DeepMind

UNIVERSITY OF OXFORD

Deep Learning for Video Understanding

Andrew Zisserman

IISc, August 2022

Video Understanding

- What is in the video?
  - objects, animals, people …
- Where is it?
  - 3D scene spatial layout
  - object shape
  - human pose …
- What is happening?
  - actions
  - activities …

Video Understanding

- What is in the video?
  - **objects, animals, people** …
- Where is it?
  - 3D scene spatial layout
  - object shape
  - human pose …
- What is happening?
  - **human actions**
  - activities …

Objective: learning to recognize without explicit supervision

Outline

Prelude: learning to recognize with strong (explicit) supervision

- Example: recognizing human actions in video

How can we learn to recognize without explicit supervision?

Part I: Using multi-modal (audio-visual) self-supervised learning

Part II: Discovering objects and their effects using self-supervised learning

Part III: Using weak supervision from videos with text

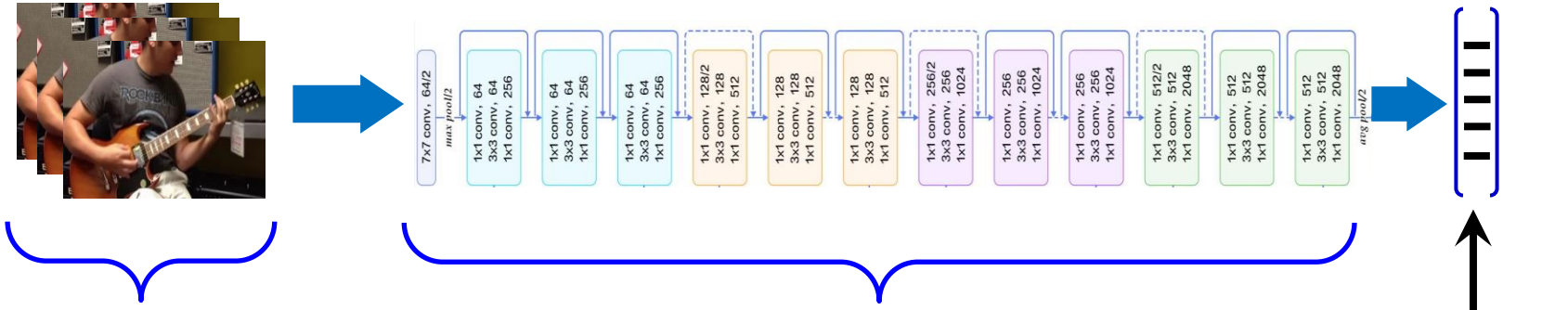
Prelude

Learning to recognize with strong (explicit) supervision

Representing human actions

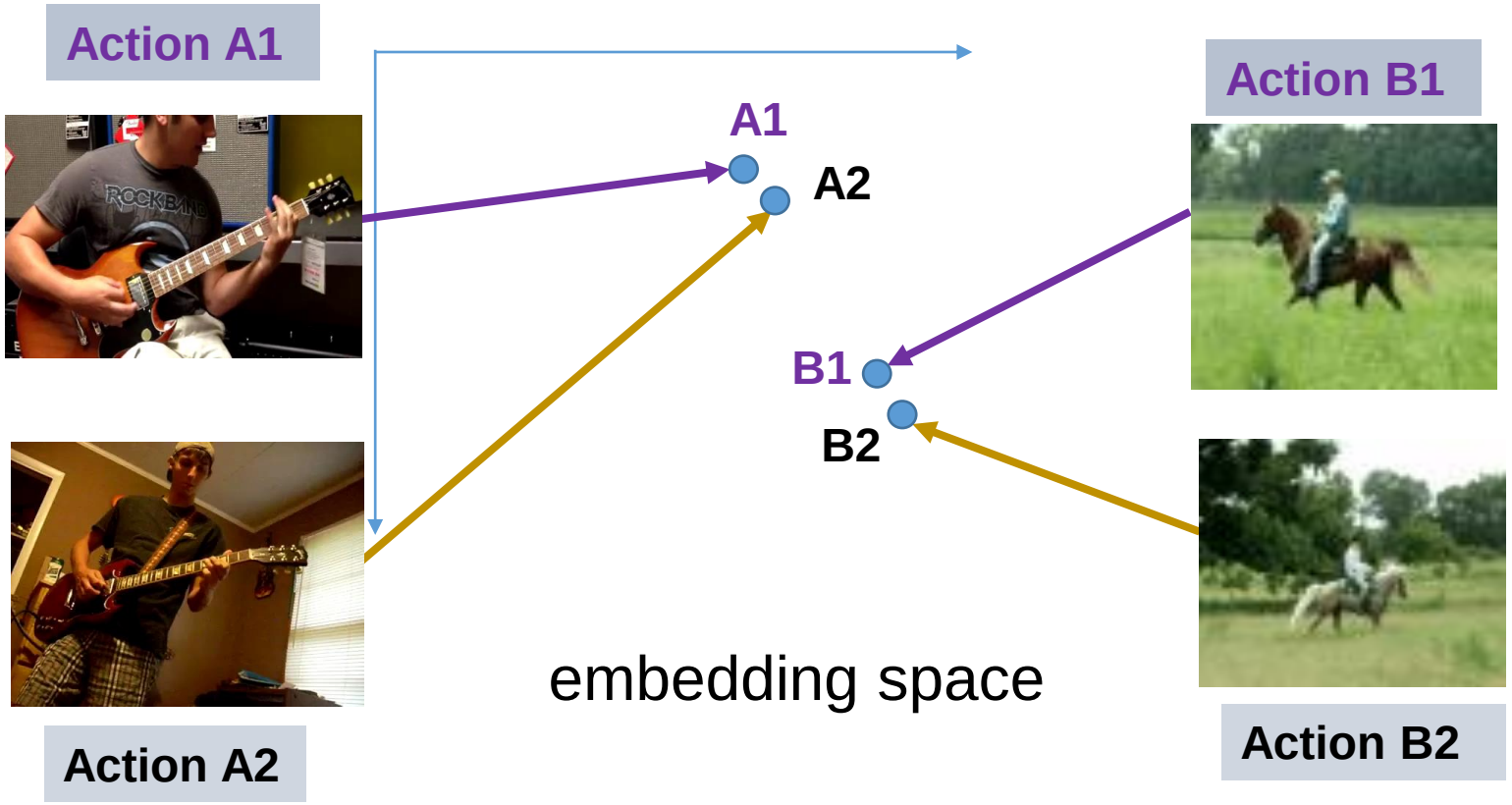
Learn a video clip embedding for action recognition

Map from video clip to a vector



Use action vector for classification, localization, retrieval …

Objective of the action embedding



Action vectors of the same (semantic) action should be

- close, and
- distinct from those of different actions

Steps for “Classical” learning with strong supervision

- A large-scale video dataset with labels – for human actions
- Choose a suitable deep network
- A loss function – cross entropy for classification
- Train network with back propagation to minimize loss

The Kinetics Human Action Dataset

The Kinetics Human Action Video Dataset



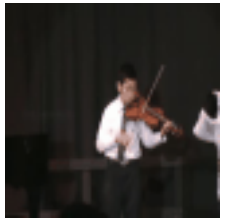
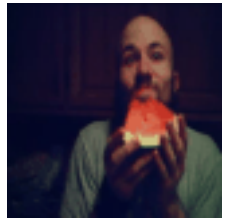













archery

country line dancing

riding or walking with horse

playing violin

eating watermelon

Kinetics datasets overview

- Stats:

