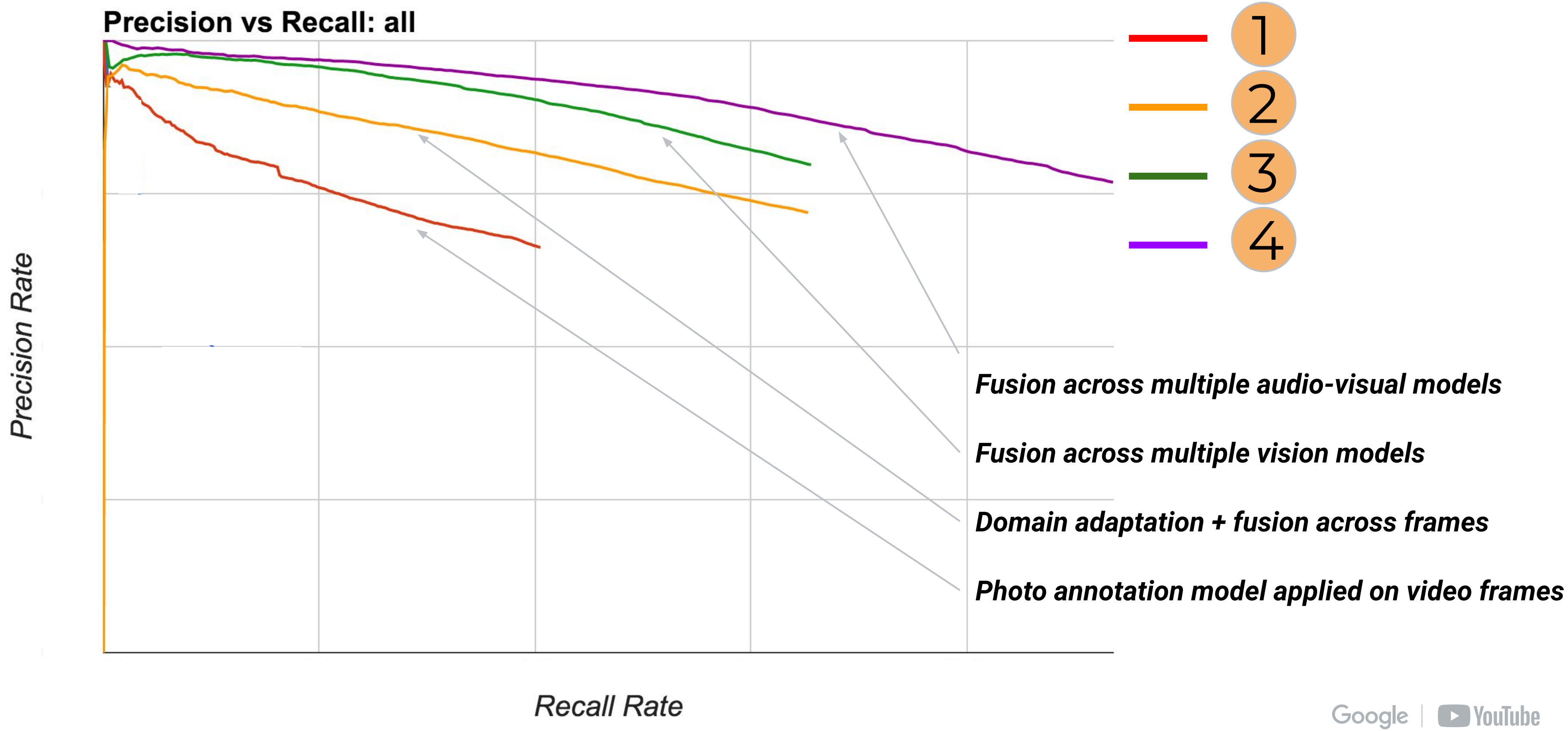
Video-level

- Max pooling
- Avg pooling
- VLAD pooling

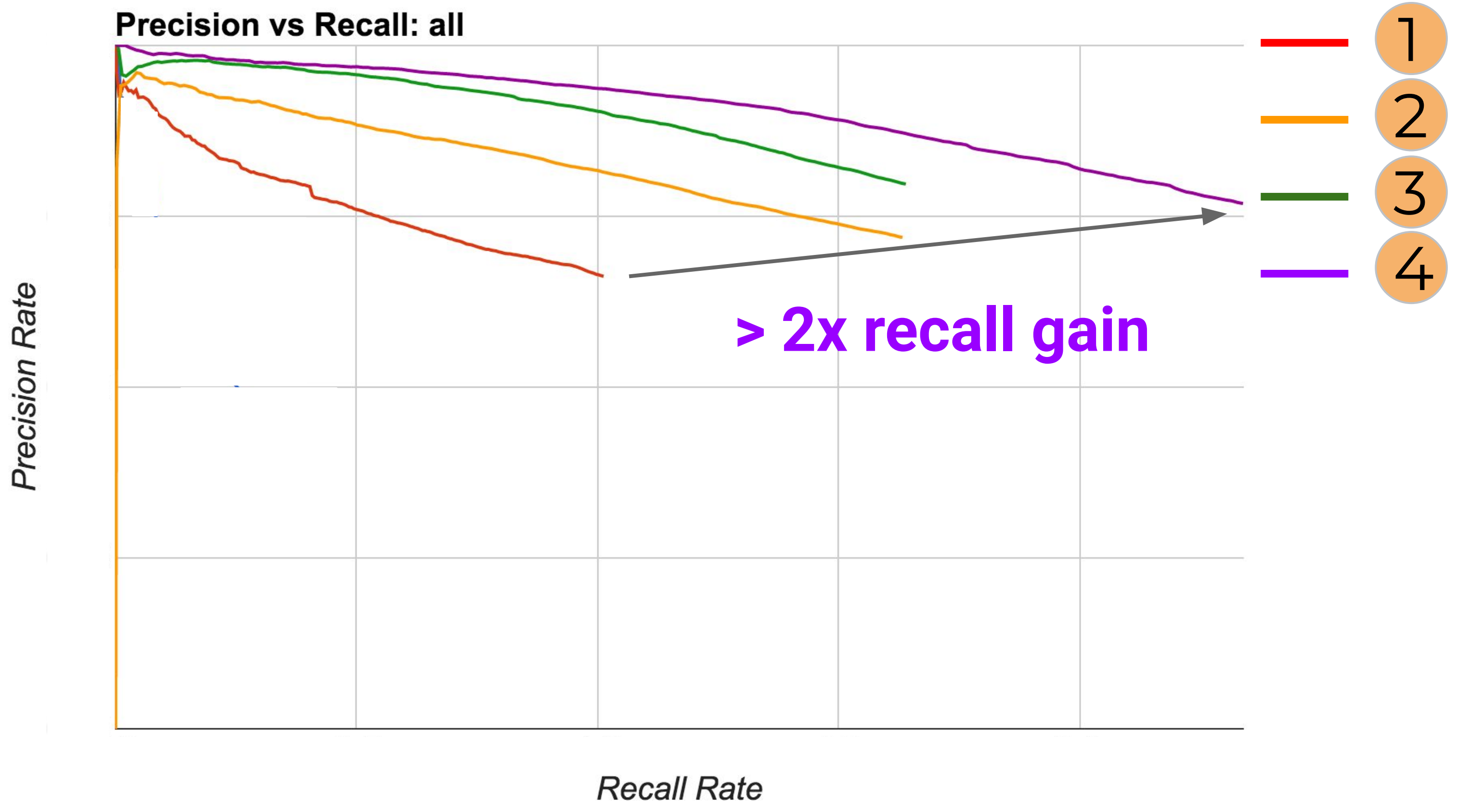


- Deep learned visual features, VLAD coding: 1024-d, 0.272 MAP
- Handcrafted audio-visual features: ~40K-d, 0.153 MAP

Improving detection in video (starting from images)

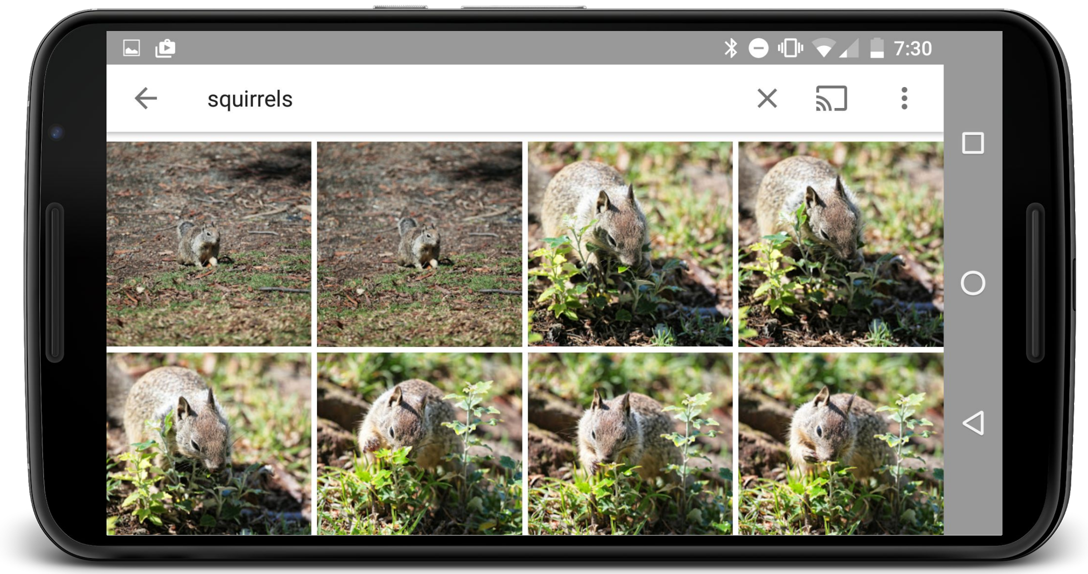


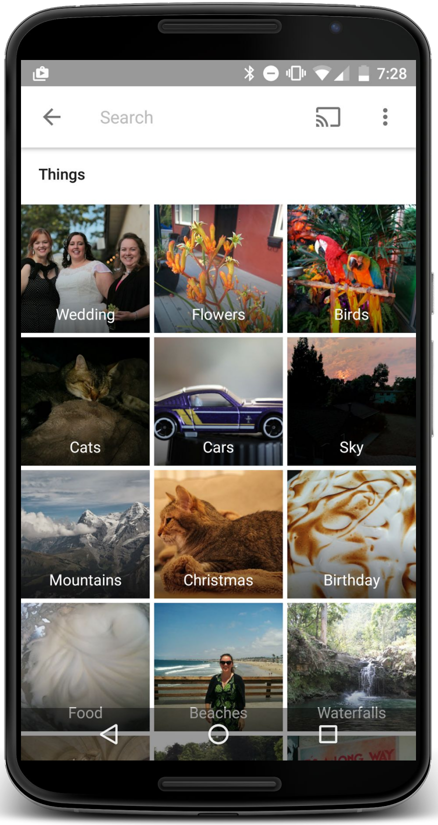
Improving detection in video (starting from images)



Images/Videos ⇒ Labels

[Veit, Alldrin, Chechik, et al.]





Automatic tagging and search!

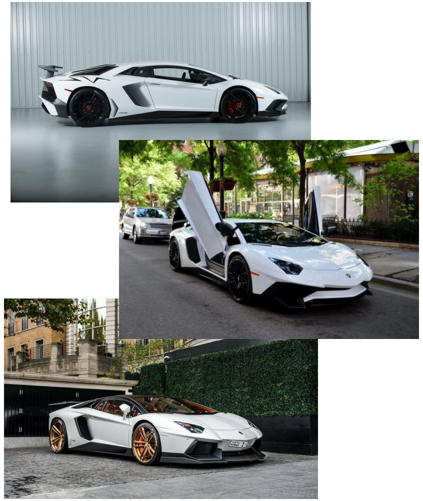
Google | YouTube

Spectrum of semantic similarity

[Juan, Lu, Li, et al.]

Instance -> fine-grained -> coarse-grained

2016 lamborghini aventador white



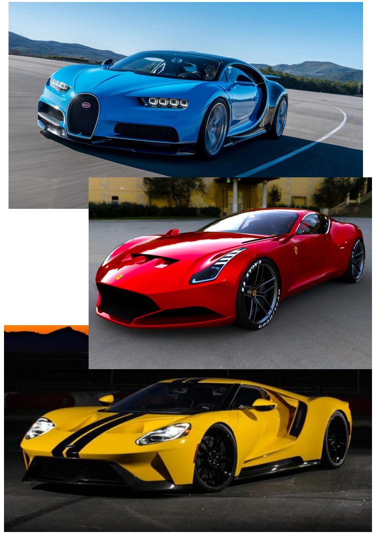
lamborghini aventador red

lamborghini aventador camouflage

lamborghini huracan

2010 lamborghini Gallardo

super car



Disentangled visual similarity

[Veit, Belongie, Karaletsos]

Disentangled visual attributes.

The model is able to rank/filter images by different attribute dimensions.

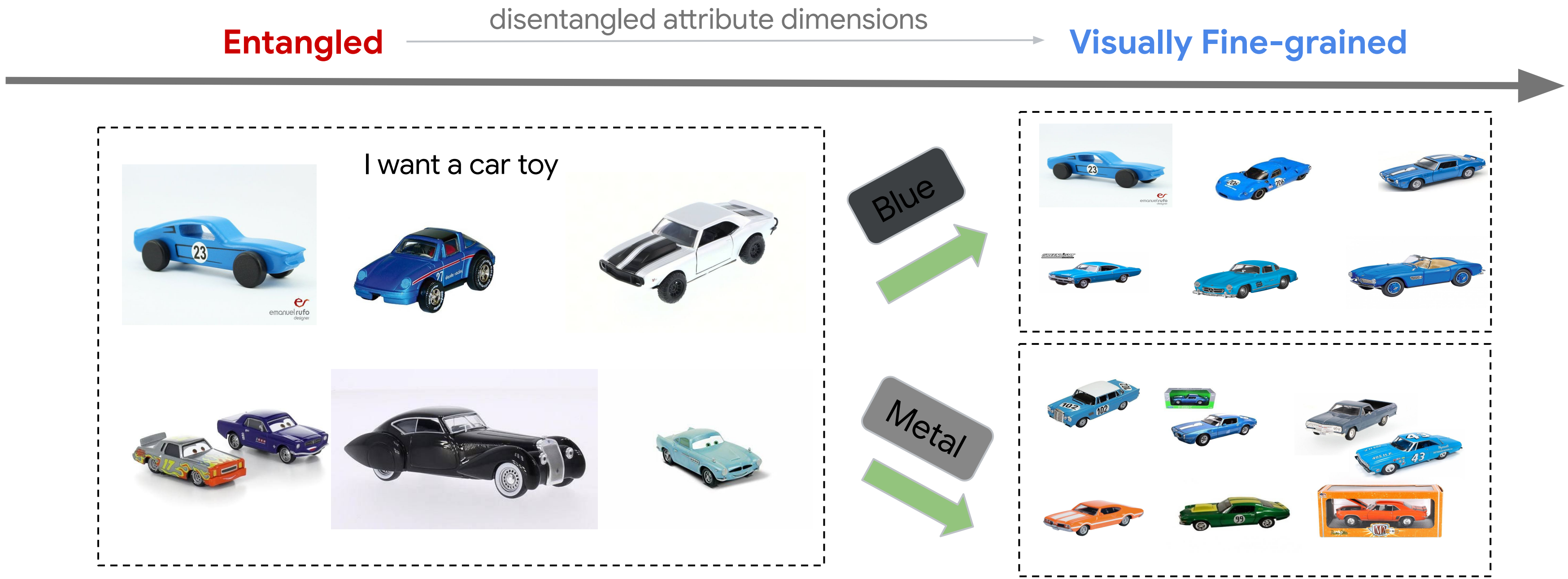


Image-text co-embedding

[Juan, Lu, Li, et al.]