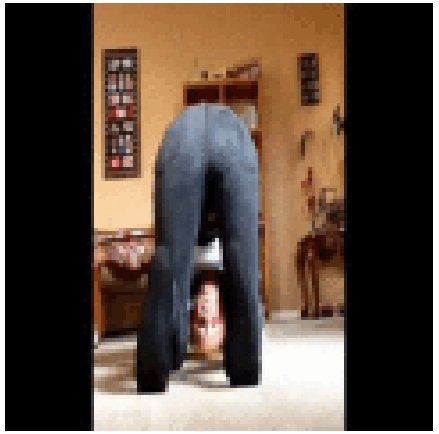
	Year	Action classes	Clips per class	Total
Kinetics-400	2017	400	400-1000	300k
Kinetics-600	2018	600	600-1000	500k
Kinetics-700	2019	700	600-1000	650k

- 10s clips
- Every clip from a different YouTube video:
  - huge variety in people, viewpoint, scenes, execution …

- *The Kinetics Human Action Video Dataset*. Kay, Carreira, Simonyan, Zhang, Hillier, Vijayanarasimhan, Viola, Green, Back, Natsev, Suleyman, Zisserman, arXiv 2017
- *A Short Note about Kinetics-600*. Carreira, Noland, Banki-Horvath, Hillier, Zisserman, arXiv 2018
- *A Short Note on the Kinetics-700 Human Action Dataset*. Carreira, Noland, Hillier, Zisserman, arXiv 2019

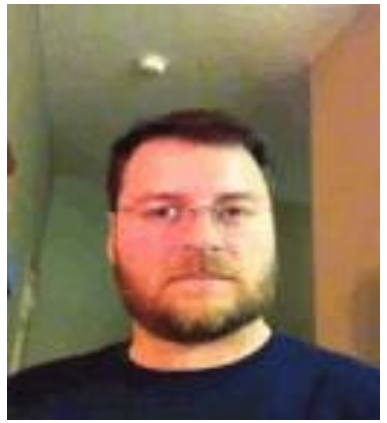
Datasets available for download from: cvdfoundation

Person Actions (Singular)

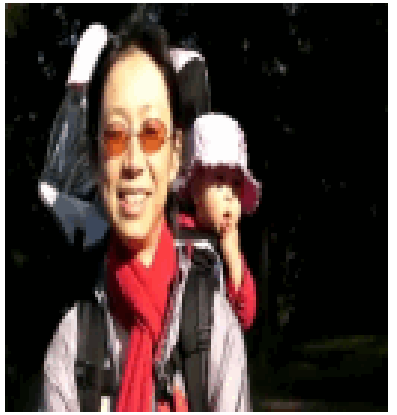



Head stand

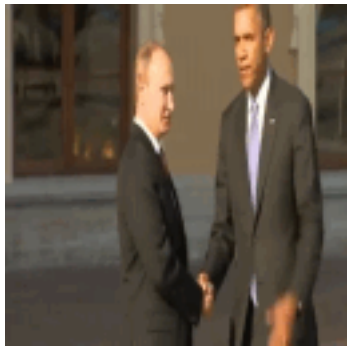




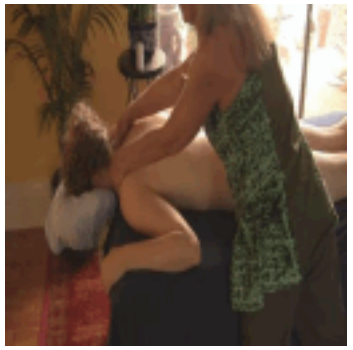
Shaking Head

Person-Person Actions

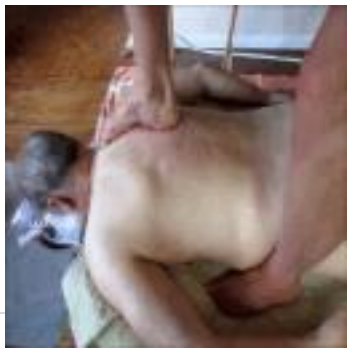




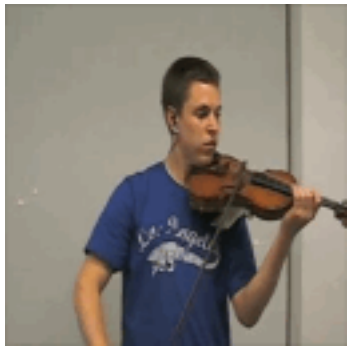
Shaking Hands

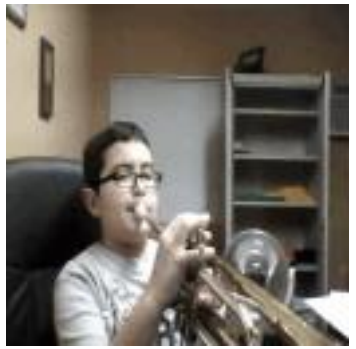

Massaging Back



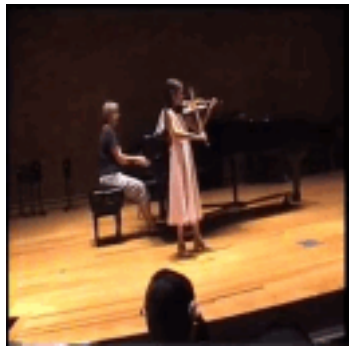



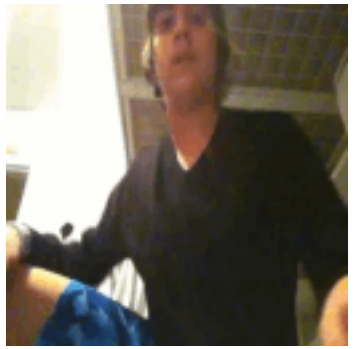
Person-Object Actions

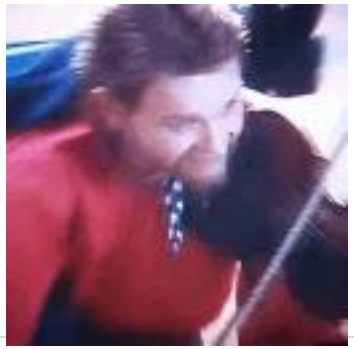


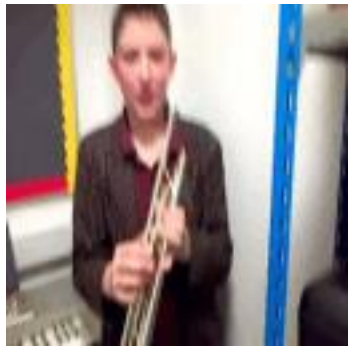












Playing Violin

Playing Trumpet

Google DeepMind

More Person-Object Actions

Using sledgehammer

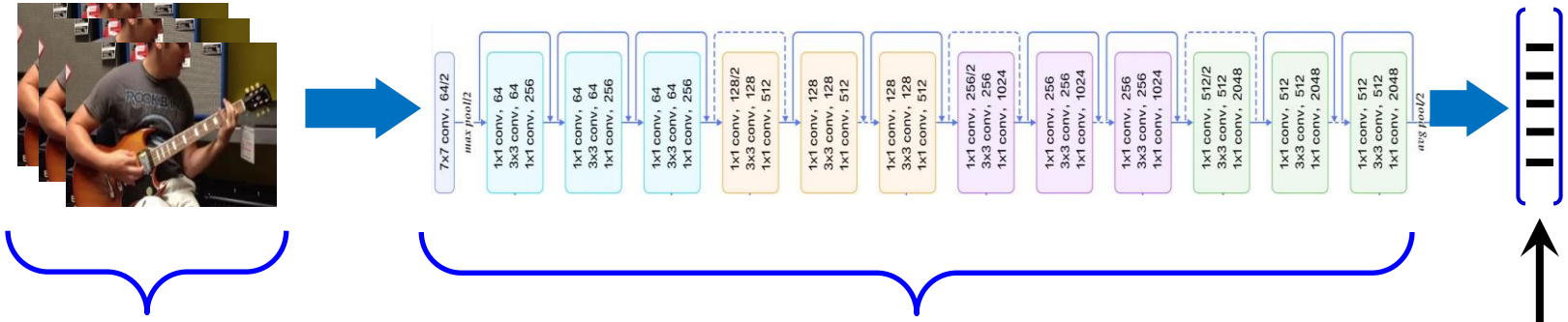
Using power drill

Also using paint roller, circular saw, wrench, others

Objective – human action classification

Learn a video clip embedding for action recognition

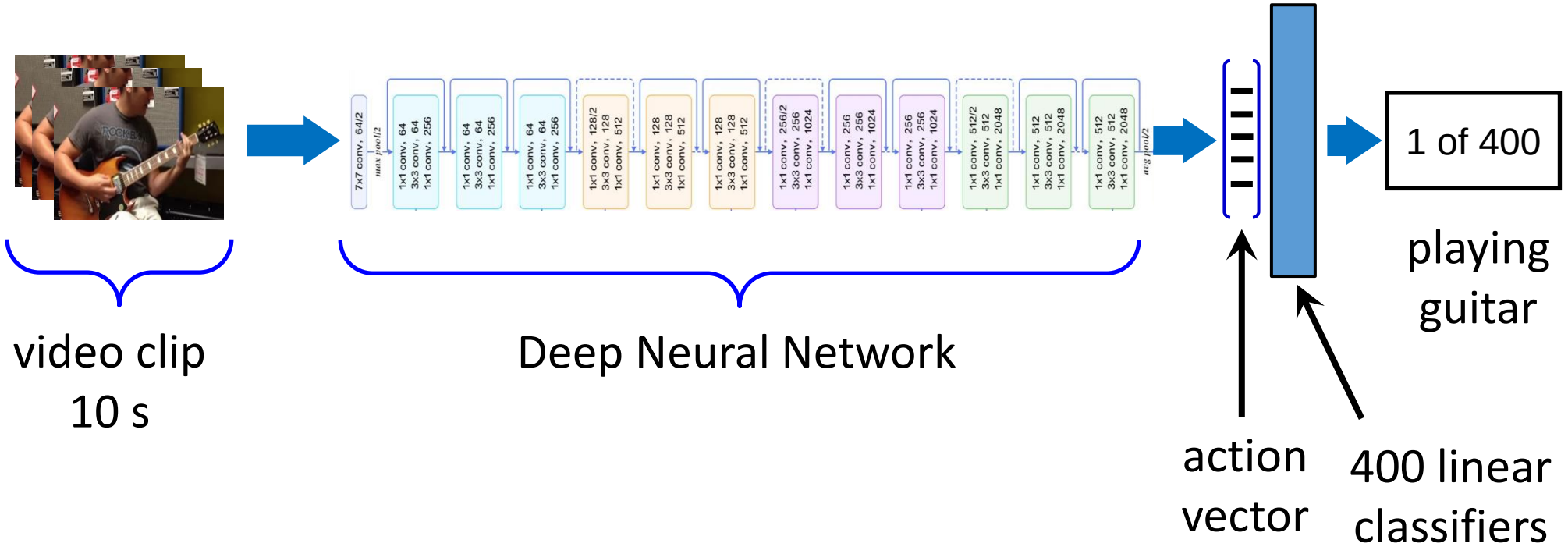
Map from video clip to a vector



Strong supervision: train network to classify actions on the Kinetics dataset

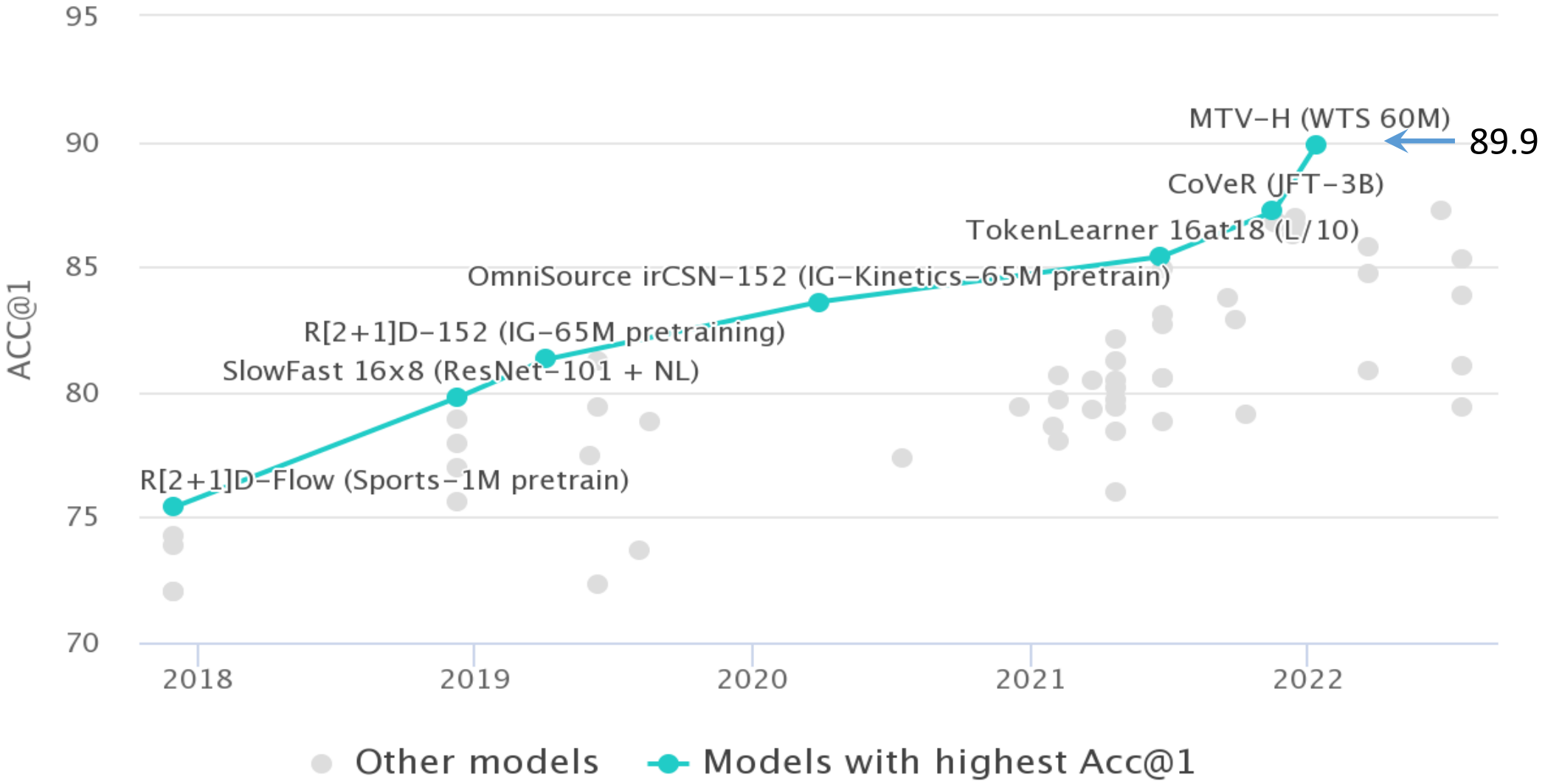
Train the network using strong supervision

- Kinetics 400: Multiway classification - add 400 linear classifiers (one for each action class)
- Train network to predict class label for each clip with a cross-entropy loss function using back-prop