Bridge the visual and text domain with multitask, multimodal learning

- **Multimodal**: Knowledge in one domain helps learning in the other
- **Multitask**: vast amount of visual-text and NLP data at Google
- Generic models to be used / fine-tuned for cross-domain inference / learning cases

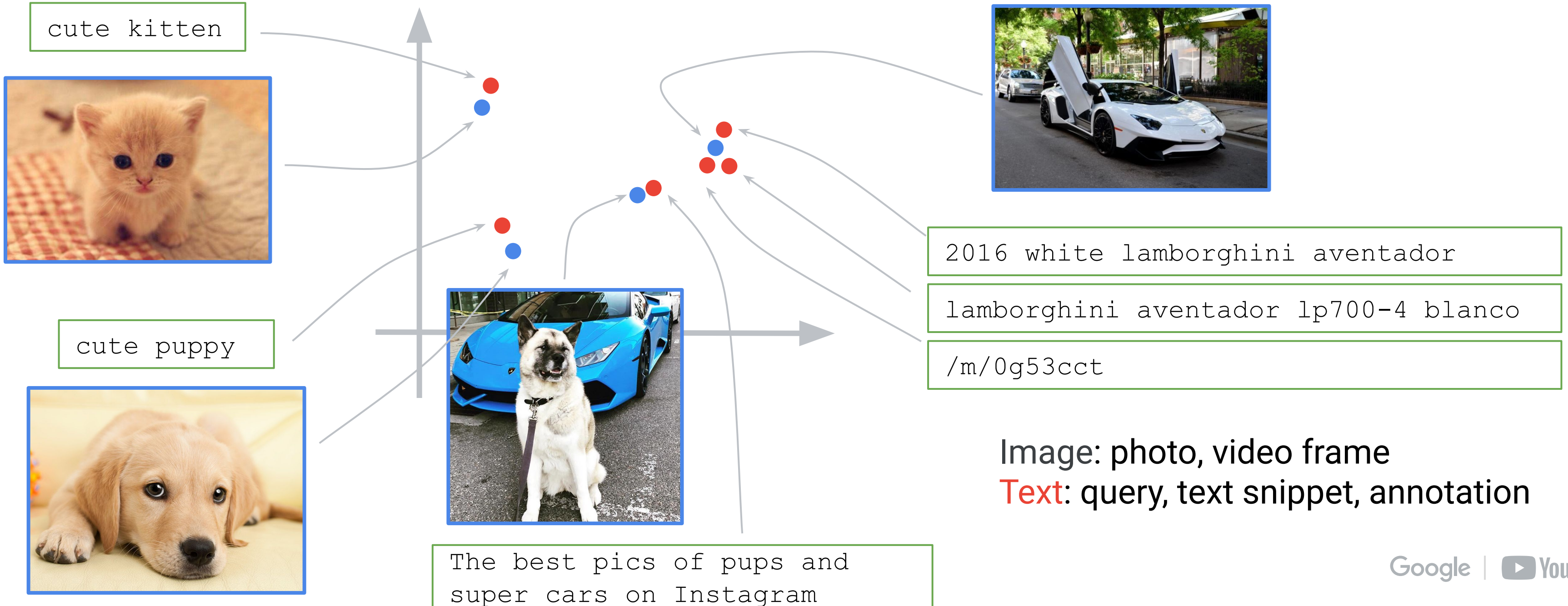


Image: photo, video frame
Text: query, text snippet, annotation

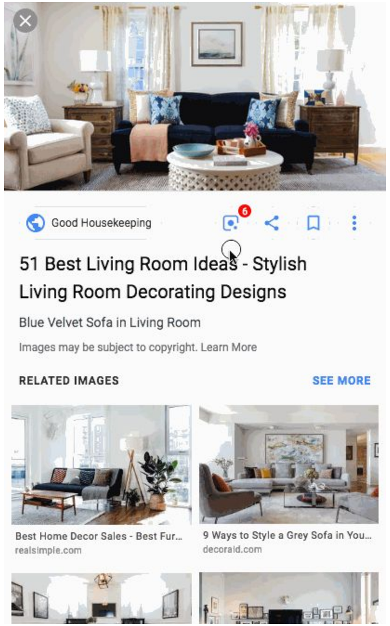
Specialized embeddings

(e.g., products)

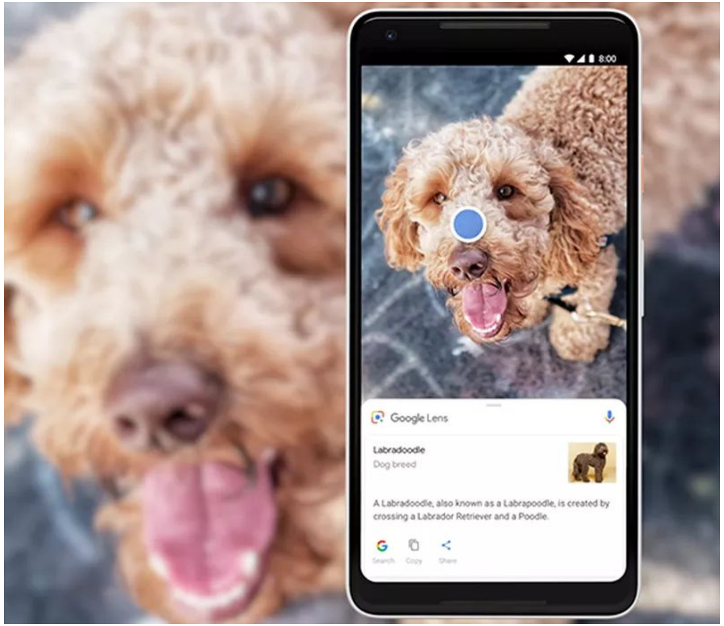
https://cloud.google.com/vision/product-search/docs/

[Juan, Lu, Li, et al.]

Use cases in Search and Lens



Lens in Google Images



Lens natural world, similar products, gleaming, similar images, among other improvements

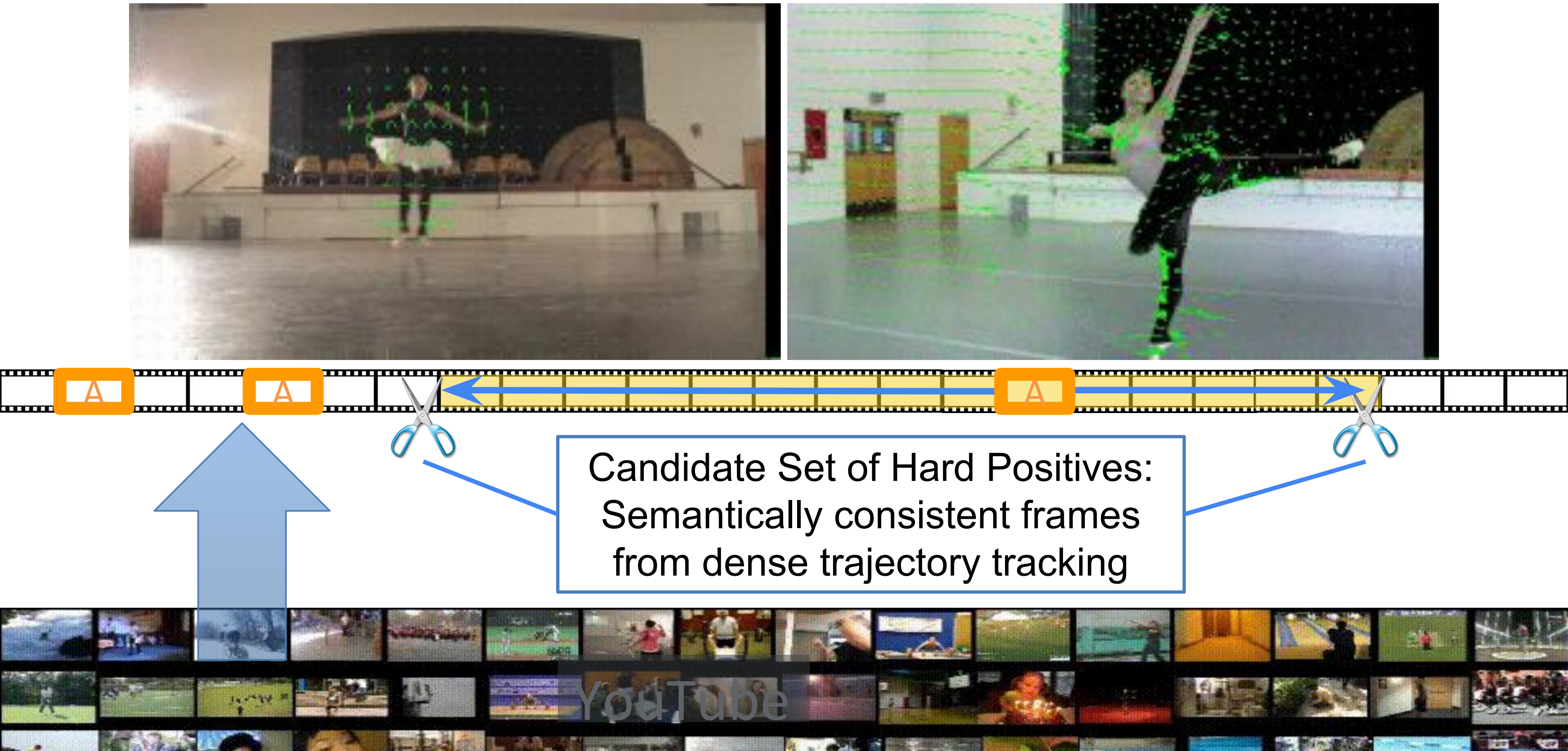
Using Video to help with Object Recognition in Images

Weak, noisy Labels from:

- meta-data, comments
- labeling using image-trained networks

[Hartmann, Grundmann, Hoffman, et al.]

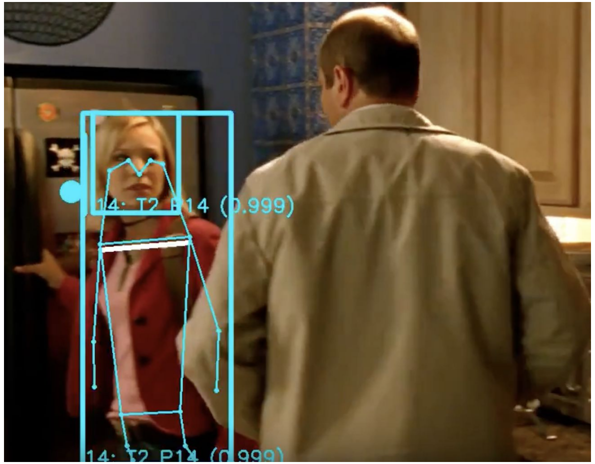
Weakly & Self-Supervised Learning from Video

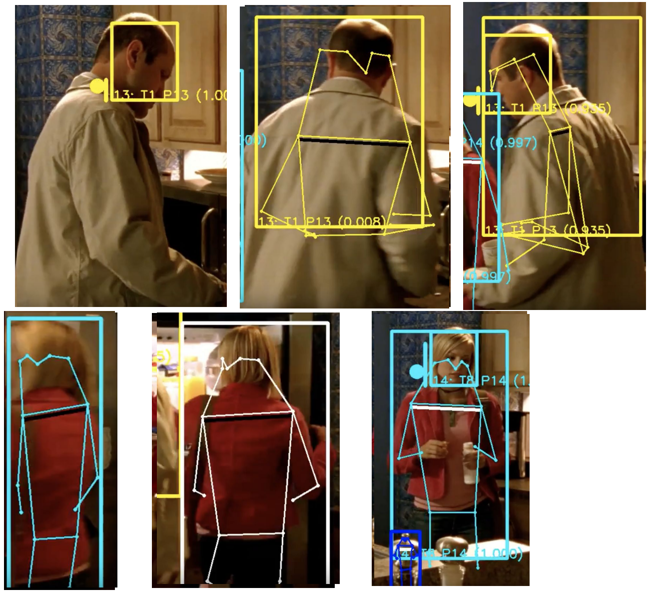


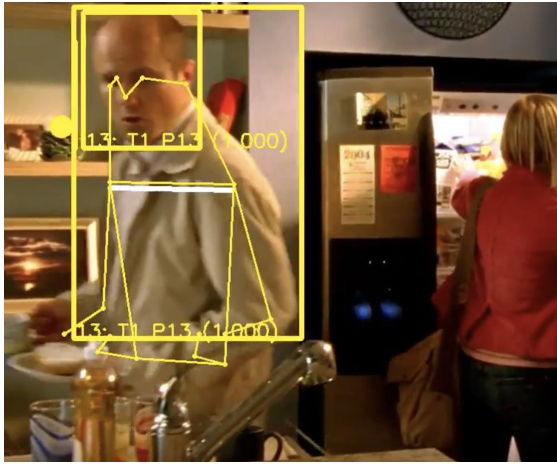
Using consistency signals for supervision

Goal: Pose-insensitive person embedding (i.e., PersonNet)

Solution: 360 degree pose samples from large image / video corpus + tracking + clustering + user feedback signals







video + tracking → 360° pose training samples

Understanding: Starting from...

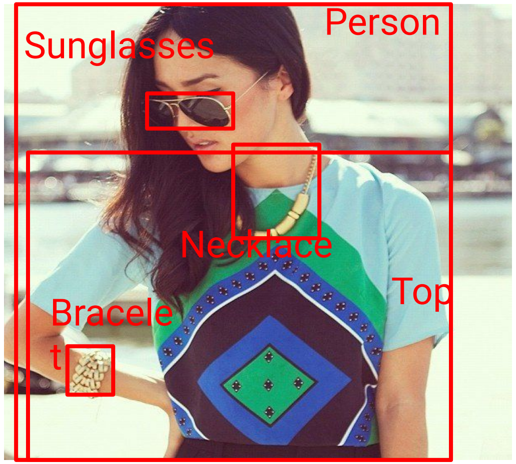
Classification

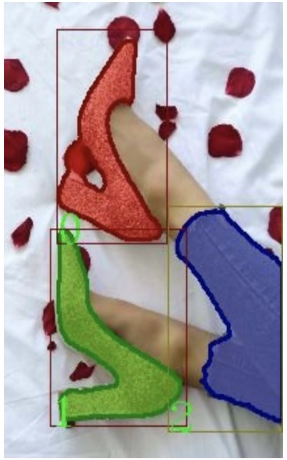
Images ⇒ Labels

Cake (0.93669)
Dessert (0.91911)
Birthday (0.89697)
Child (0.89183)
Fondant (0.88340)
Birthday cake (0.87525)

Detection

Images ⇒ Labelled Boxes or Regions





Embedding

Images ⇒ Features

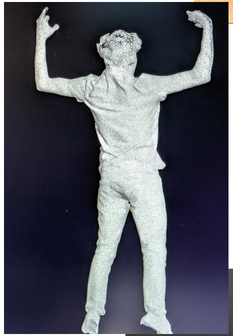
Similar Images and Labels

Understanding: ... adding in...

3D perception

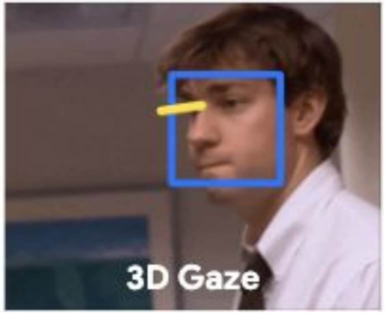
Images ⇒ 3D relationships

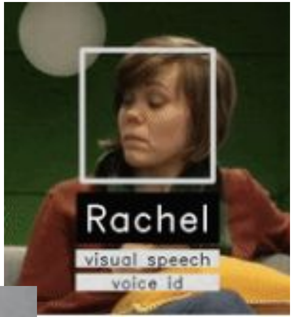
single image

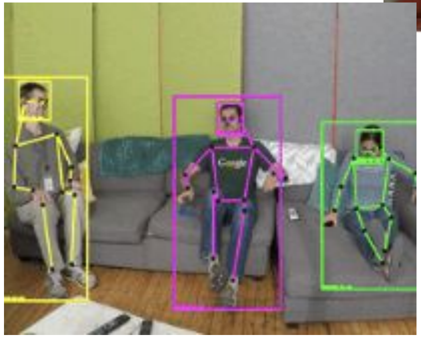


Person-centric Models

Focus, Pose, Speaker Models







Action/Interaction Recognition

shoot ball

kick ball

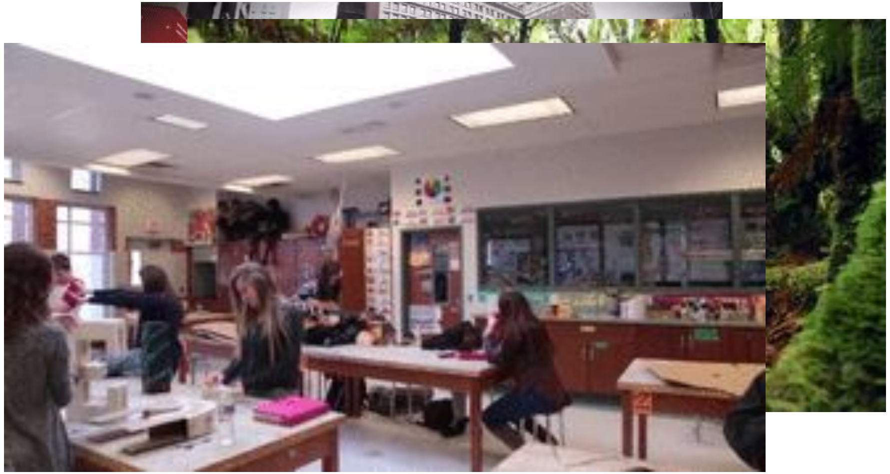
pour

[Zhou, Brown, Snavely, Lowe]

YouTube ⇒ 3D models of scenes & people

What 3D structure can we learn from watching internet video?