- Outcome: network generates action (feature) vectors that allow correct classifications

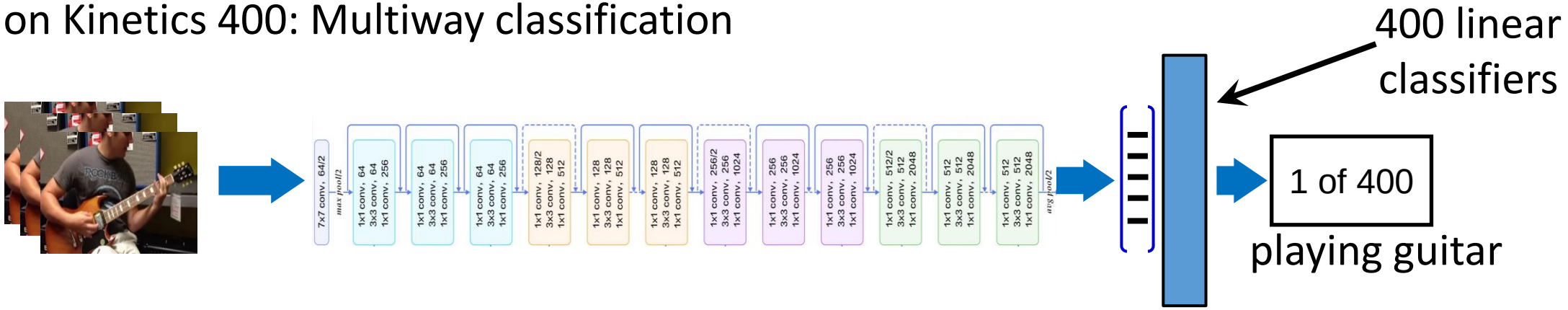
Performance on Kinetics-400 val



Compilation of results from `papers with code' website paperswithcode.com

Use trained network for `downstream' tasks

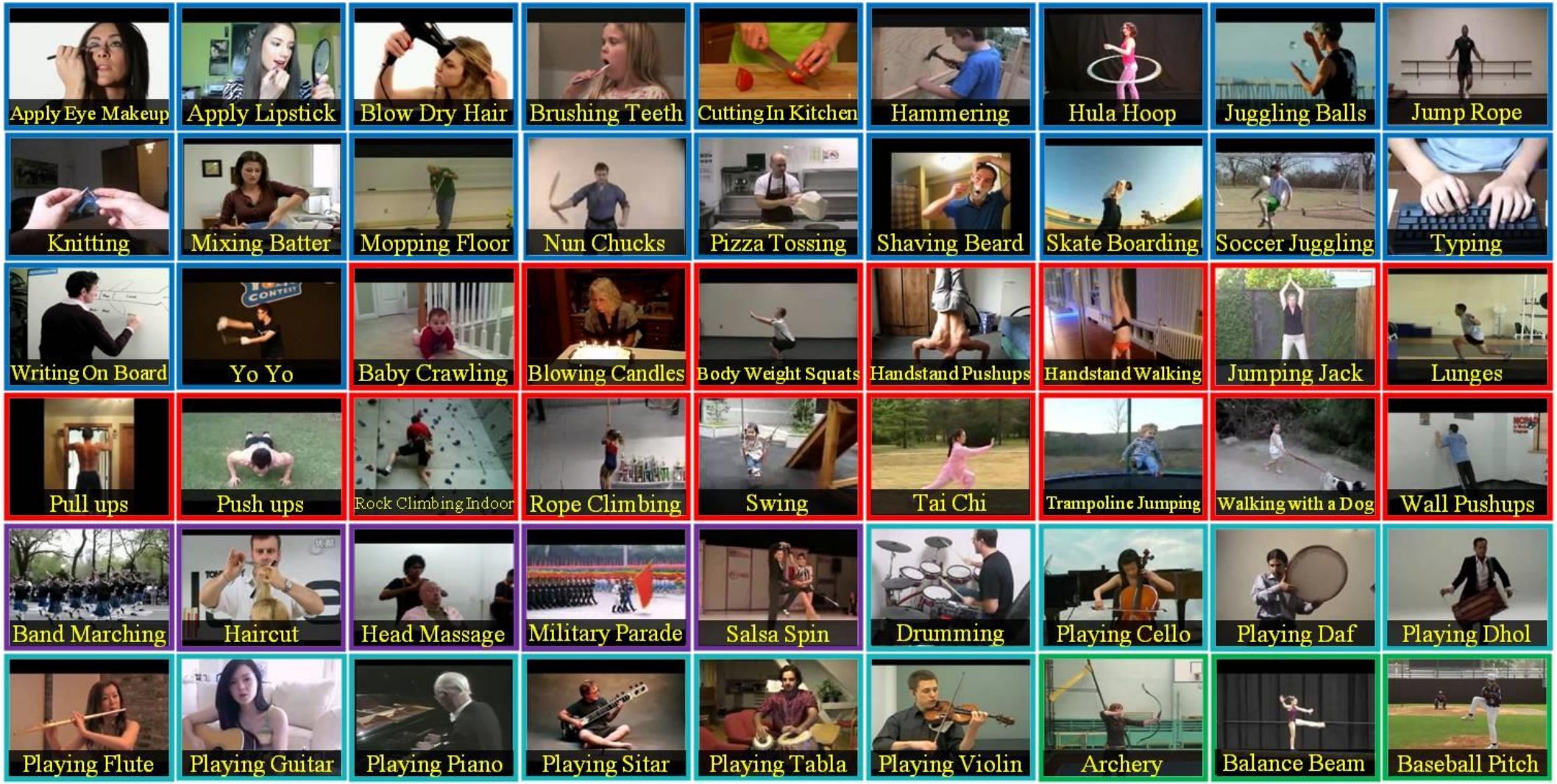
1. Train on Kinetics 400: Multiway classification



2. Use network for new task, e.g. classification on a new video dataset: UCF-101

- Train only linear classifiers, or fine-tune network starting from pre-trained weights

UCF-101 video human action dataset



101 human action classes

Performance on UCF-101

Compilation of results from `papers with code' website paperswithcode.com

Steps for “Classical” learning with strong supervision

- A large-scale video dataset with labels – for human actions
- Choose a suitable deep network
- A loss function – cross entropy for classification
- Train network with back propagation to minimize loss

Part I

Learning Human Action Representations using Audio-Visual Self-Supervision

Why Self-Supervision?

Rather than “strong supervision”, e.g. on Kinetics:

1. Expense of producing a new labelled dataset for each new task

2. Some areas are supervision-starved, e.g. medical data, where it is hard to obtain annotation

3. Untapped/availability of vast numbers of unlabelled images/videos
   - Facebook: one billion images uploaded per day
   - 300 hours of video are uploaded to YouTube every minute

4. How infants may learn …

Self-Supervised Learning

The Scientist in the Crib: What Early Learning Tells Us About the Mind
by Alison Gopnik, Andrew N. Meltzoff and Patricia K. Kuhl