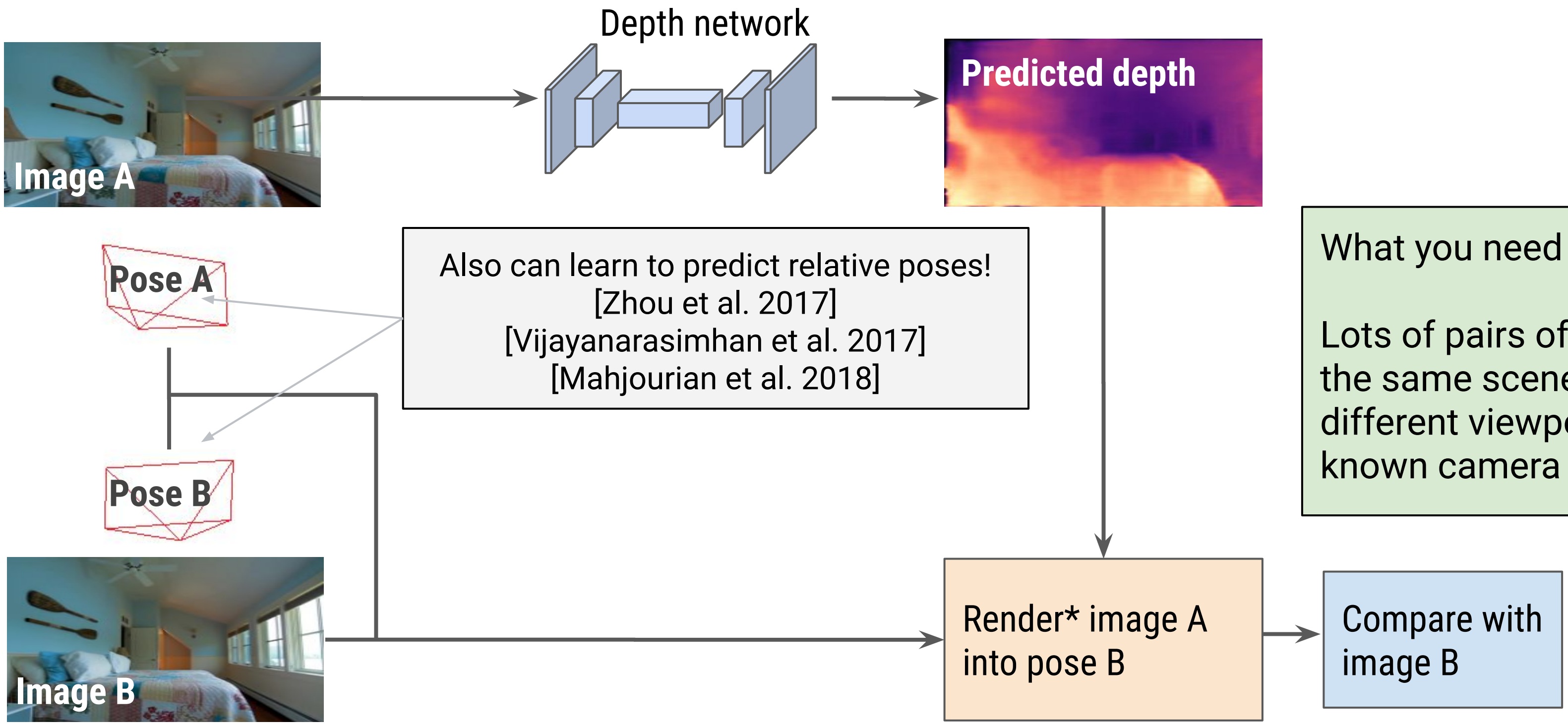
Training: Multiple views



Testing: Single Image

[Zhou, Brown, Snavely, Lowe]

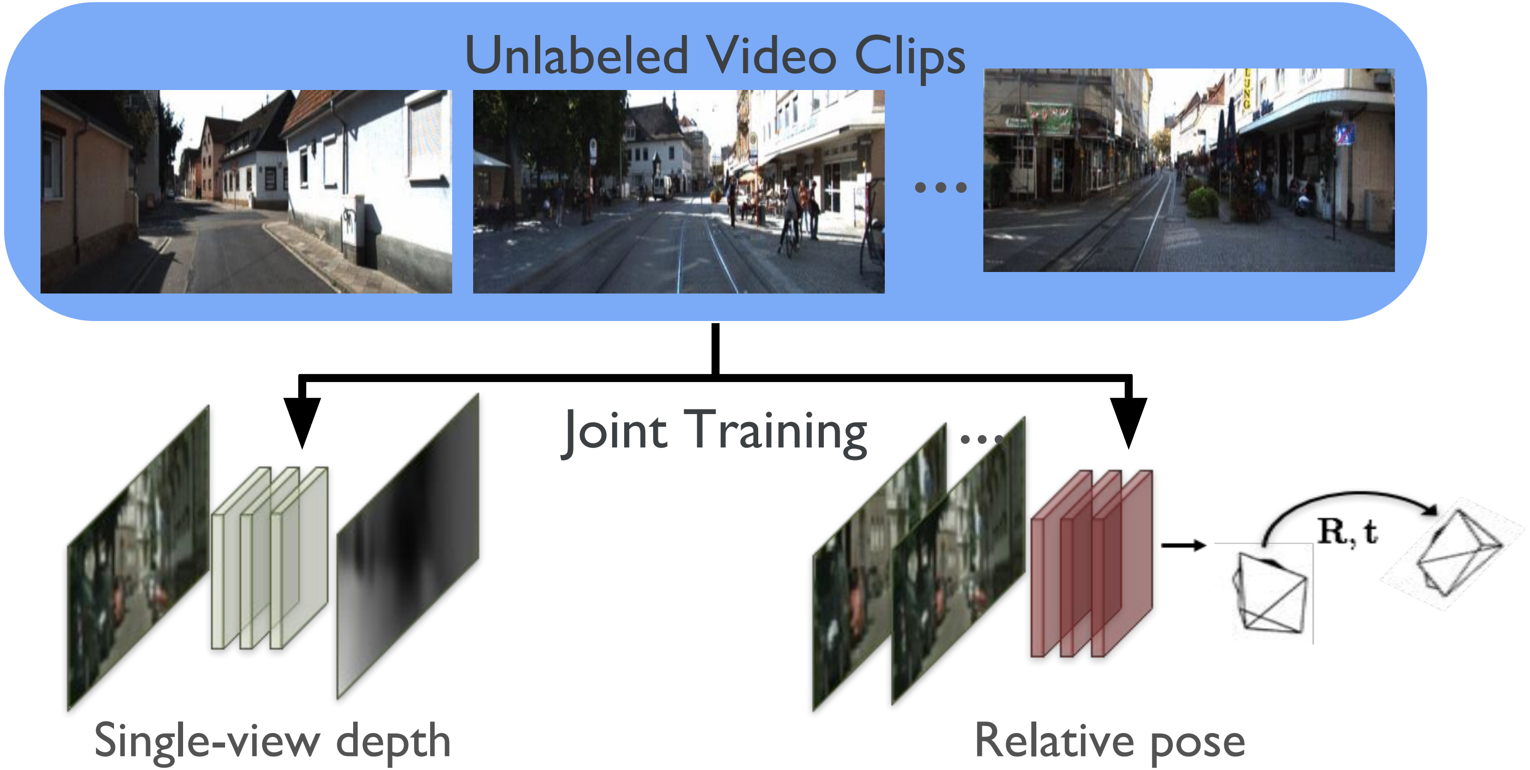
Beyond direct supervision



What you need for training:

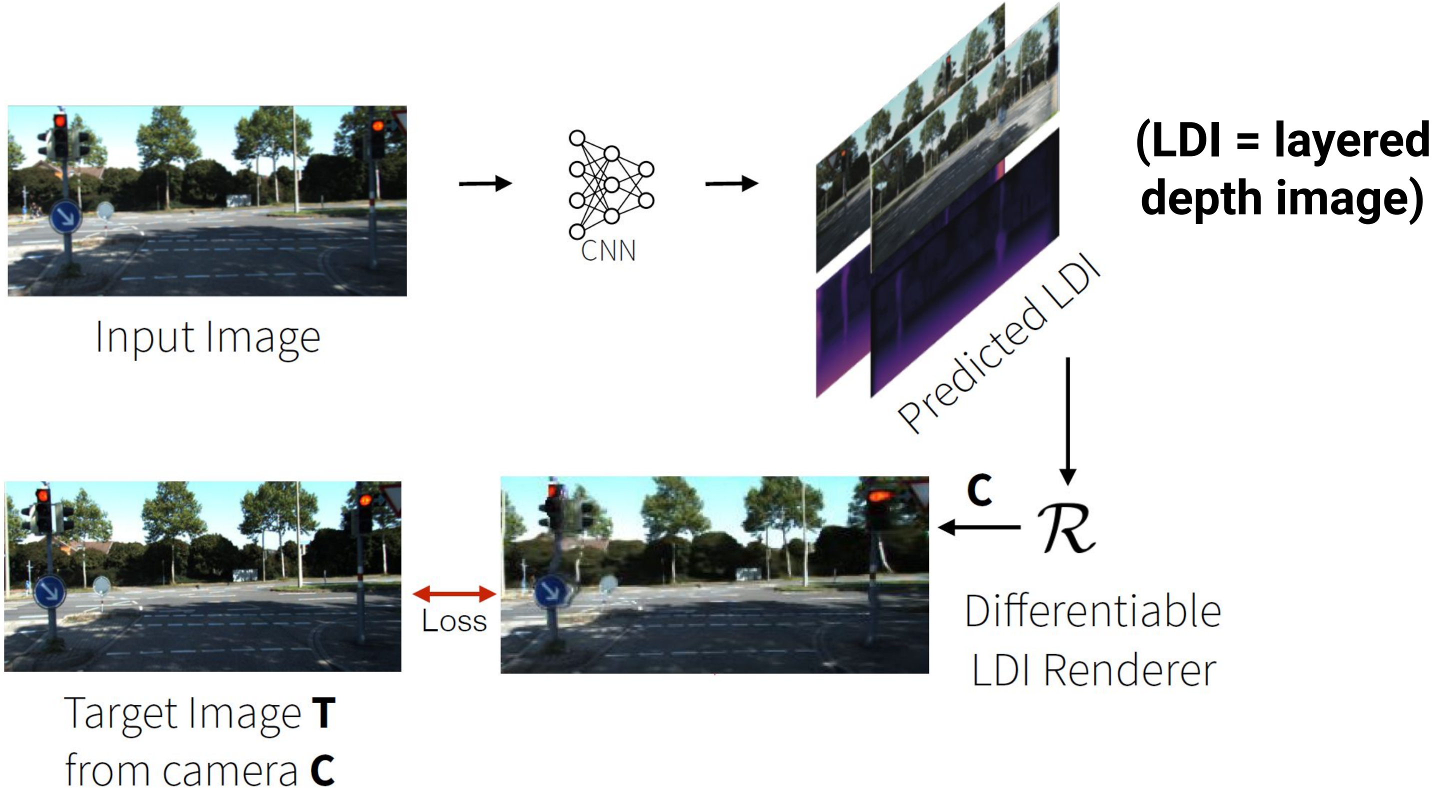
Lots of pairs of images of the same scene from different viewpoints, with known camera pose

Learning depth and camera pose via view synthesis



[Zhou, Brown, Snavely, Lowe. *Unsupervised Learning of Depth and Ego-Motion from Video*. CVPR 2017]

Beyond depth maps: Learning layered models



[Tulsiani, Tucker, Snavely. *Layer-structured 3D Scene Inference via View Synthesis*. ECCV 2018.]

Reference input view

Plane 0

Plane 9

Plane 13

Plane 16

Plane 24

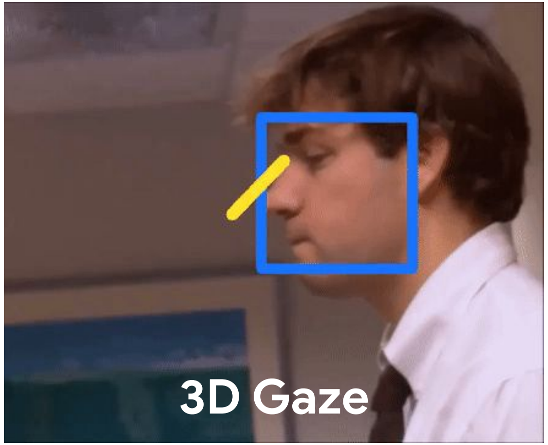
Plane 26

Output

Output

[Mori, Pantofaru, Kothari, et al.]

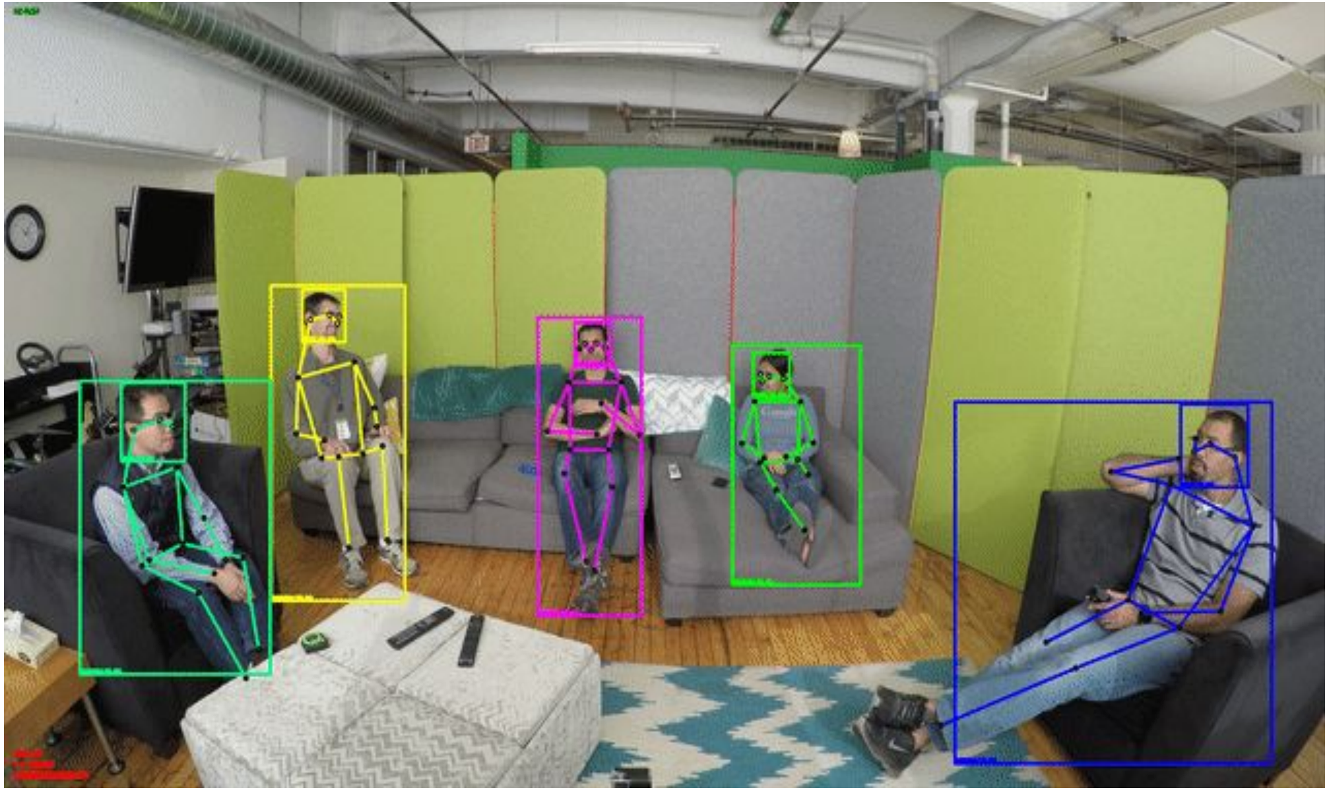
Focus on people



- Better binary and 3D gaze models
- Dynamic gestures
- Active speaker detection
- Speech Detection and Diarization

[Papandreou, Zhu, Kanazawa, et al.]

Focus on people



- Combined person & face SSD detection models

- Recurrent models for detection

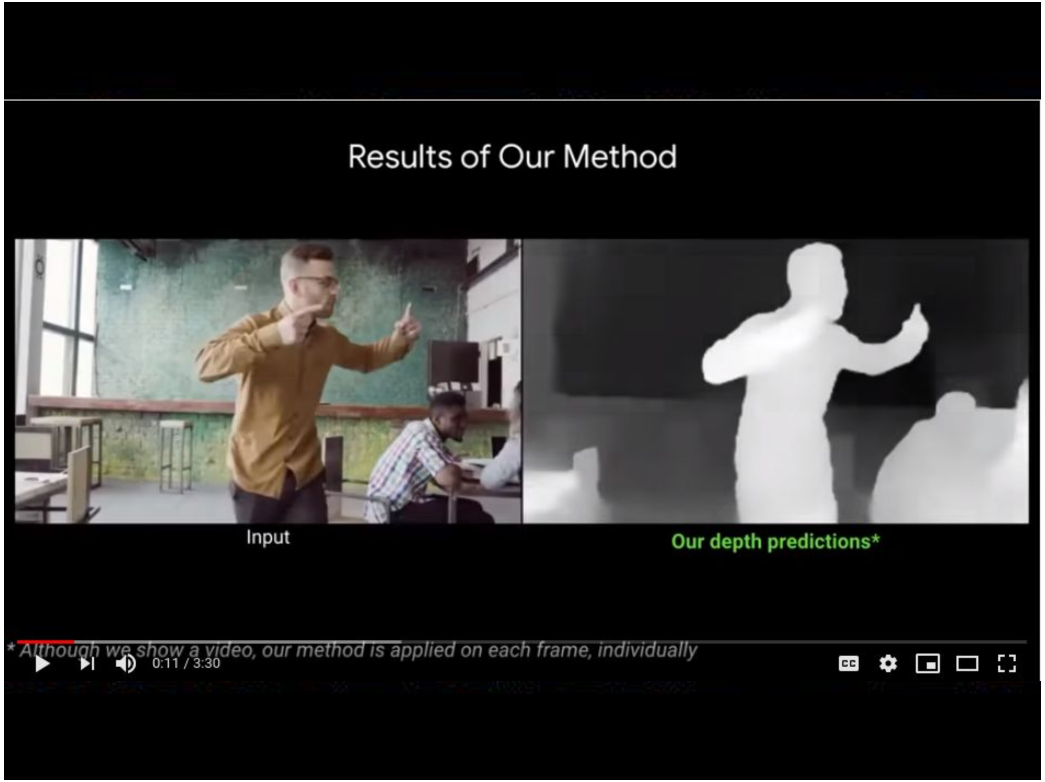
- Probabilistic tracking

- Rotation invariance

Depth Estimation for Moving People

Z. Li, T. Dekel, F. Cole, R. Tucker,
N. Snavely, C. Liu, W.T. Freeman,
CVPR 2019 Honorable Mention

Goal: depth estimation with moving camera and moving people



Depth Estimation for Moving People

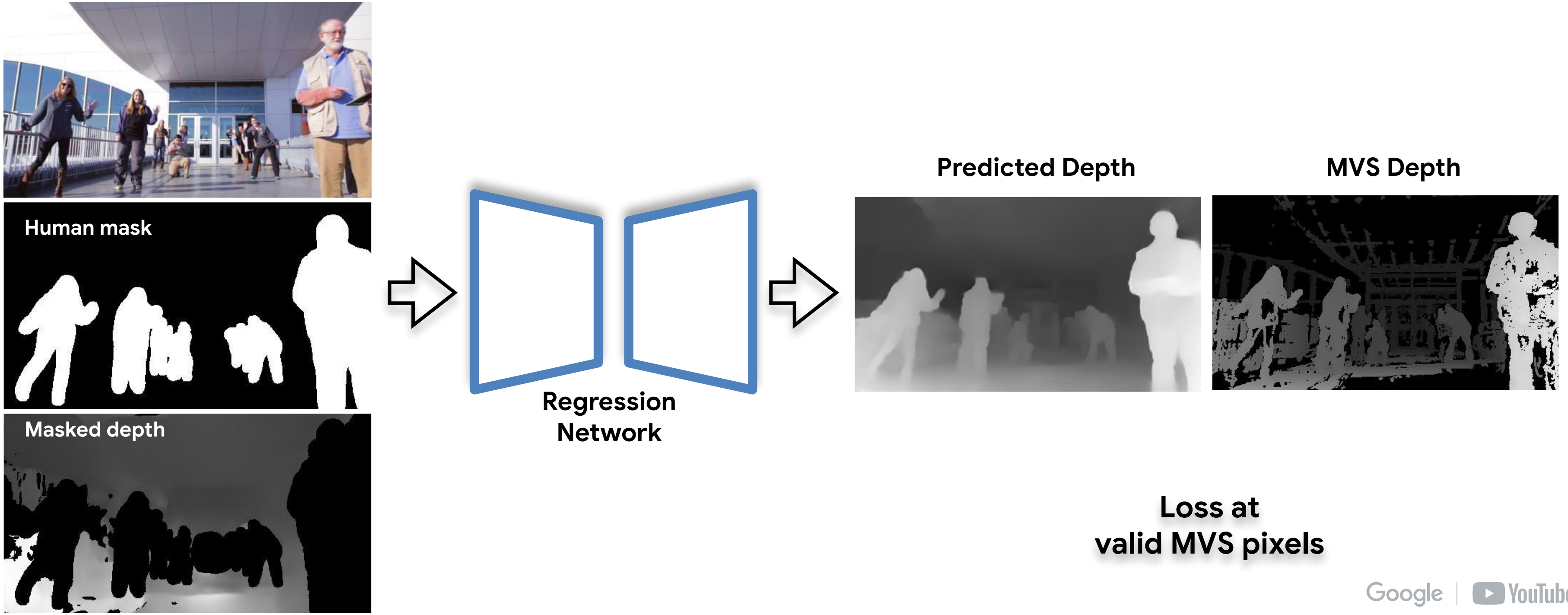
Z. Li, T. Dekel, F. Cole, R. Tucker,
N. Snavely, C. Liu, W.T. Freeman,
CVPR 2019 Honorable Mention

Idea: leverage MannequinChallenge dataset of ***frozen*** people!

Depth Estimation for Moving People

Z. Li, T. Dekel, F. Cole, R. Tucker,
N. Snavely, C. Liu, W.T. Freeman,
CVPR 2019 Honorable Mention

Approach: compute depth for static scene with multiview stereo, and predict depth for moving people with a regression network

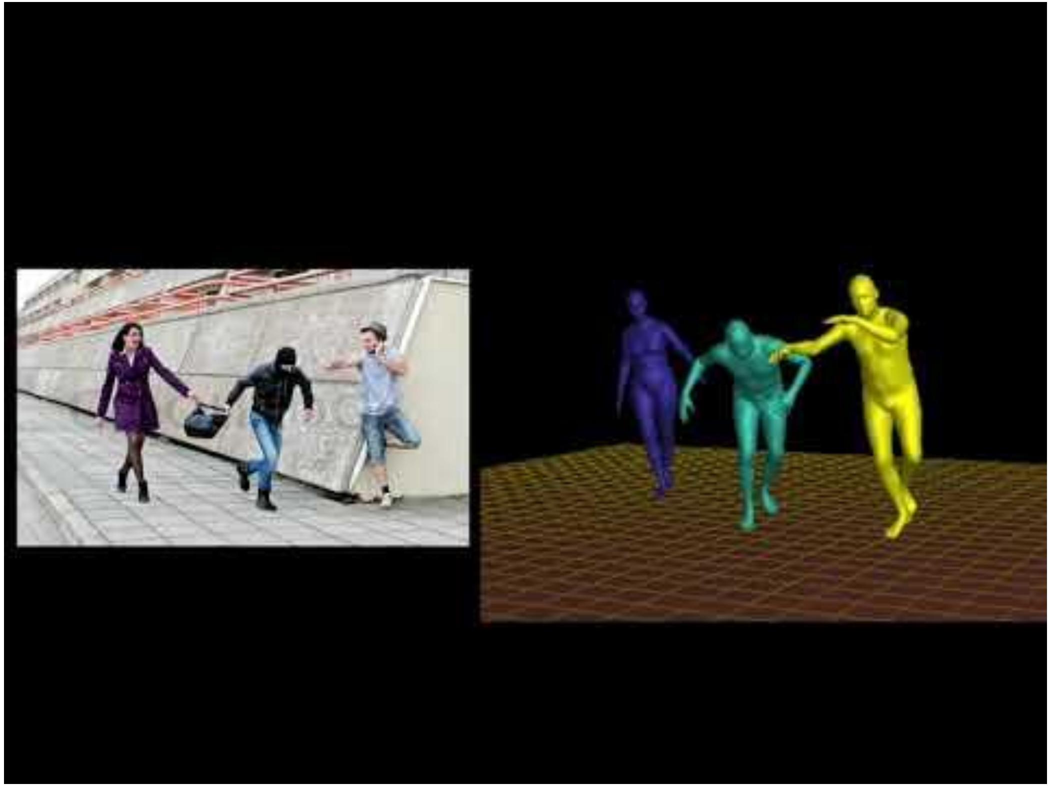


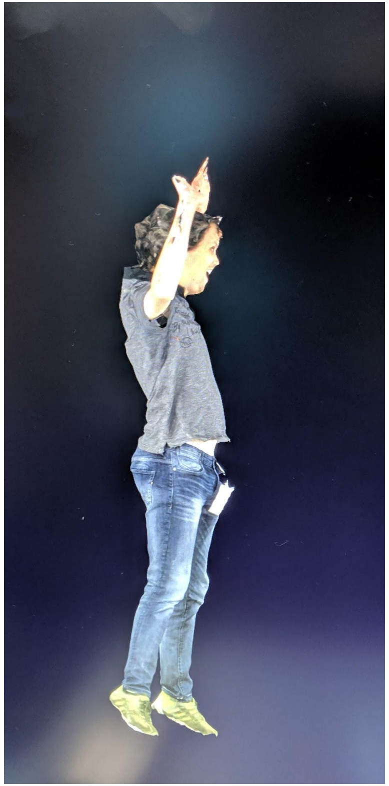
Z. Li, T. Dekel, F. Cole, R. Tucker,
N. Snavely, C. Liu, W.T. Freeman,
CVPR 2019 Honorable Mention

Depth Estimation for Moving People

Result: depth estimation with moving camera and ***moving*** people

Focus on People: 3D Shape Models

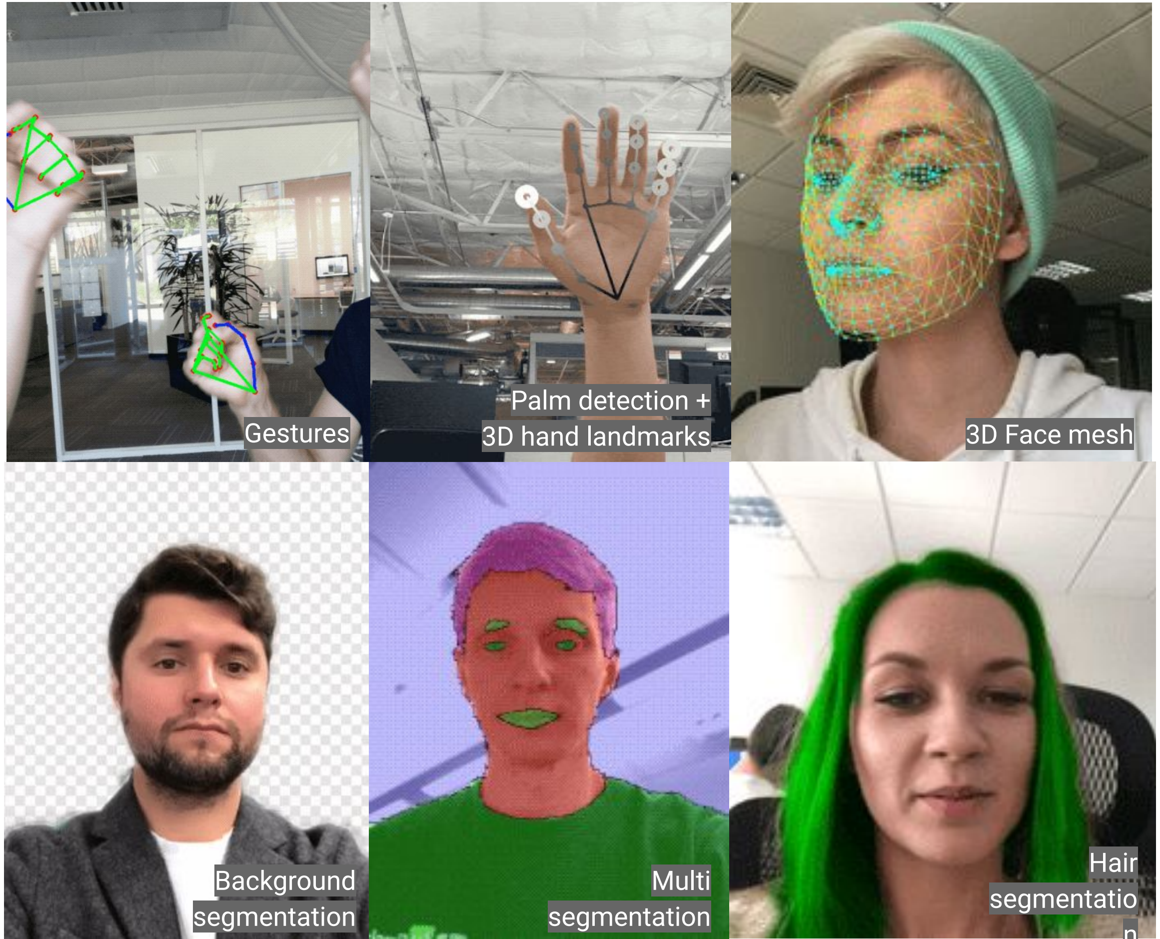




Physical relationships between people in 3D

[Wei, Ye, Mullen, et al.], [Kartynnik, Ablavatski, Grishchenko, Grundmann]

Focus on people: Real-time on Mobile



https://developers.google.com/ar/develop/java/augmented-faces/

Real time:

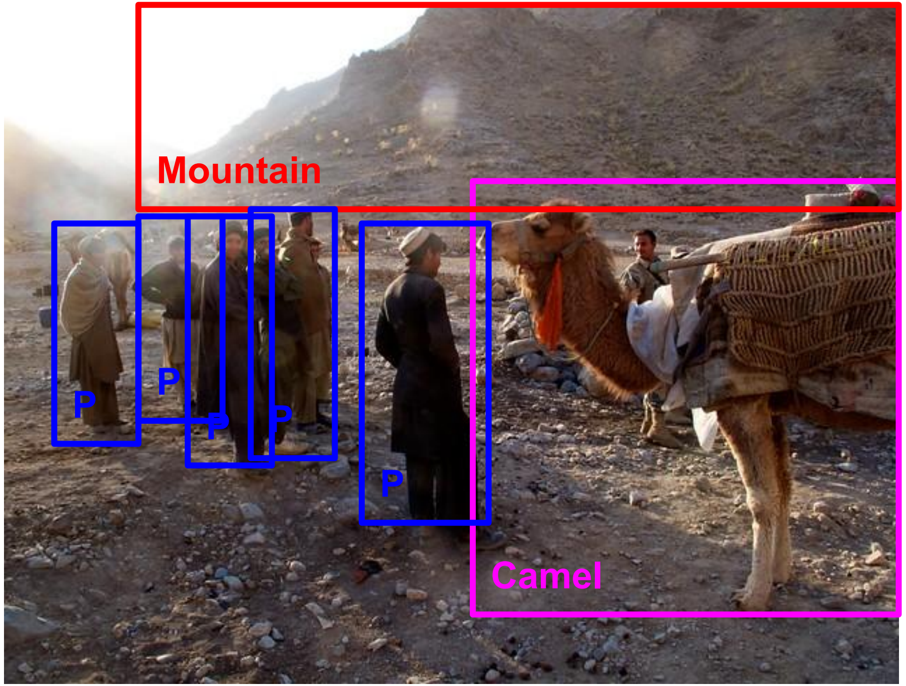
- Hand tracking

- Face tracking

- Expression parsing

Action Recognition

Object Recognition != Action Recognition



Examples of "Paint" action in AVA

[Sun, Ross, Vondrick, Pantofaru, et al.]

[Todd Huffman from Open Images v5]

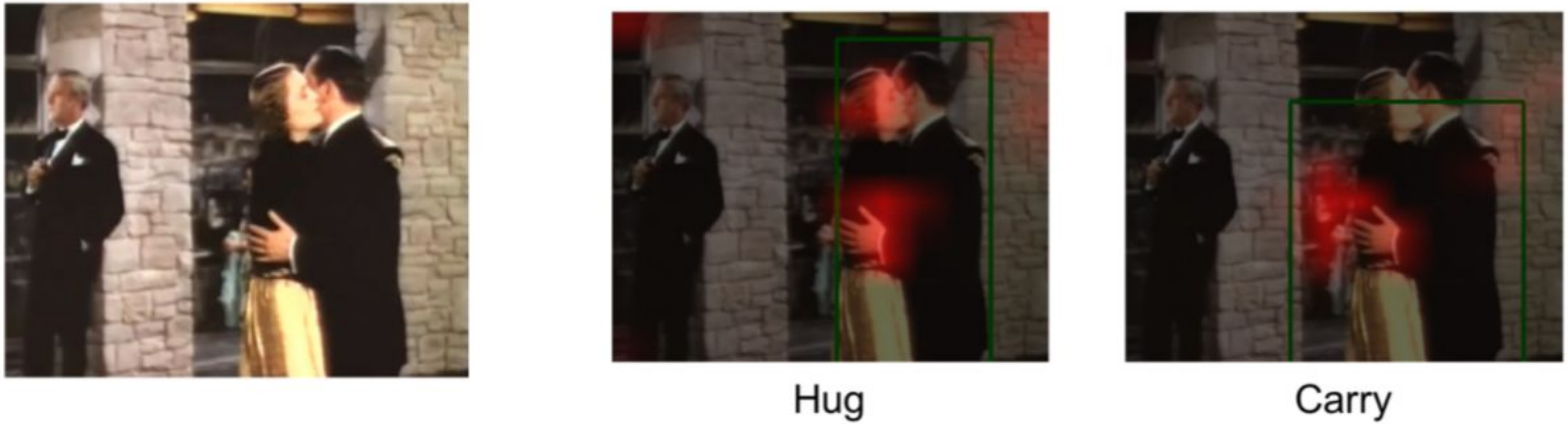
Actor-Centric Relation Network (ACRN)

- Faster RCNN looks only at the actors (appearance, pose, etc.)
- **Opportunity: model relationship between actor and other objects/people**



[Sun, Shrivastava, Vondrick, Schmid, Murphy, Sukthankar, ECCV'18]

Actor-Centric Relation Network (ACRN)



[Sun, Shrivastava, Vondrick, Schmid, Murphy, Sukthankar, ECCV'18]

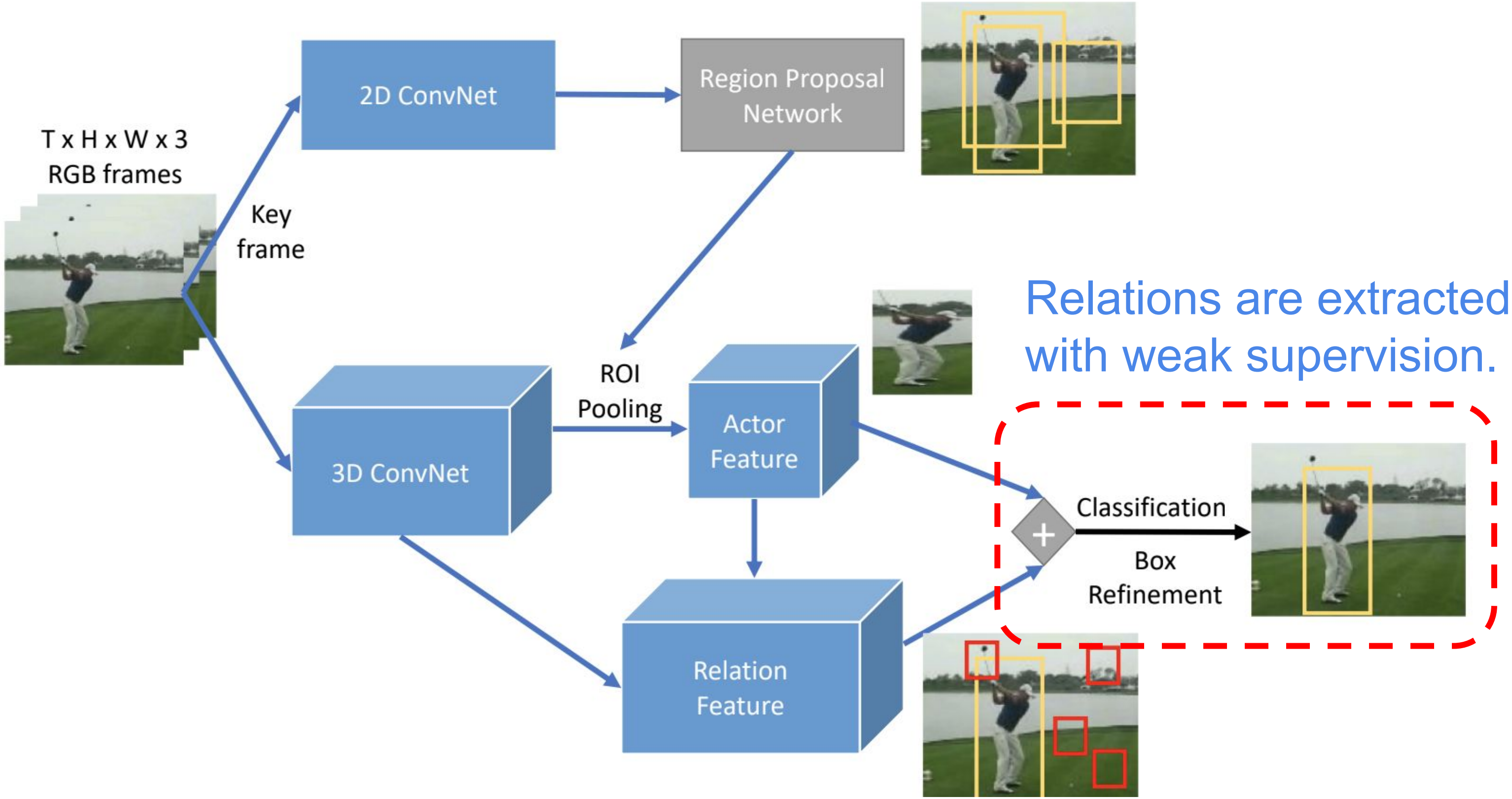
Actor-Centric Relation Network (ACRN)

Catch ball

Throw ball

[Sun, Shrivastava, Vondrick, Schmid, Murphy, Sukthankar, ECCV'18]

ACRN Architecture

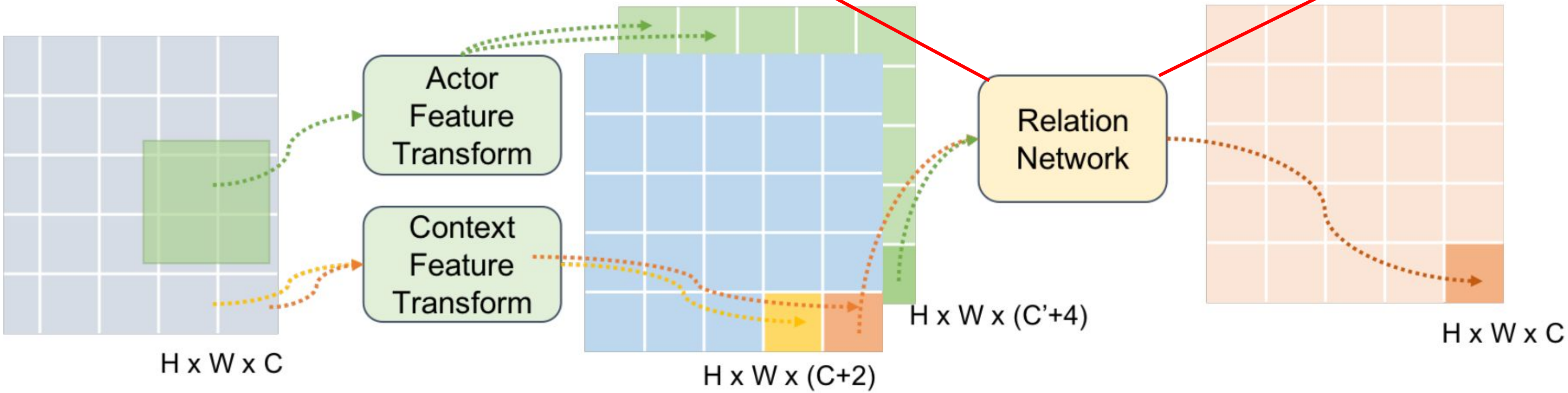


Relations are extracted with weak supervision.