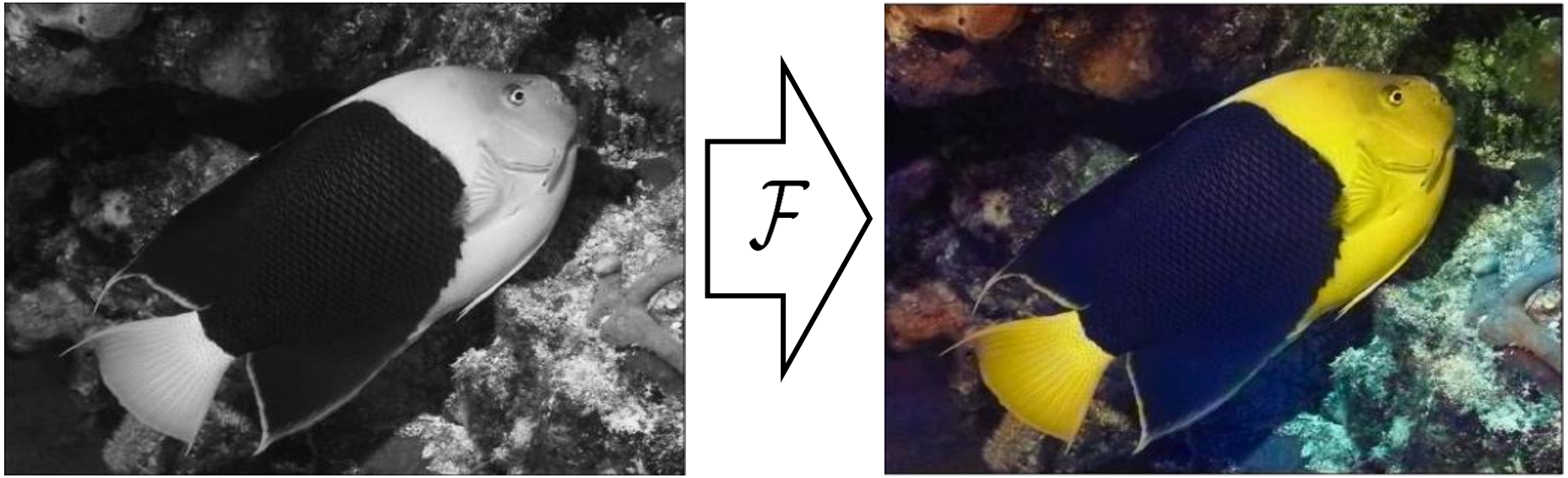
The Development of Embodied Cognition: Six Lessons from Babies
by Linda Smith and Michael Gasser

What is Self-Supervision?

- A form of unsupervised learning where the data provides the **supervision**
- In general, withhold some part of the data, and task the network with predicting it
- The task defines a proxy loss, and the network is forced to learn what we really care about, e.g. a semantic representation, in order to solve it
- Inspiration from NLP learning methods such as word2vec

Image example 1: colourization

Train network to predict pixel colour from a monochrome input



Grayscale image: *L* channel

$$\mathbf{X} \in \mathbb{R}^{H \times W \times 1}$$

Concatenate (L,ab)

$$(\mathbf{X}, \hat{\mathbf{Y}})$$

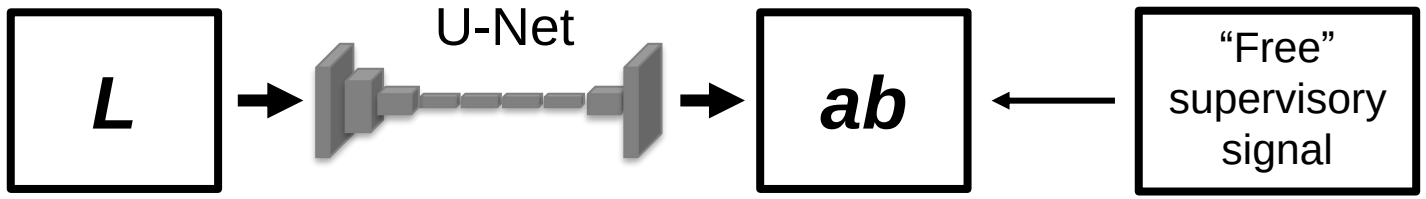
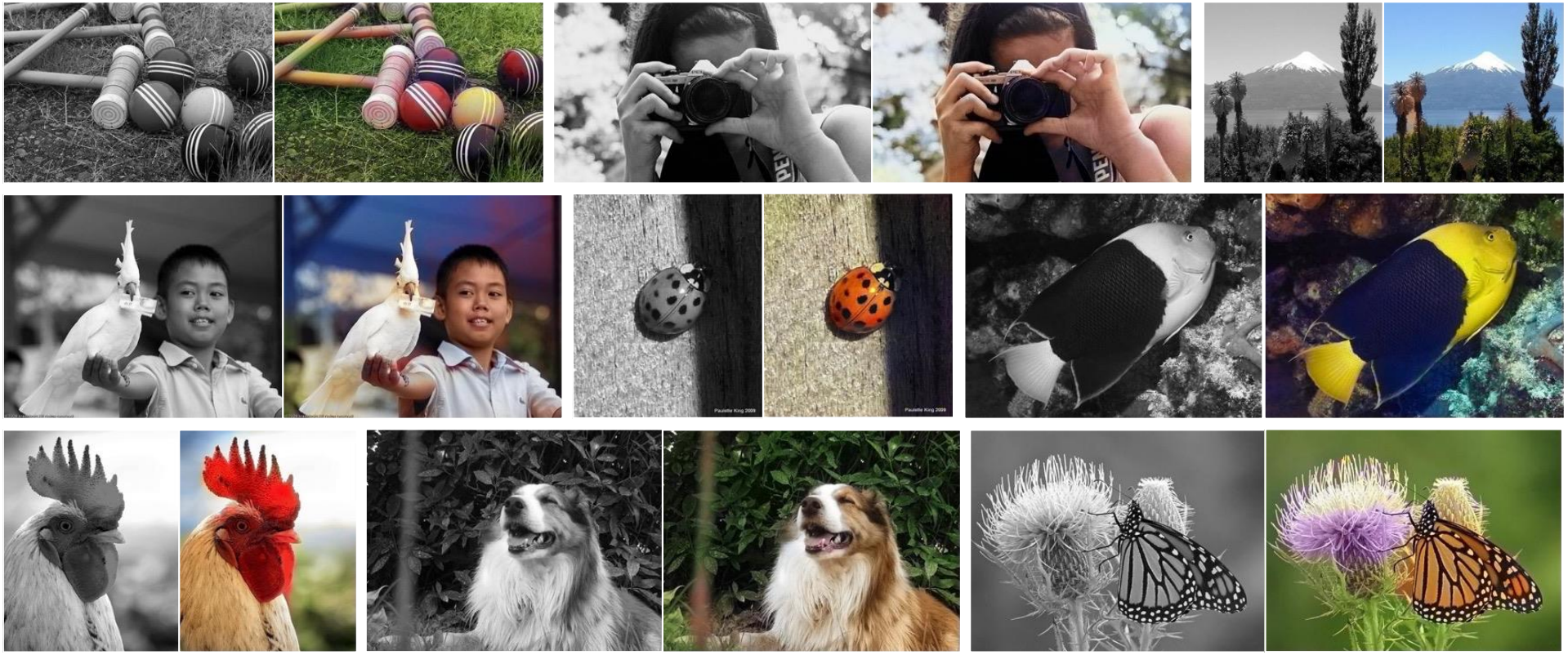


Image example 1: colourization

Train network to predict pixel colour from a monochrome input



“Colorful Image Colorization”, Richard Zhang, Phillip Isola, Alexei A. Efros, ECCV 2016