[Sun, Shrivastava, Vondrick, Schmid, Murphy, Sukthankar, ECCV'18]

ACRN Architecture

$$\text{ACRN}(\mathcal{A}_i, I) = f_\phi \left(\sum_{j,k} g_\theta \left(a_i, o_{j,k} \right) \right)$$



- Pairwise relation between actor and "objects"
- No explicit objectness proposals, use feature cells
- Implemented as 1x1 convolutions

Related work: Santoro et al., A simple neural network module for relational reasoning. NeuIPS 2017.

[Sun, Shrivastava, Vondrick, Schmid, Murphy, Sukthankar, ECCV'18]

Visualizations

shoot ball

kick ball

pour

[Sun, Shrivastava, Vondrick, Schmid, Murphy, Sukthankar, ECCV'18]

Visualizations

Google | YouTube

Understanding: Starting from...

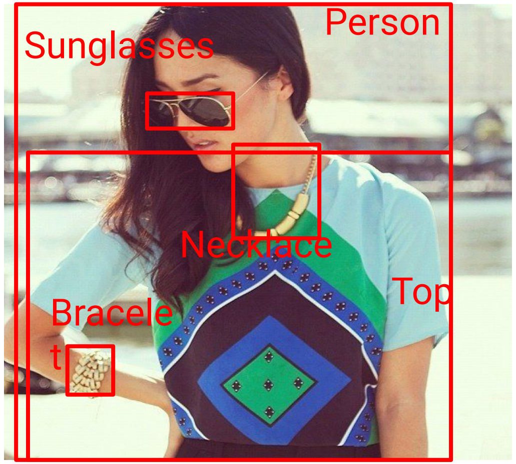
Classification

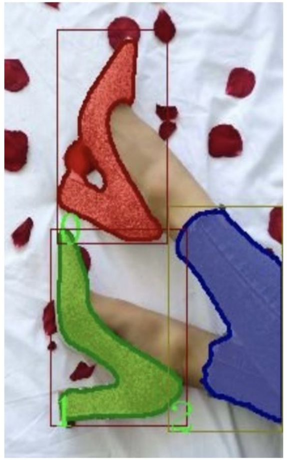
Images ⇒ Labels

Cake (0.93669)
Dessert (0.91911)
Birthday (0.89697)
Child (0.89183)
Fondant (0.88340)
Birthday cake (0.87525)

Detection

Images ⇒ Labelled Boxes or Regions





Embedding

Images ⇒ Features

Similar Images and Labels

Google | YouTube

Understanding: ... adding in...

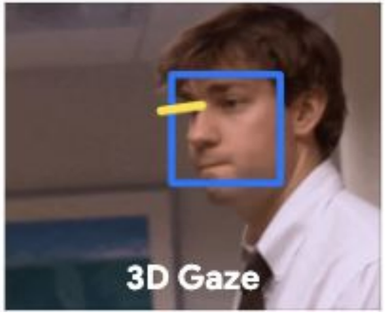
3D perception

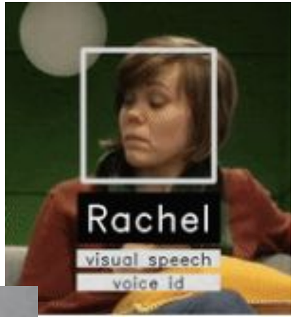
Images ⇒ 3D relationships

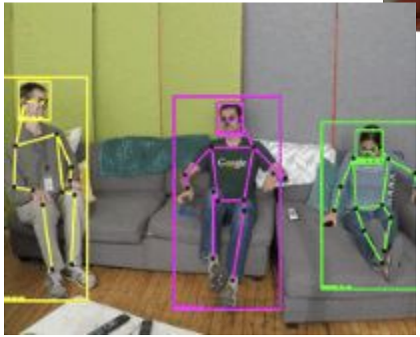
single image

Person-centric Models

Focus, Pose, Speaker Models







Action/Interaction Recognition

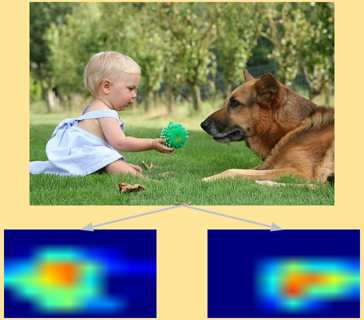
shoot ball

kick ball

pour

Understanding: ... but more to go?

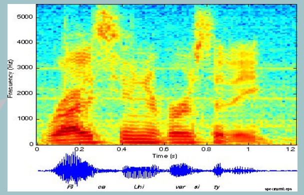
Visual (Scene) Features



Person Dog

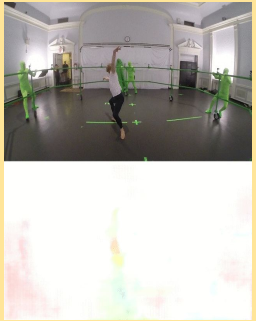
Visual (OCR Text) Features

Audio and Speech Features

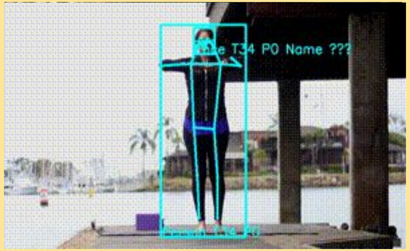


"Green Revolution"
"agricultural yields"
"irrigation techniques"

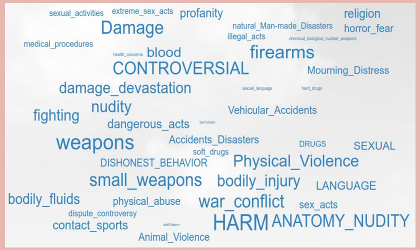
Visual (Motion) Features



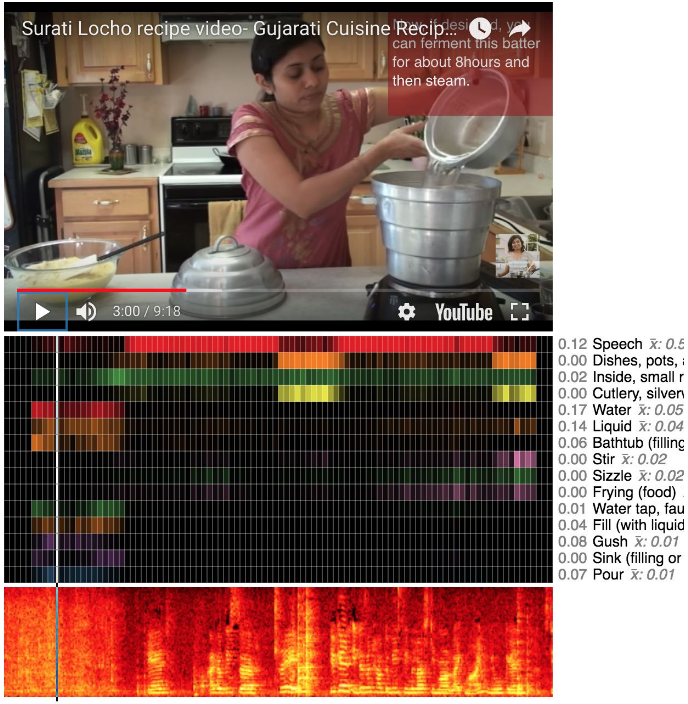
Visual (Person) Features



(Multi-modal) Semantic and Content Safety Features



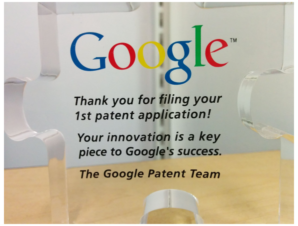
Much of video labeling/segmentation



different than understanding?

not focused on intent (or significance)

Text understanding: more focused on intent



Comprehensive **OCR** for lots of languages

Lens Dining

Lens Tip Calculator

[Vaswani, Shazeer, Parmar, et al.]

Machine Translation:
All about both focus and context

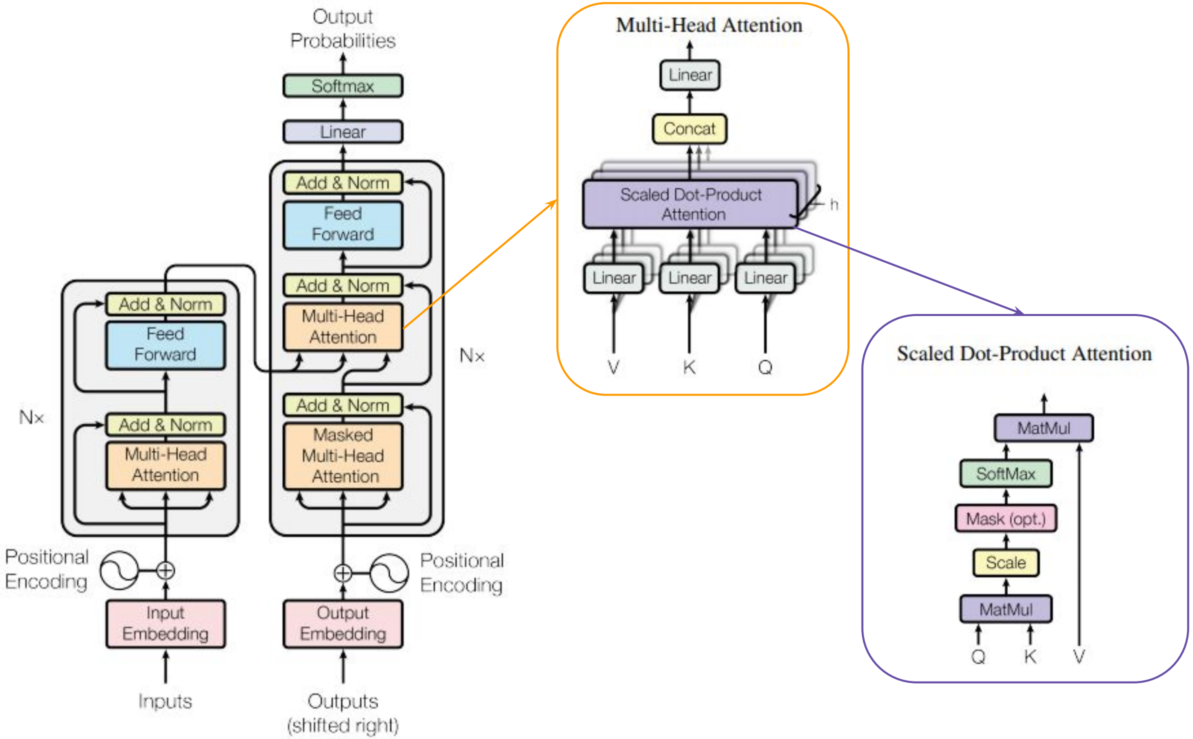
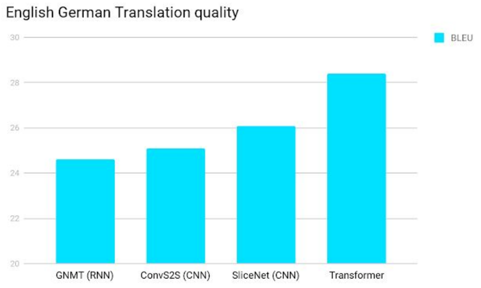
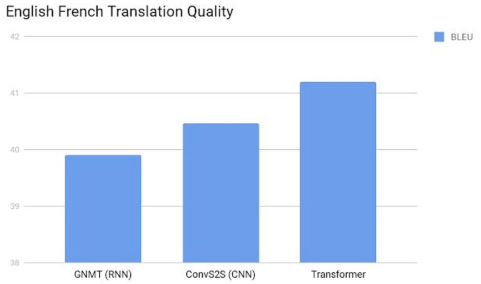


Figure 1: The Transformer - model architecture.



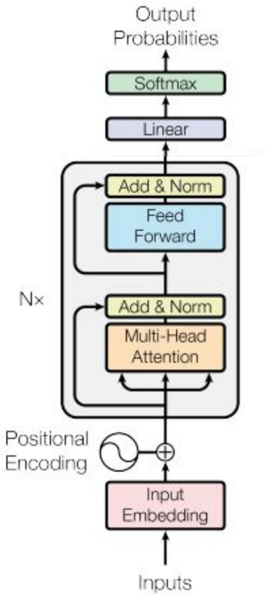
BLEU scores (higher is better) of single models on the standard WMT newstest2014 English to German translation benchmark.



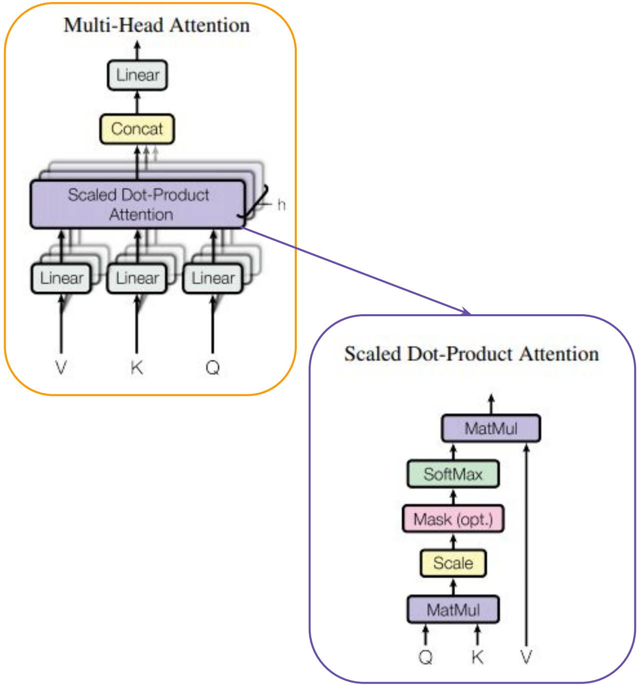
BLEU scores (higher is better) of single models on the standard WMT newstest2014 English to French translation benchmark.

Machine Translation:
All about both focus and context

[Devlin, Chang, Lee, Toutanova]



BERT



Results with BERT

To evaluate performance, we compared BERT to other state-of-the-art NLP systems. Importantly, BERT achieved all of its results with almost no task-specific changes to the neural network architecture. On SQuAD v1.1, BERT achieves 93.2% F1 score (a measure of accuracy), surpassing the previous state-of-the-art score of 91.6% and human-level score of 91.2%:

SQuAD1.1 Leaderboard

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	**87.433**	**93.160**
2 Sep 09, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.356	91.202
3 Jul 11, 2018	QANet (ensemble) *Google Brain & CMU*	84.454	90.490

BERT also improves the state-of-the-art by 7.6% absolute on the very challenging GLUE benchmark, a set of 9 diverse Natural Language Understanding (NLU) tasks. The amount of human-labeled training data in these tasks ranges from 2,500 examples to 400,000 examples, and BERT substantially improves upon the state-of-the-art accuracy on all of them:

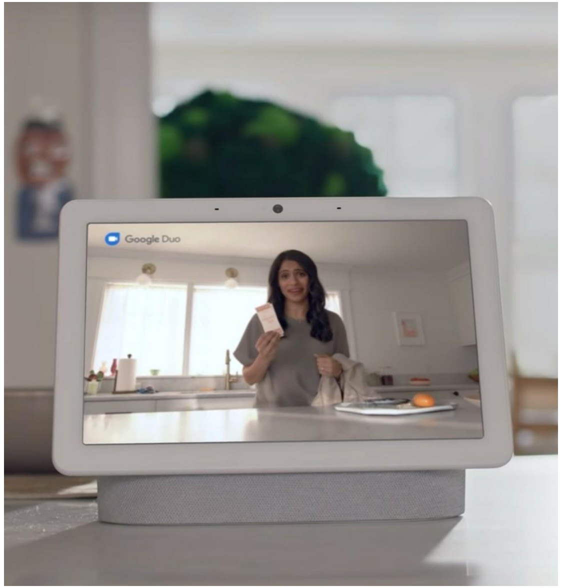
Rank	Model	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	QNLI	RTE
1	BERT: 24-layers, 1024-hidden, 16-heads	80.4	60.5	94.9	85.4/89.3	87.6/86.5	89.3/72.1	86.7	91.1	70.1
2	Singletask Pretrain Transformer	72.8	45.4	91.3	75.7/82.3	82.0/80.0	88.5/70.3	82.1	88.1	56.0
3	BiLSTM+ELMo+Attn	70.5	36.0	90.4	77.9/84.9	75.1/73.3	84.7/64.8	76.4	79.9	56.8

Understanding: ... need both focus and context

Significance in situated video

use context

Interaction in live video



Intent in authored video

[Beery, Wu, Rathod, et al.]