Image example 2: Masked Auto-Encoding

Train network to predict masked patches from input image

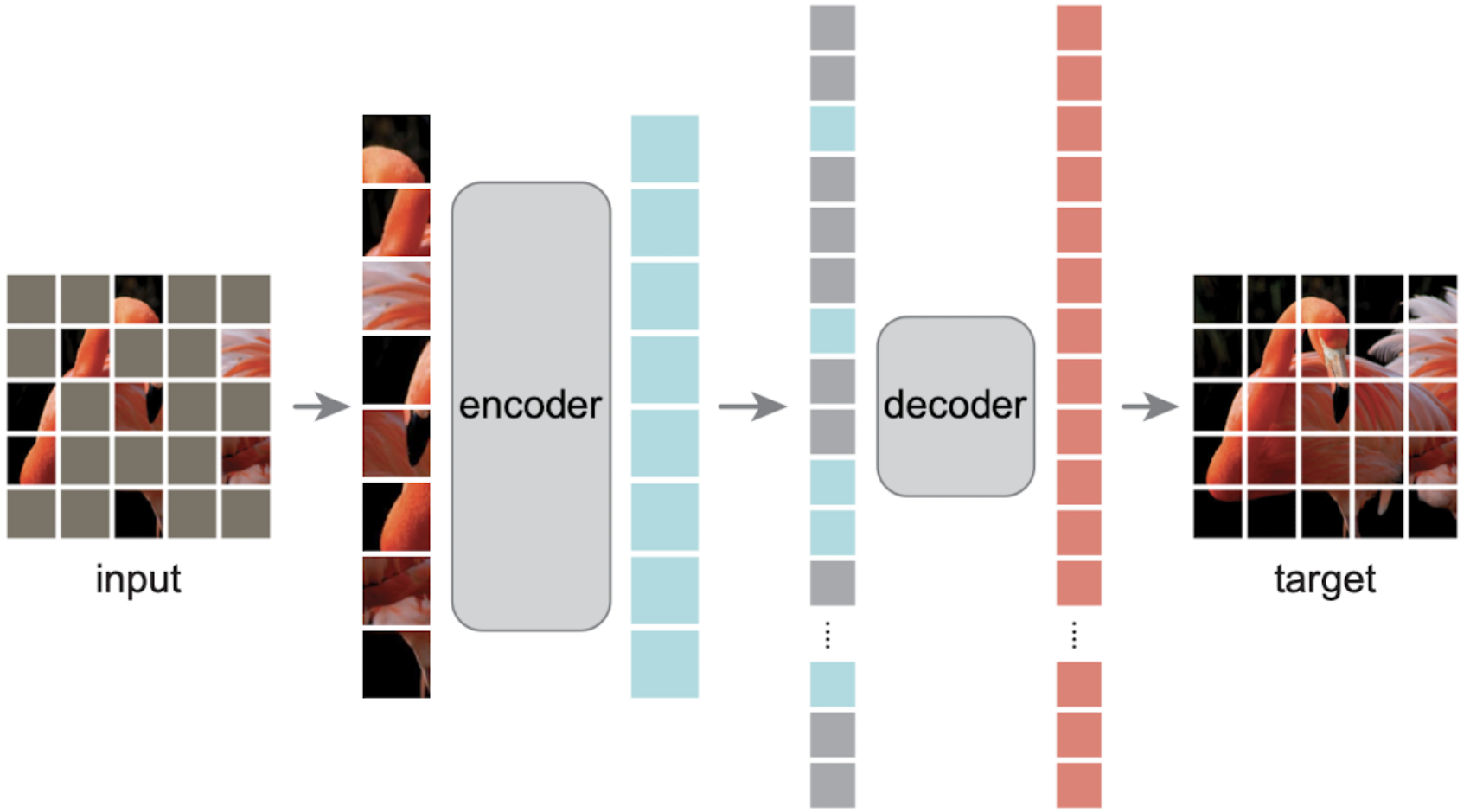
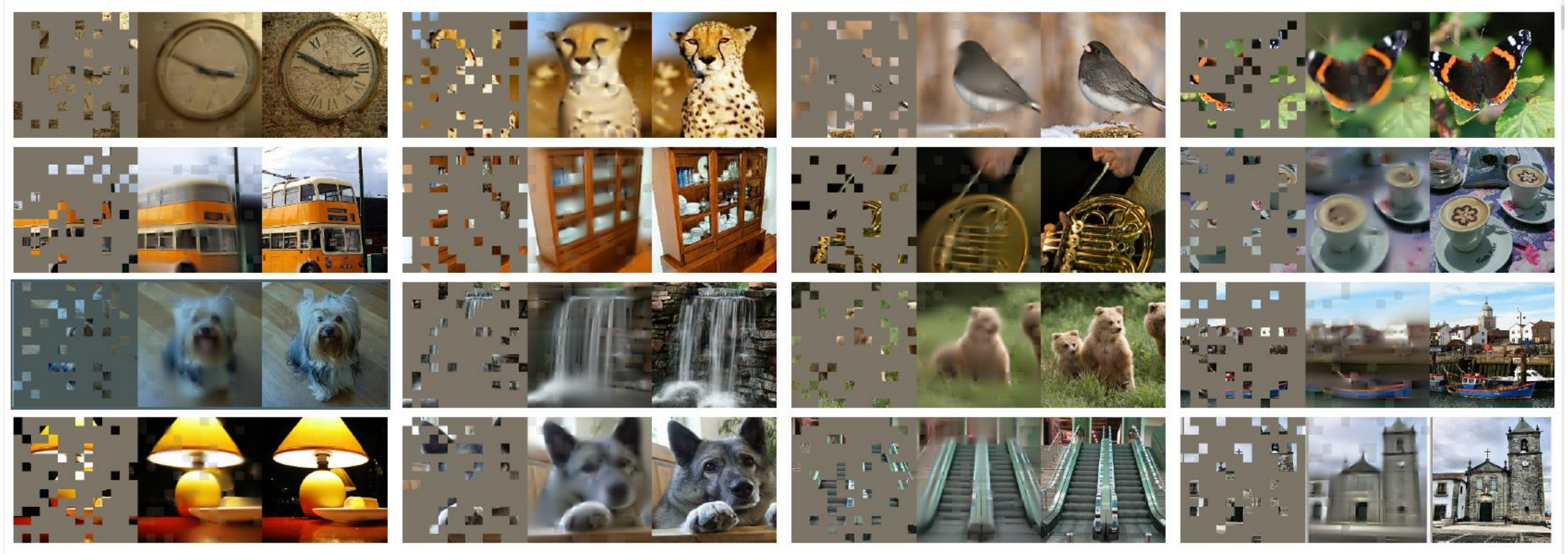