Static passive monitoring cameras

- Sparse, irregular frame rate
- Power, computational, and memory constraints.
- Many images are empty
- Always looking at the same background, objects of interest often habitual

[Beery, Wu, Rathod, et al.]

Data Challenges

(1) Illumination

(2) Blur

(3) ROI Size

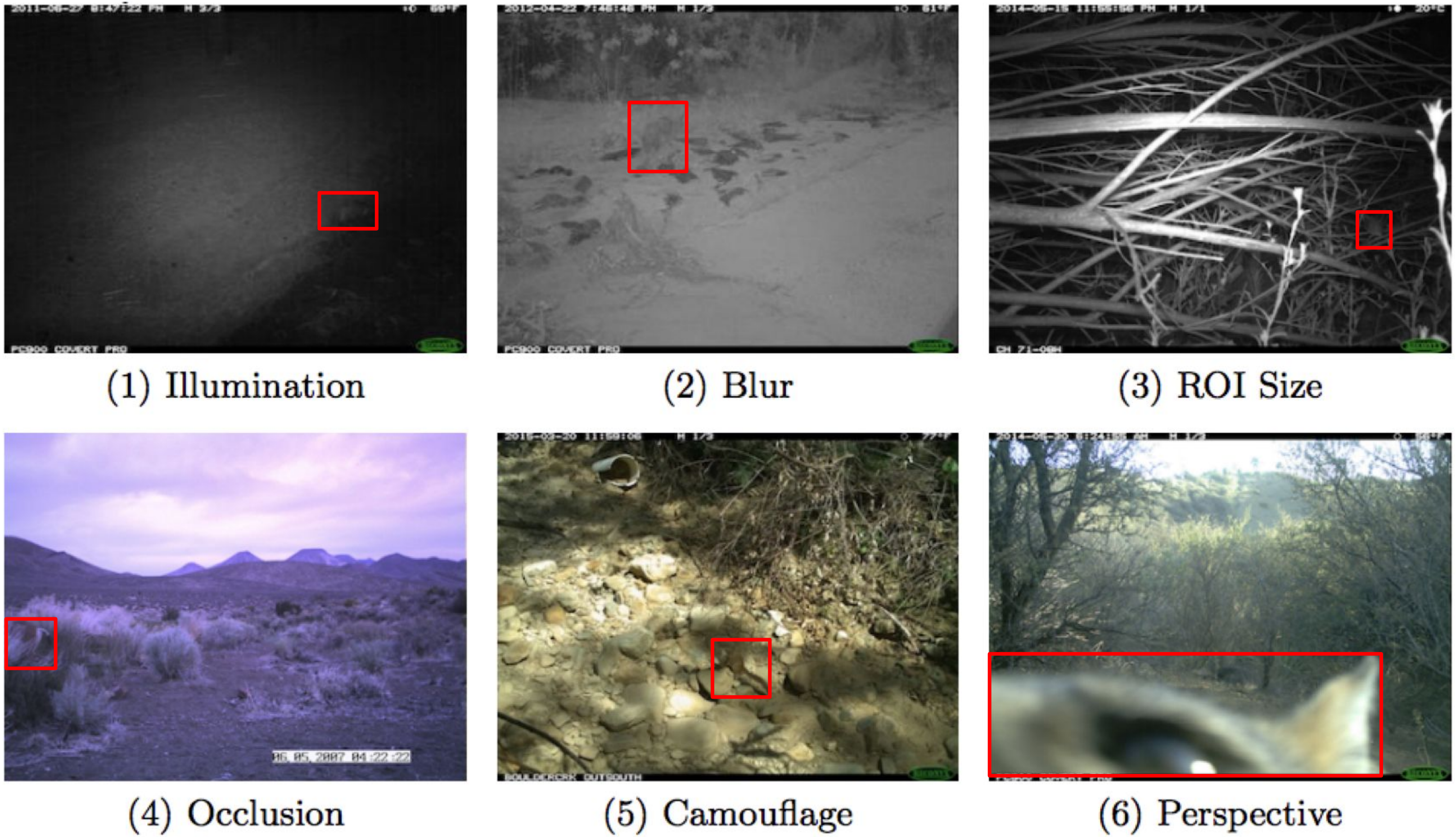
(4) Occlusion

(5) Camouflage

(6) Perspective

All these images have an animal in them

[Beery, Wu, Rathod, et al.]



[Beery, Wu, Rathod, et al.]

Cameras are static; Objects are habitual!

We want per-camera models that leverage long-term temporal context to:

[Beery, Wu, Rathod, et al.]

Cameras are static; Objects are habitual!

We want per-camera models that leverage long-term temporal context to:

1. Ignore salient false positives

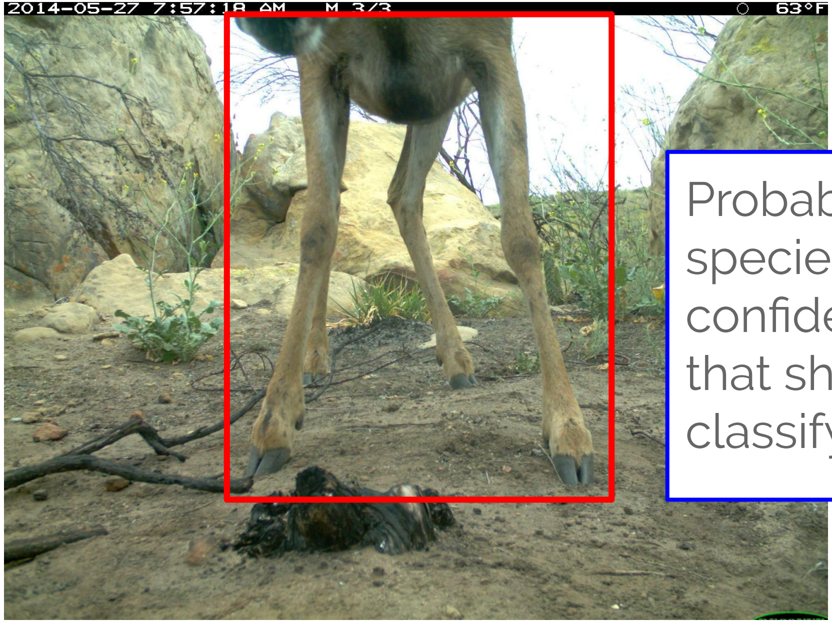
These rocks have not moved in a month. *Probably not animals.*

[Beery, Wu, Rathod, et al.]

Cameras are static; Objects are habitual!

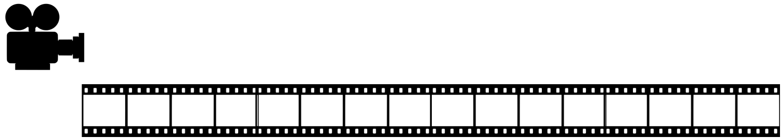
We want per-camera models that leverage long-term temporal context to:

1. Ignore salient false positives
2. Improve per-location object classification



Probably the same species; If we're confident about one, that should help us classify the other

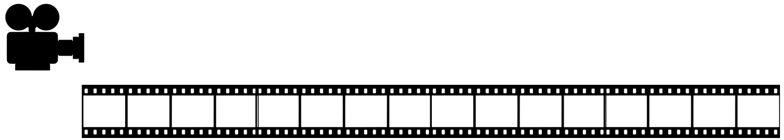
Our approach (high level)



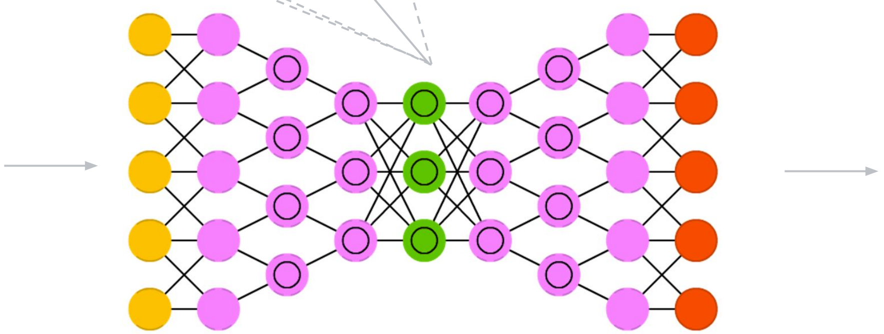
1. Build (nonparametric) per-camera model representing what a camera has seen (a.k.a. *memory bank*)

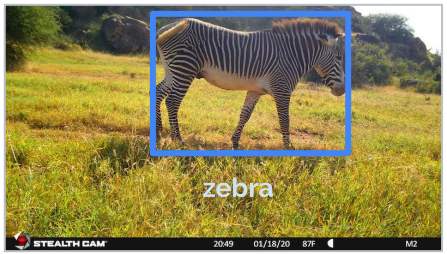
[Beery, Wu, Rathod, et al.]

Our approach (high level)



1. Build (nonparametric) per-camera model representing what a camera has seen (a.k.a. *memory bank*)

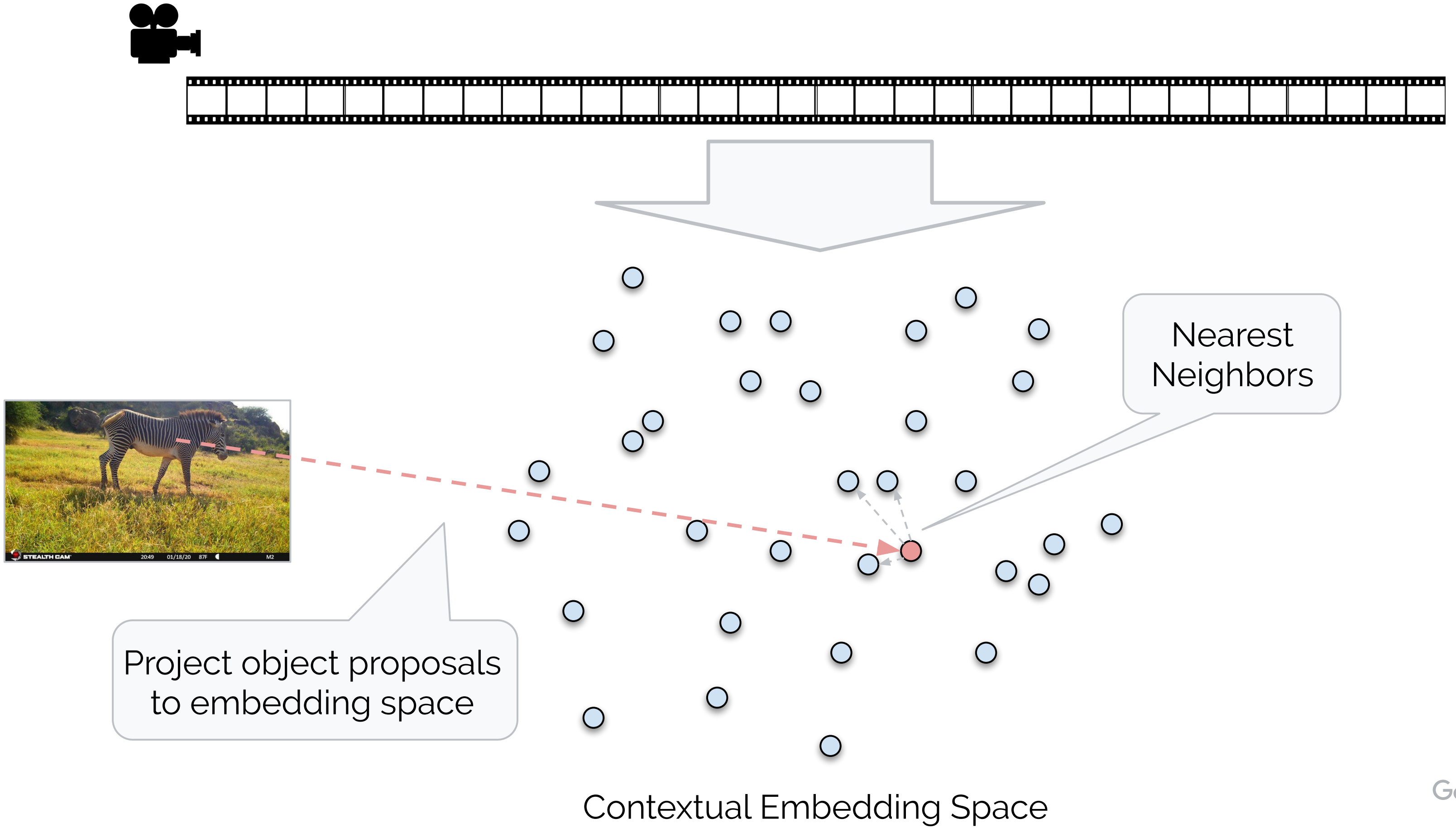




2. Give model running on current frame a way to reference into the memory bank

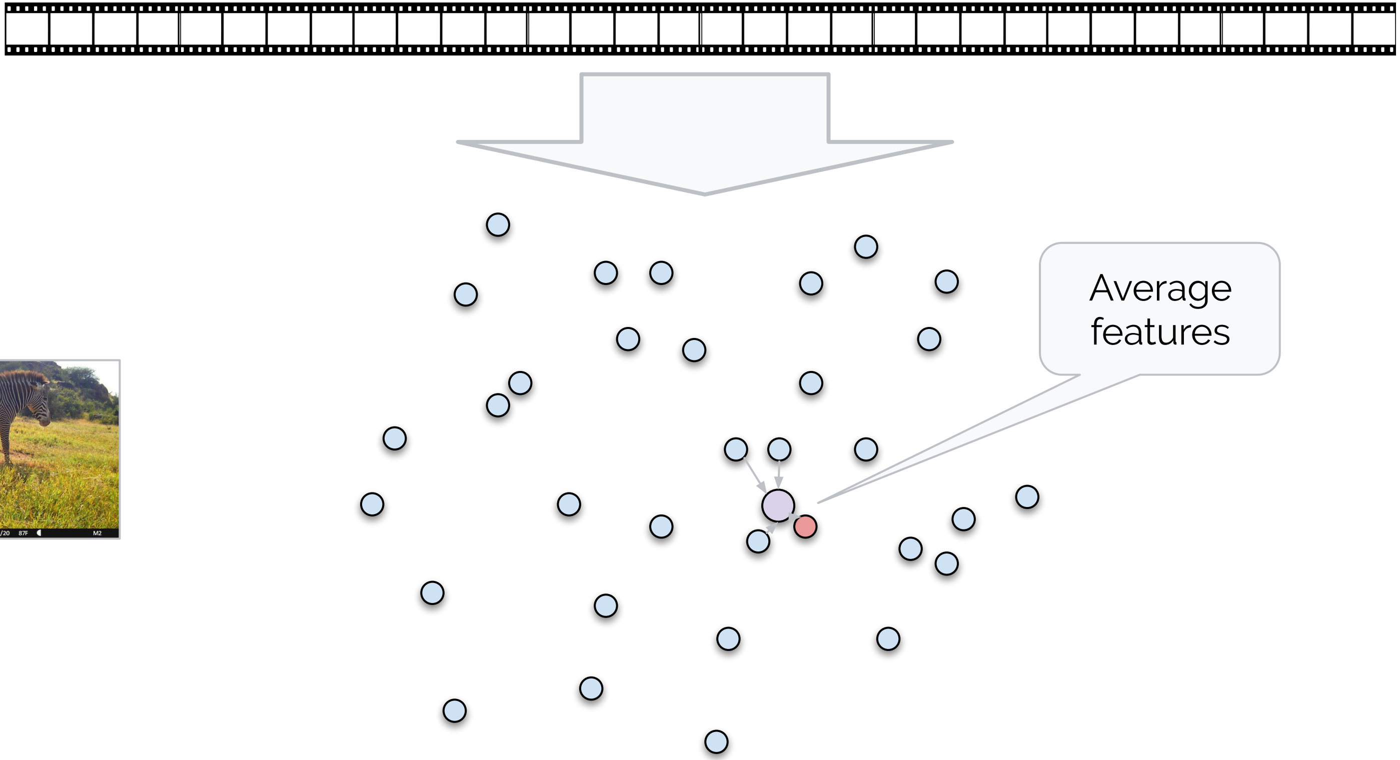
[Beery, Wu, Rathod, et al.]

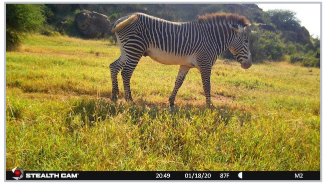
Aggregating Features from Memory Bank: Simplified



[Beery, Wu, Rathod, et al.]

Aggregating Features from Memory Bank: Simplified

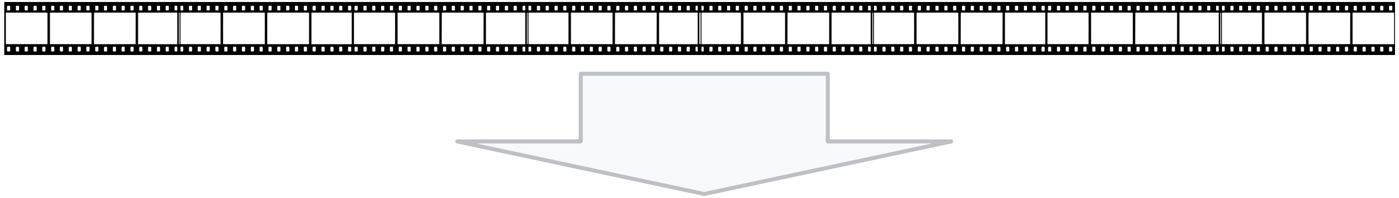


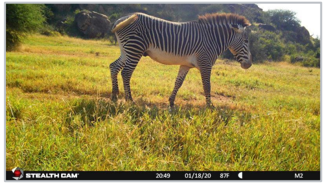


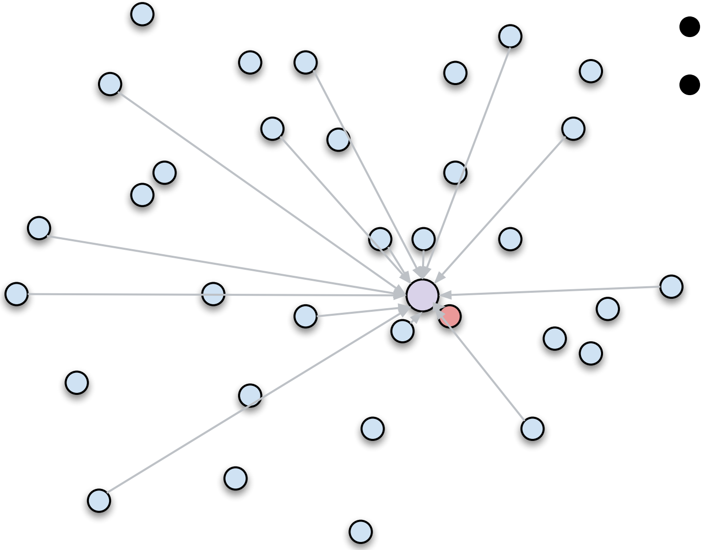
Contextual Embedding Space

[Beery, Wu, Rathod, et al.]

Aggregating Features via Attention





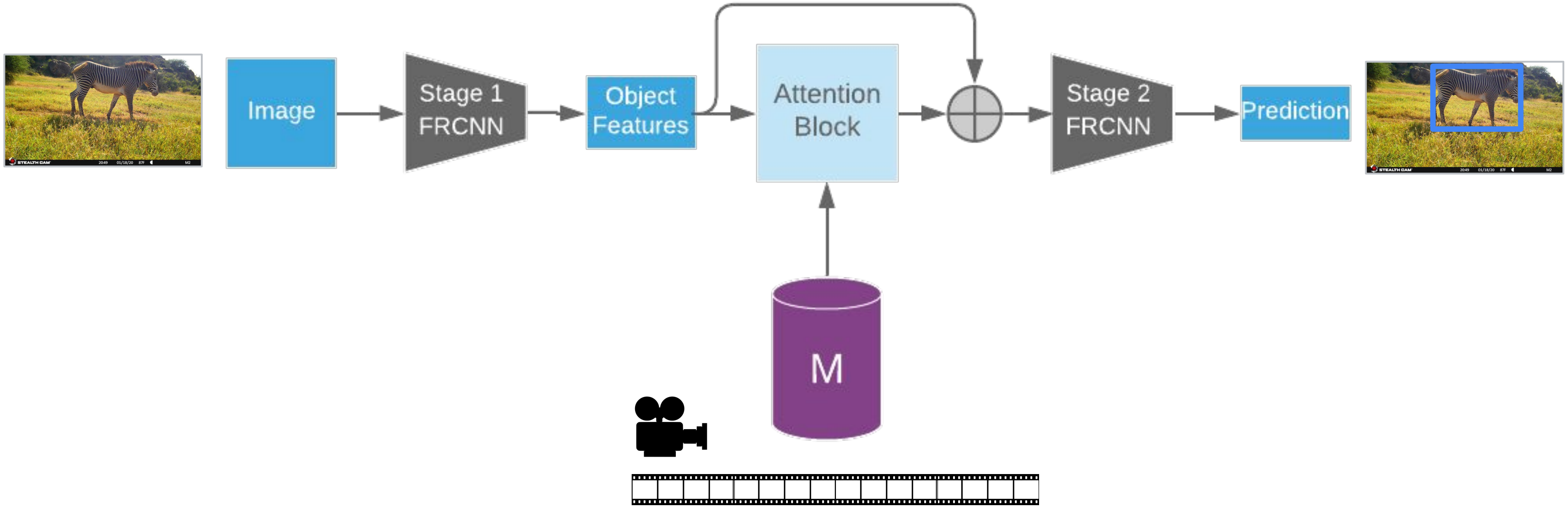


- Learn similarity metric
- Weighted average over entire memory bank

Contextual Embedding Space

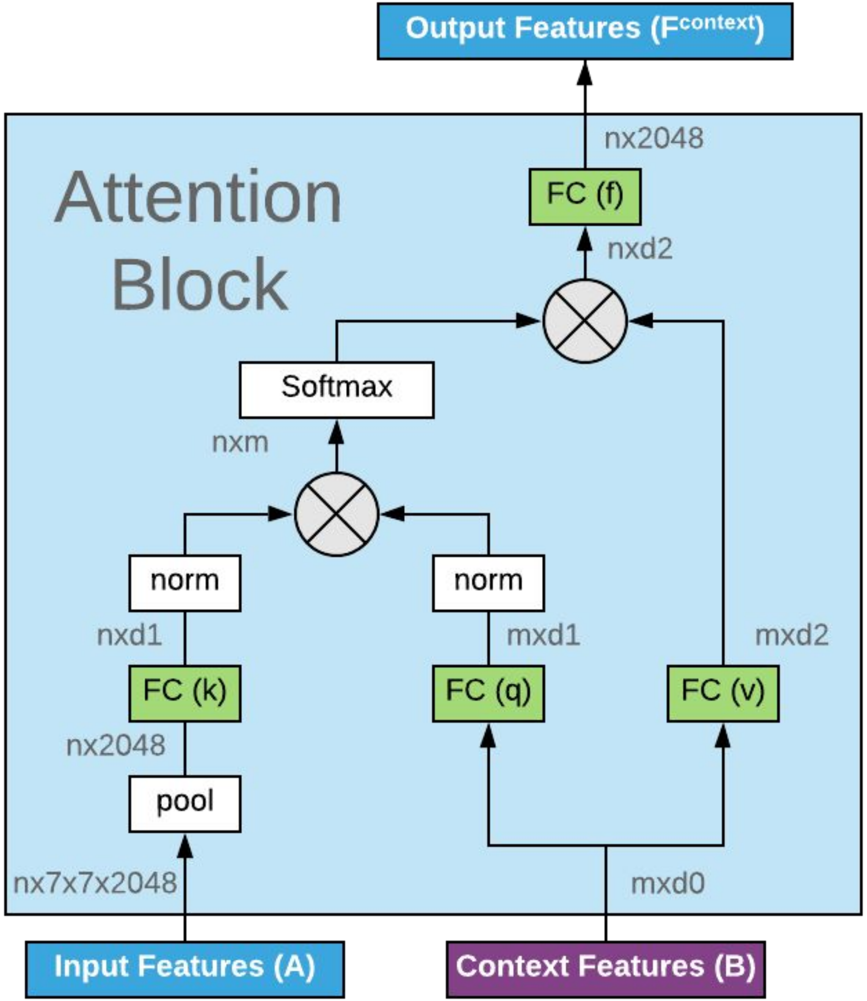
[Beery, Wu, Rathod, et al.]

Context R-CNN Architecture



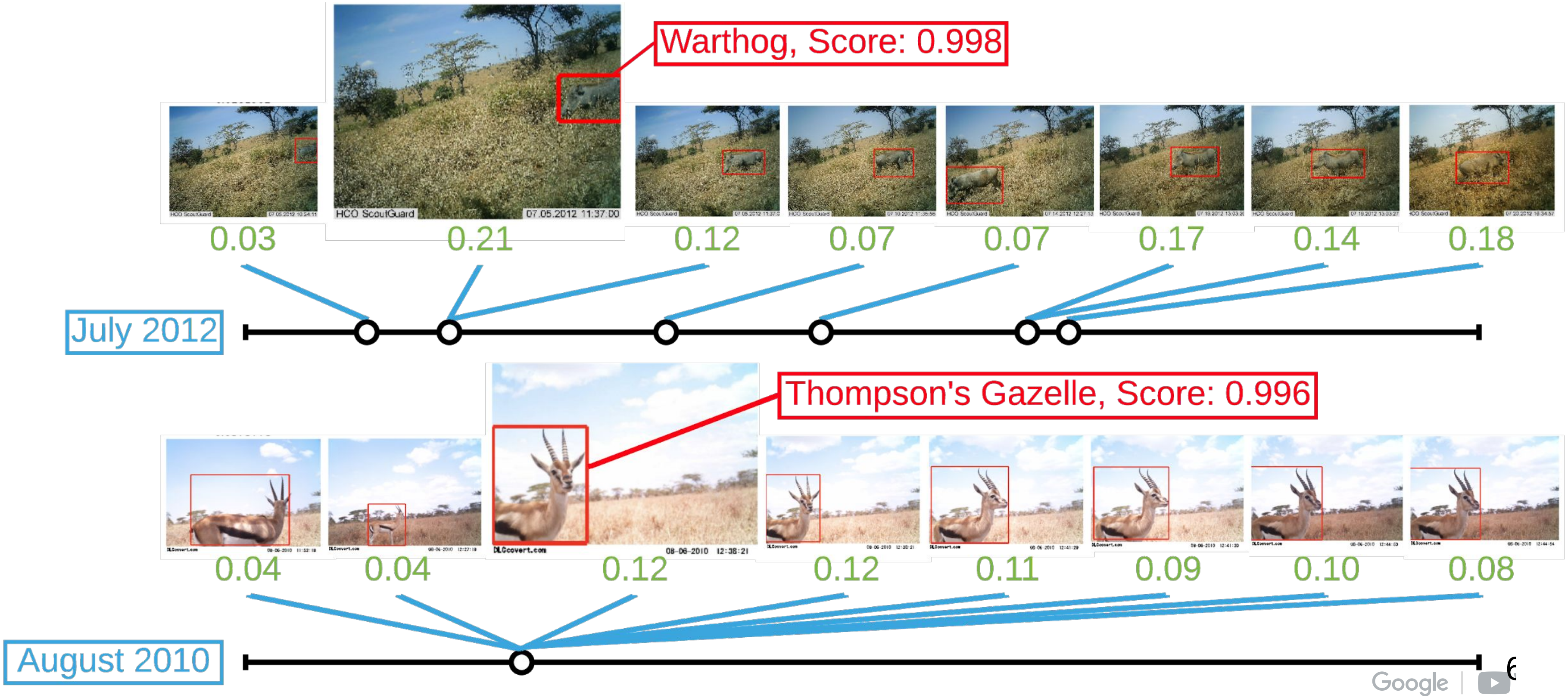
[Beery, Wu, Rathod, et al.]

Context is incorporated based on relevance

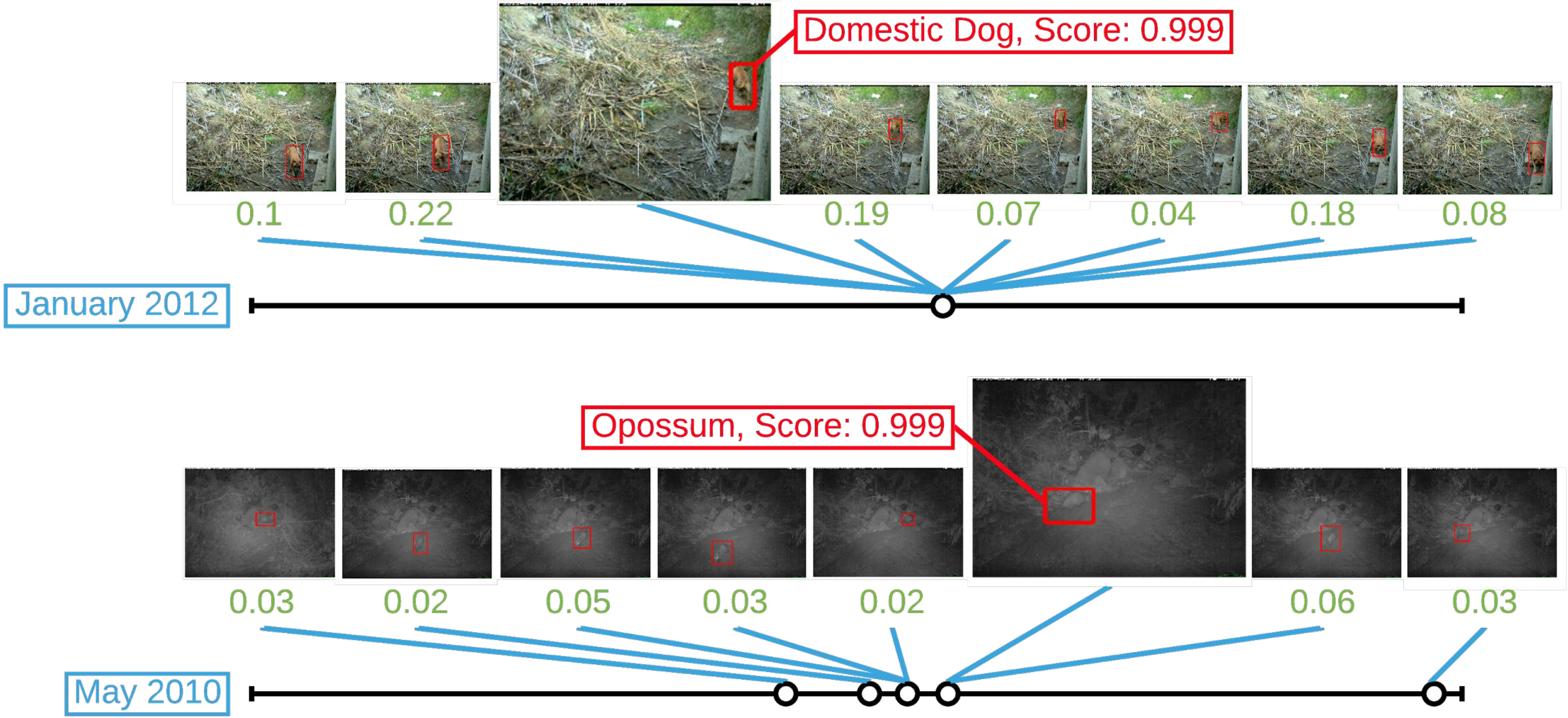


[Beery, Wu, Rathod, et al.]

Attention is temporally adaptive to relevance



[Beery, Wu, Rathod, et al.]

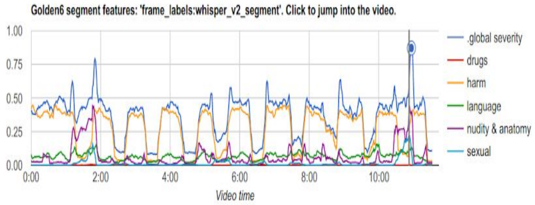


[Beery, Wu, Rathod, et al.]

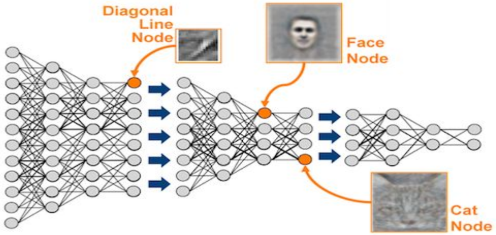
Background classes are learned without supervision

Authored video content: understand intent!

Semantic **annotation**



Objectionable content detection and localization



Learned **representations**

Authored video content: understand intent!

https://www.blog.google/outreach-initiatives/accessibility/get-image-descriptions/

Machine-generated image descriptions

Machine-generated description for this image: "Appears to be: Person playing guitar on the sofa."

Machine-generated description for this image: "Appears to be: Fruits and vegetables at the market."