Image example 2: Masked Auto-Encoding

Train network to predict masked patches from input image



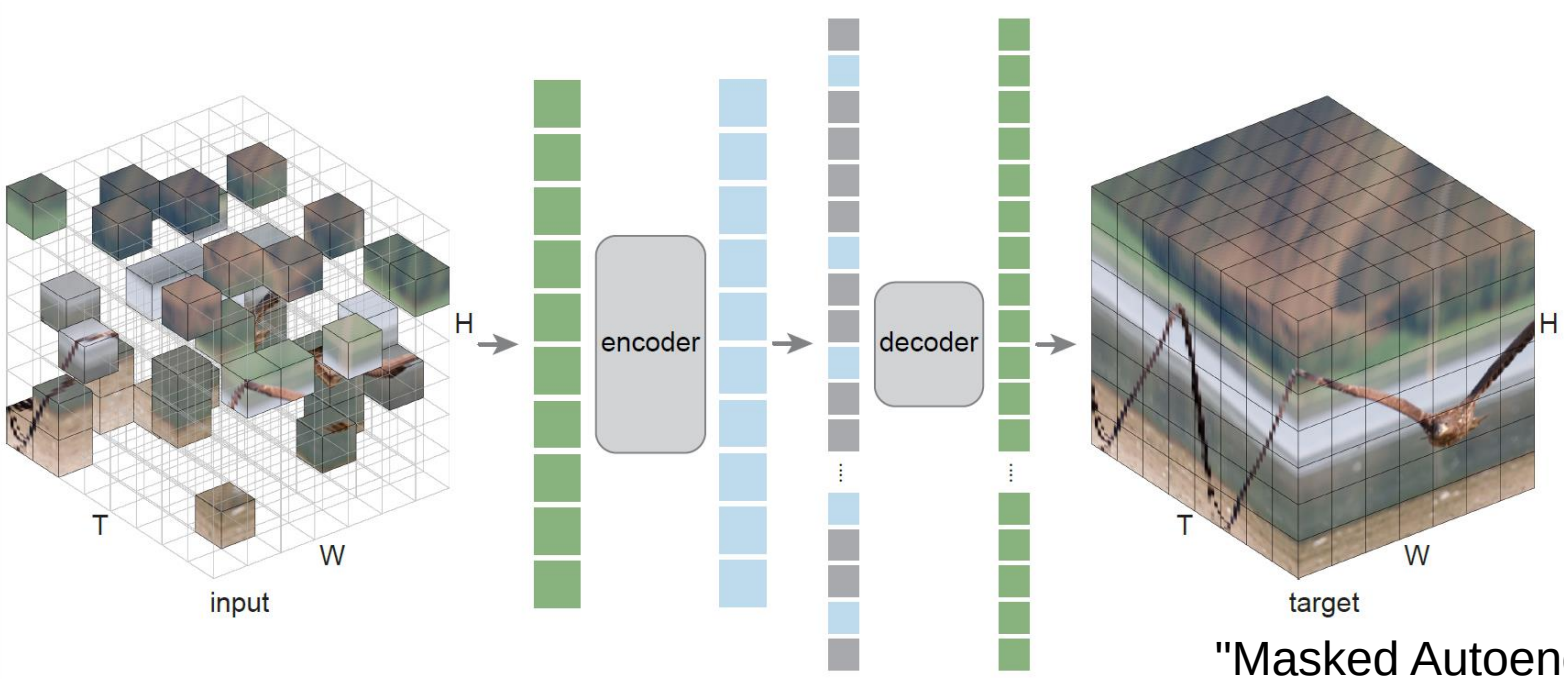
"Masked autoencoders are scalable vision learners", Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick, CVPR 2022

"SimMIM: A Simple Framework for Masked Image Modeling", Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu, CVPR 2022

Self-supervised learning for video

Video inherits all the image proxy tasks (e.g. instance discrimination, clustering, MAE) at the frame level, e.g.

Masked Auto-Encoding for Videos: train network to predict masked spatio-temporal patches from input video



"Masked Autoencoders as Spatio-temporal Learners",
Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, Kaiming He,
ArXiv, 2022

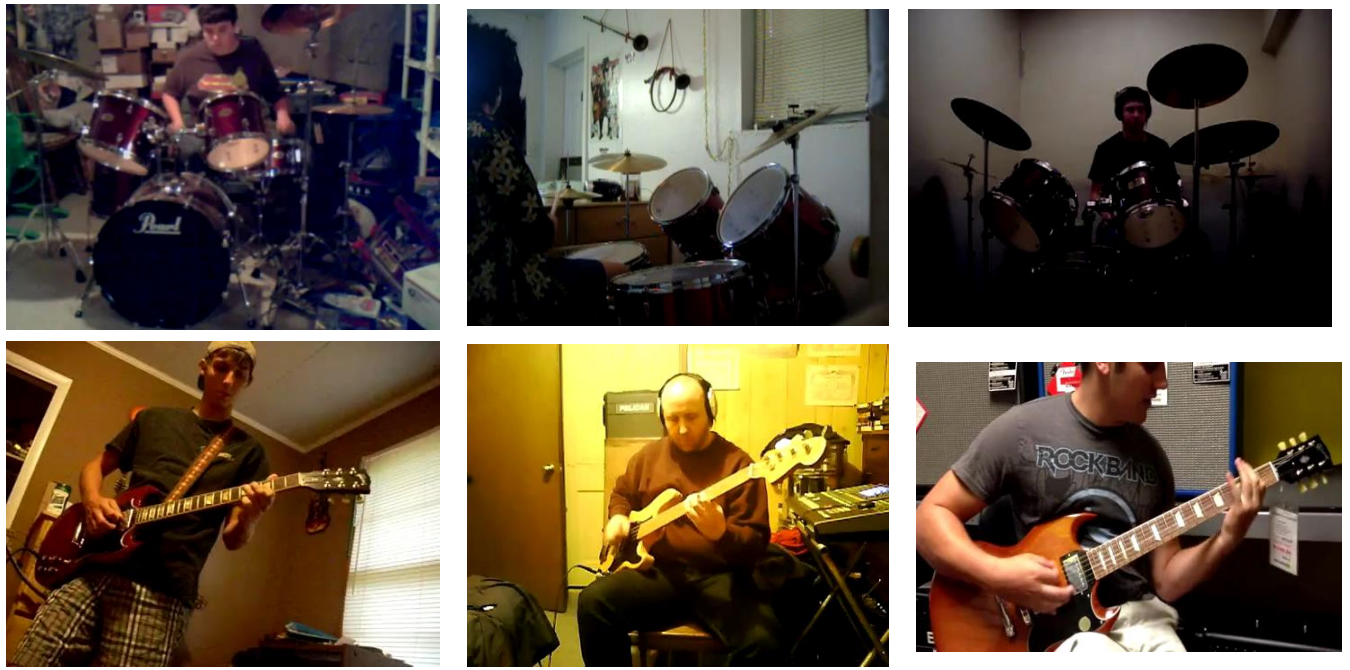
but also has proxy tasks particular to the video domain …

Self-supervised learning for video

Video, beyond images, naturally …

- extends and develops sequentially in time,
- has multiple modalities (audio stream),
- has motion (optical flow stream)

Audio-Visual Co-supervision



Sound and frames are:

- Semantically consistent
- Synchronized

Audio-Visual Co-supervision

Objective: use vision and sound to learn from each other

- Sound and frames are (i) semantically consistent, and (ii) synchronized

- Two types of proxy task:
    1. Predict audio-visual **correspondence**
    2. Predict audio-visual **synchronization**

Audio-Visual Co-supervision

Objective: use vision and sound to learn from each other

- Sound and frames are (i) semantically consistent, and (ii) synchronized

- Two types of proxy task:
    1. Predict audio-visual **correspondence**
    2. Predict audio-visual **synchronization**

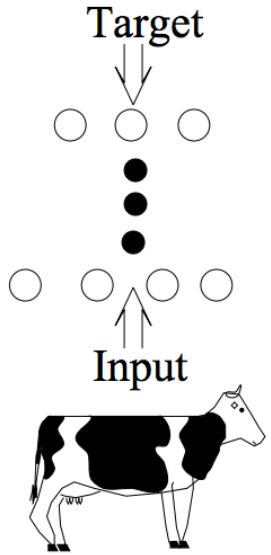
Background

Virginia de Sa. Learning Classification with Unlabeled Data. NIPS 1994

Supervised

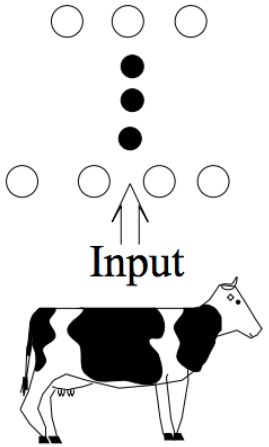
- implausible label

"COW"