Person playing guitar on the sofa

Fruits and vegetables at the market

Google | YouTube

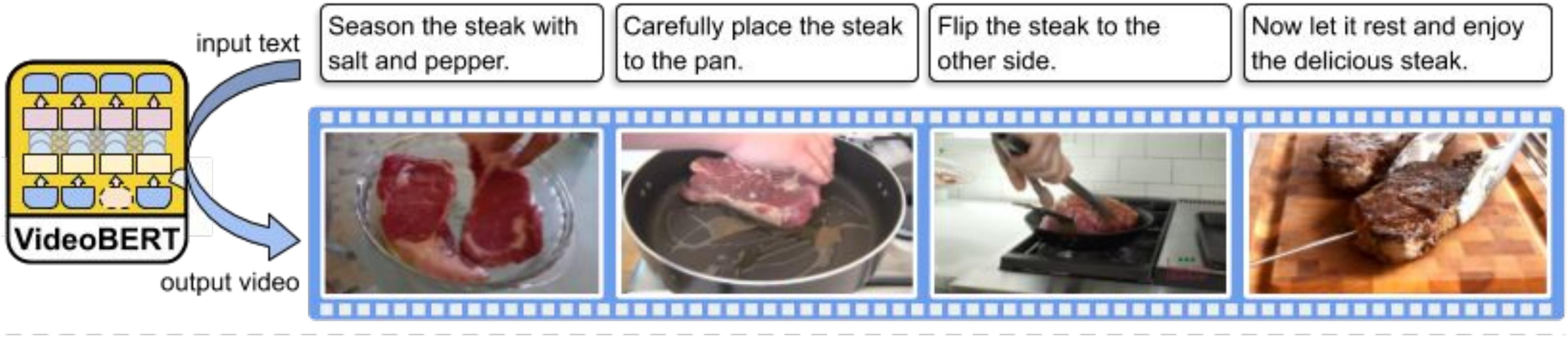
Early Steps:

Exploiting Speech to Train Video Representations

- Cross-modal weak supervision: ASR ⇔ vision
  (exploits co-occurring but noisy speech to supervise representation)
- Multimodal input: audio + video
- Some progress towards longer video sequences

Video BERT [Sun et al., ICCV'19] and Contrastive Bidirectional Transformer [Sun et al., arxiv 2019]

Video BERT



- Example 1: Given recipe text, generate sequence of visual tokens (retrieved from different videos)

Video BERT [Sun et al., ICCV'19] and Contrastive Bidirectional Transformer [Sun et al., arxiv 2019]

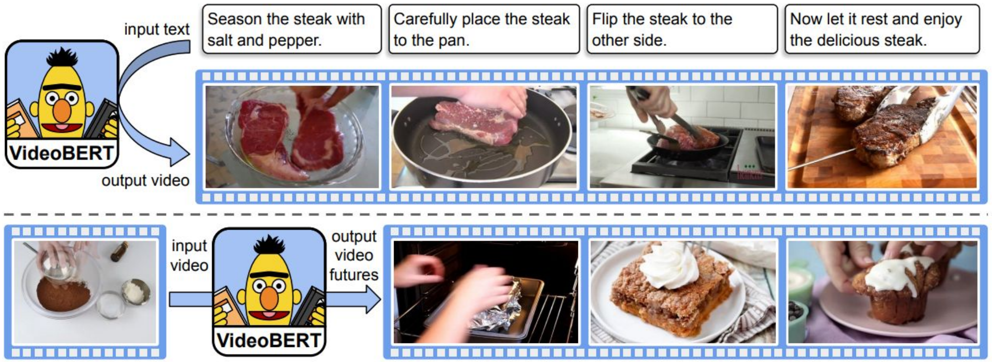
Video BERT

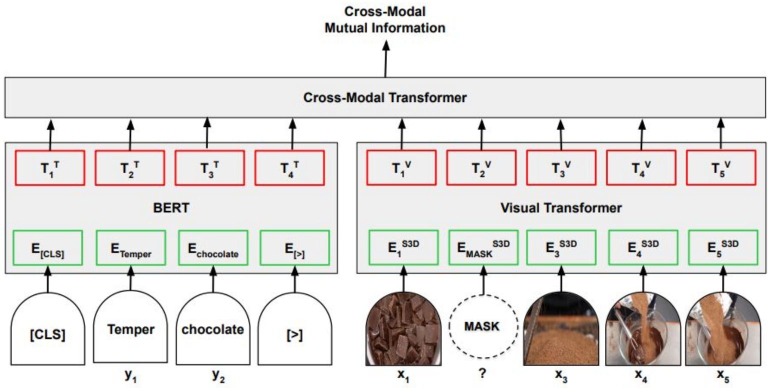
- Example 2: Given a video token, predict possible futures
(cocoa & flour mix can get baked and turn into a brownie or cupcake)

Video BERT [Sun et al., ICCV'19] and Contrastive Bidirectional Transformer [Sun et al., arxiv 2019]

[Sun, Schmid, 2019]

Joint video and language representations



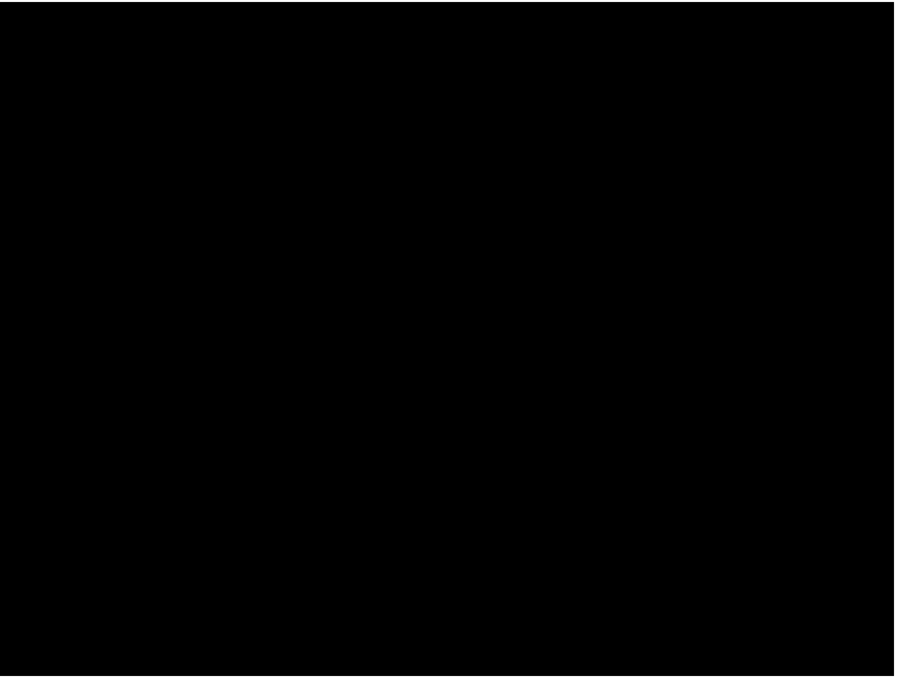


- Understand long videos with **self- (time) and cross-modal (ASR) supervisions**.
- Leverage powerful models (BERT) and pre-training tasks (masked LM).
- Cross-modal applications: zero-shot action classification, action anticipation, etc.
- Opportunities:
    - Automatic video data mining given large vocabulary (Video Search timed anchors).
    - Generic feature vectors for long videos (VCA).

Live interactions

Significance and intent both come into play

Early Steps in Situated Perception



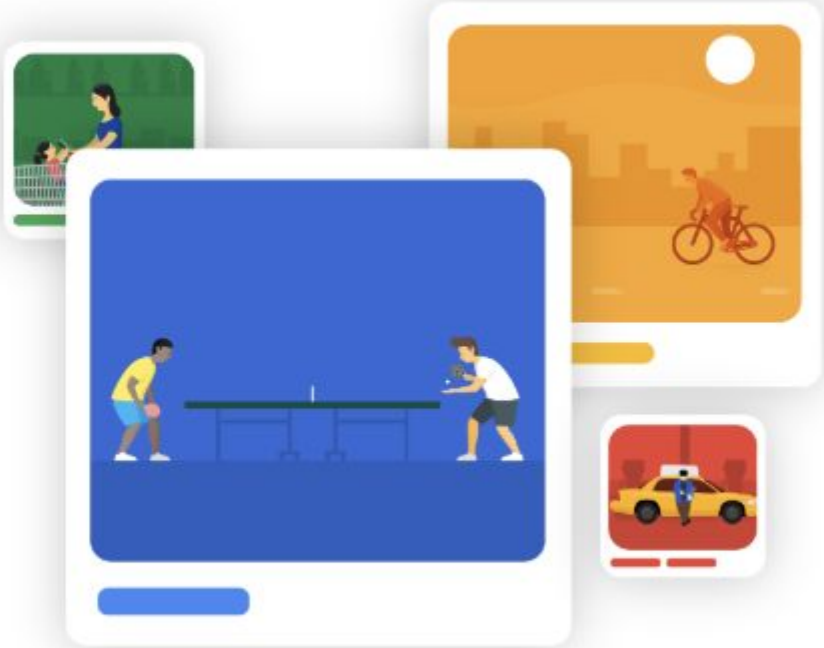
BodyPix - Real-time Segmentation in your Browser

https://github.com/tensorflow/tfjs-models/tree/master/body-pix

Natural gestures & smart camera in Google Nest Hub Max

Image credit: https://www.dailydot.com/wp-content/uploads/2019/05

Need to understand what to leave out



Media retargeting as a (small) window into this problem

https://ai.google.com/research/ConceptualCaptions

Retargeting in time

Video

Preview (6s)

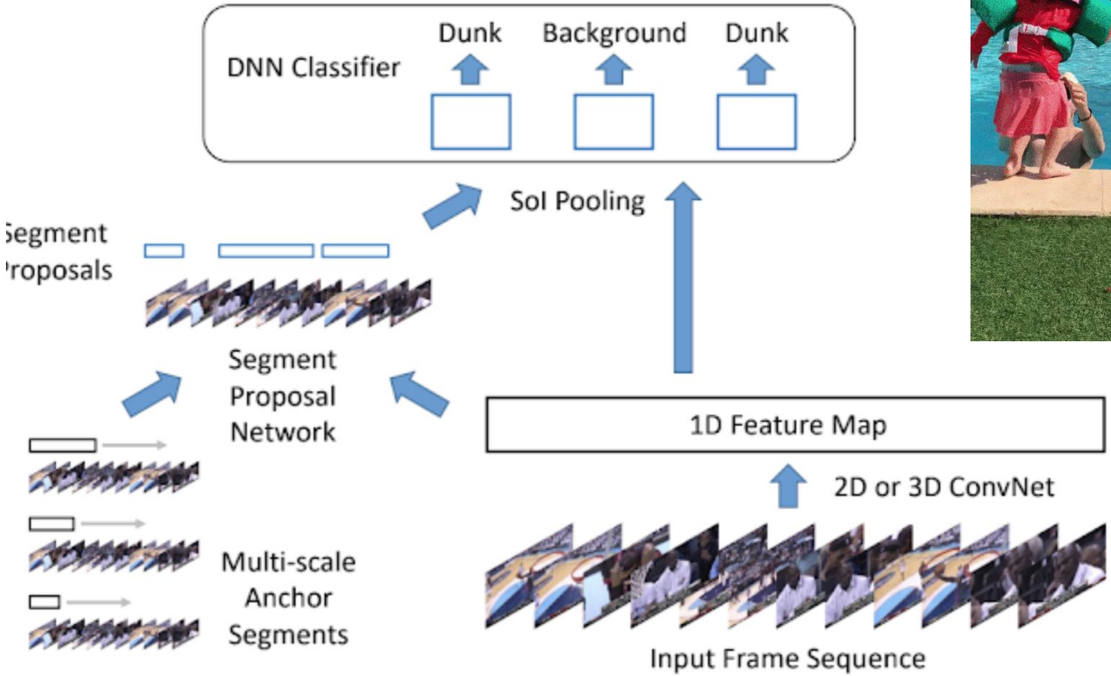
Summary (9s)

Video ad (1 minute)

Bumper ad

Retargeting in time

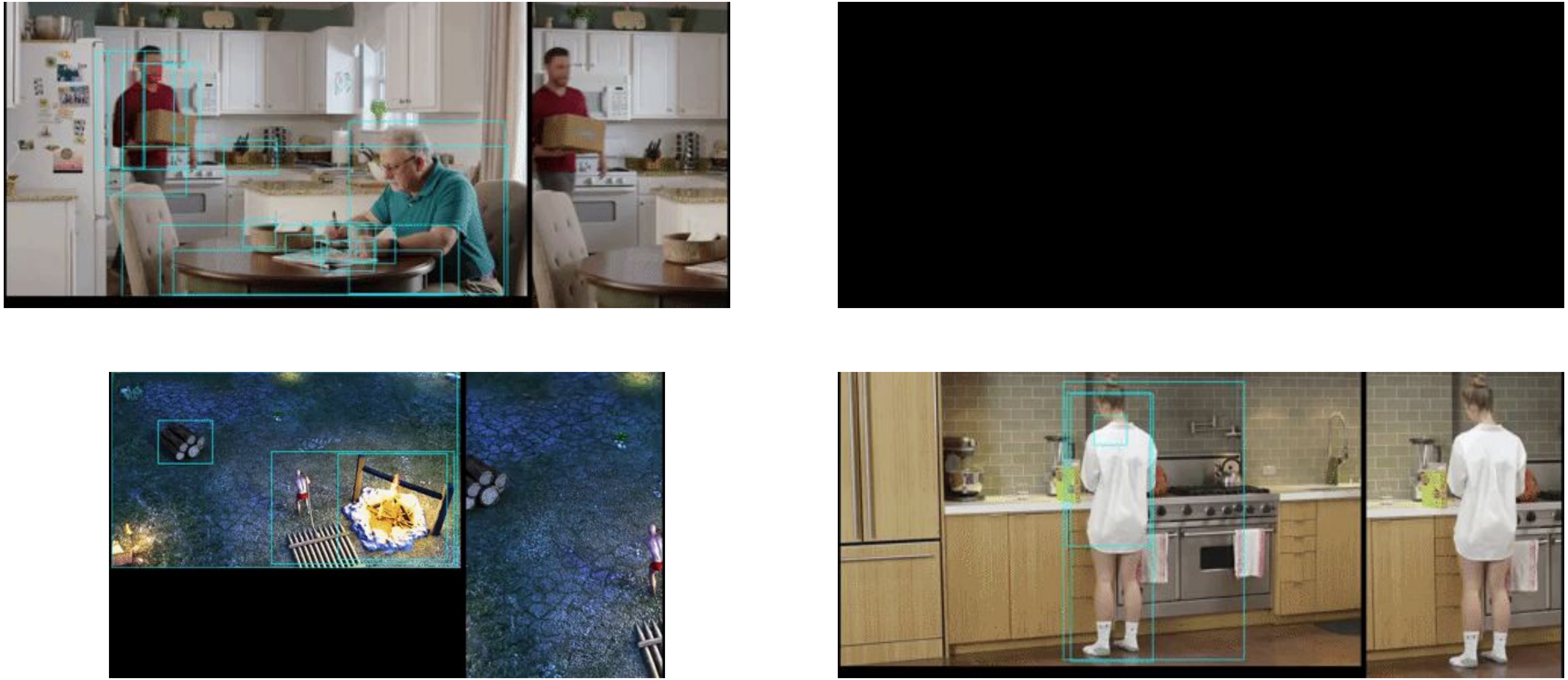
[Chao, Vijayanarasimhan, Seybold, et al.]



Architecture for temporal action localization

Special moments in video

Retargeting in space

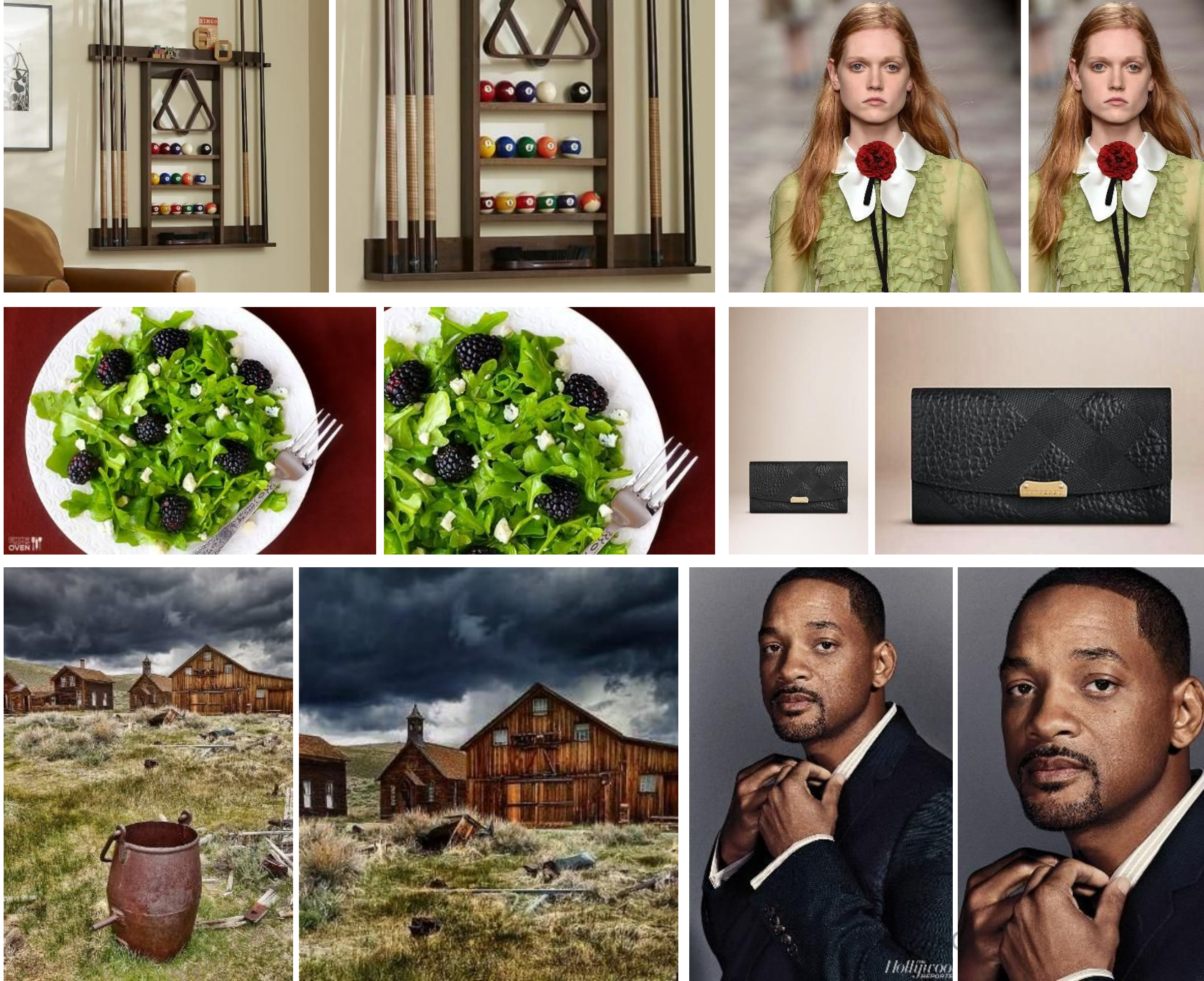


L: original landscape; R: AutoFlip
[https://www.blog.google/products/ads/level-your-gaming-business-new-innovations-apps

Retargeting in Space

Deep-learned, aesthetic cropping

Improves quality and generates fewer bad crops by training on millions of professional-quality photos



[Fang, Zhang]

Directions in ML

Understanding **intent** and **importance** in audio and video

- detection and categorization are only part way

  also have: 3D understanding, people-centric information, action/interaction recognition

- missing piece: need to determine **what to leave out**
  - want a **synopsis** for **authored** (and situated) media
  - want less constrained interactions for **live** situations

Generating new **creative** content can help highlight shortcomings

(as well as providing useful content)