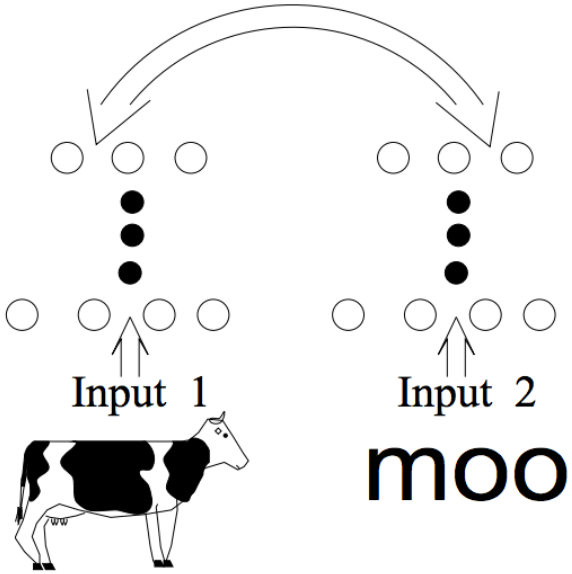
Unsupervised

- limited power



Self-Supervised

- derives label from a co-occurring input to another modality



Audio-Visual Co-supervision

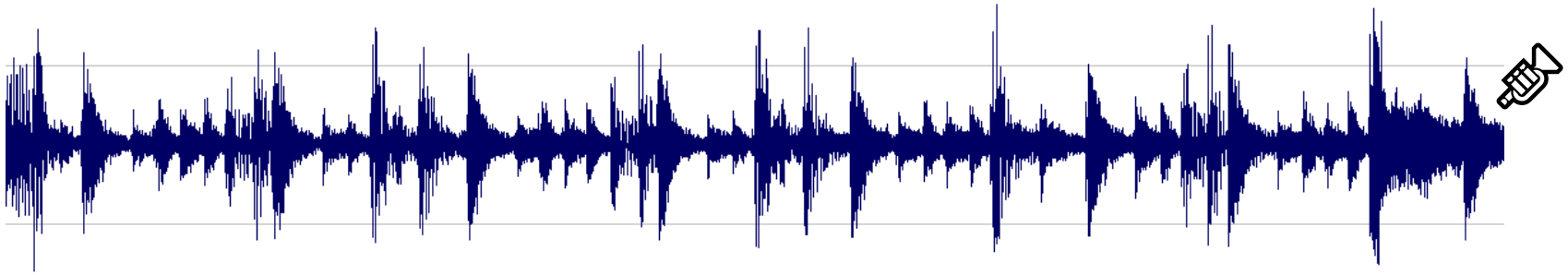
Train a network to predict if **video** and **audio** clip correspond

Correspond?

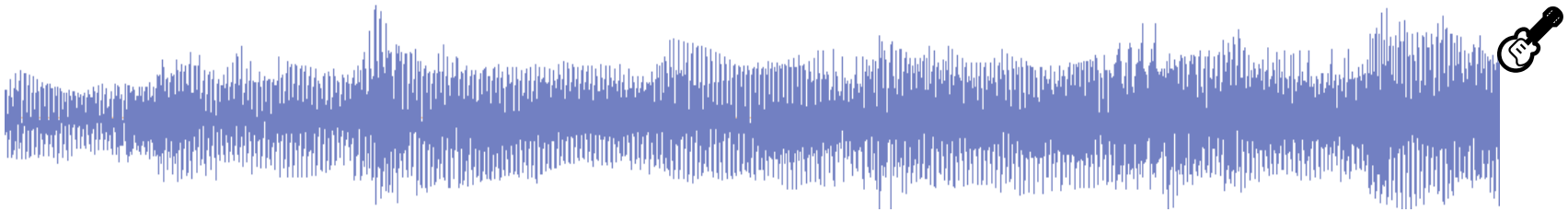
“Objects that Sound”, Arandjelović and Zisserman, ICCV 2017 & ECCV 2018
“Cooperative Learning of Audio and Video Models …”, Korbar, Tran,Torresani, NeurIPS 2018

Obtaining positives and negatives

video 1



video 2

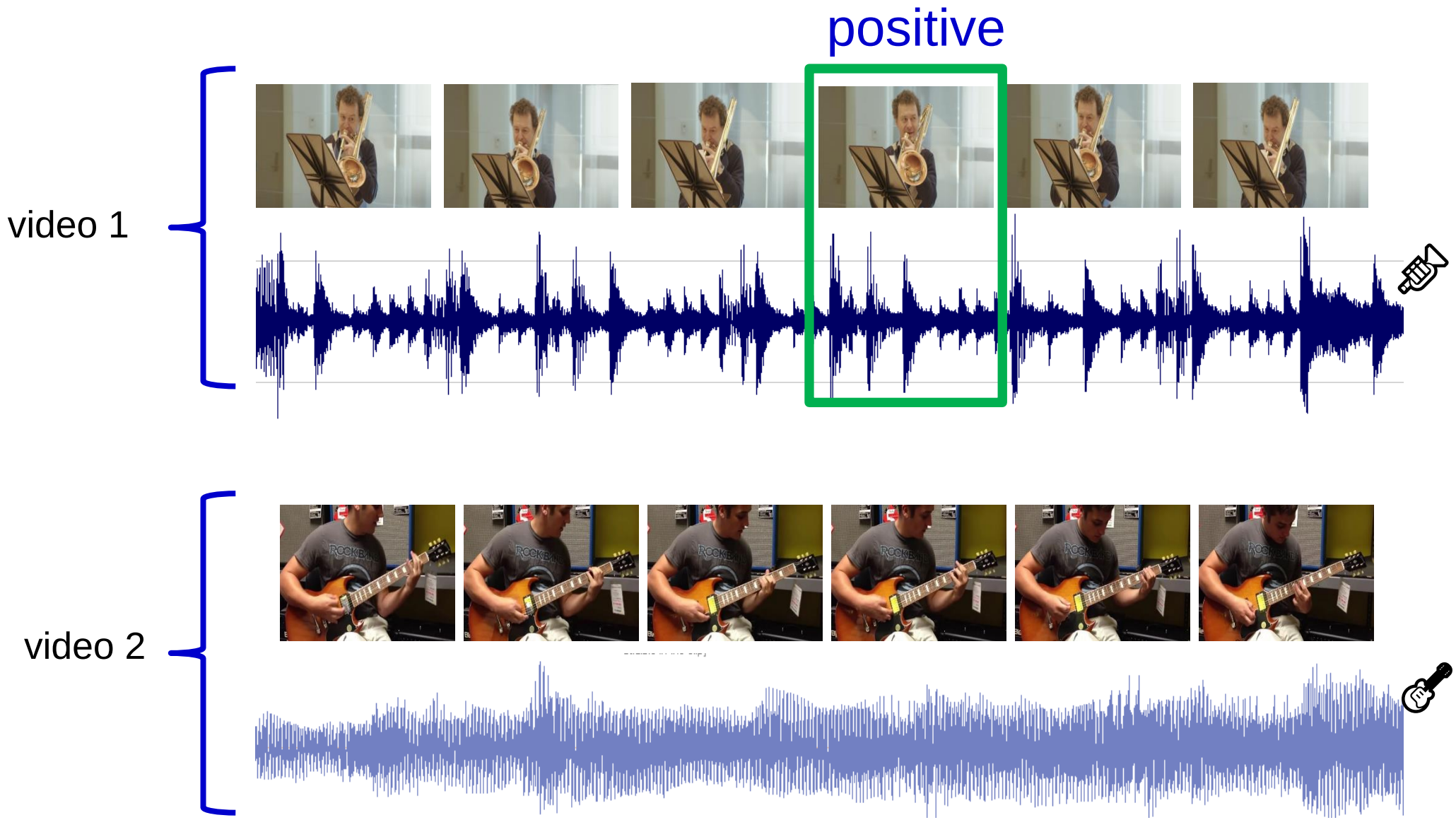


Advantage of Self-Supervised: no human annotation required!

Obtaining positives and negatives

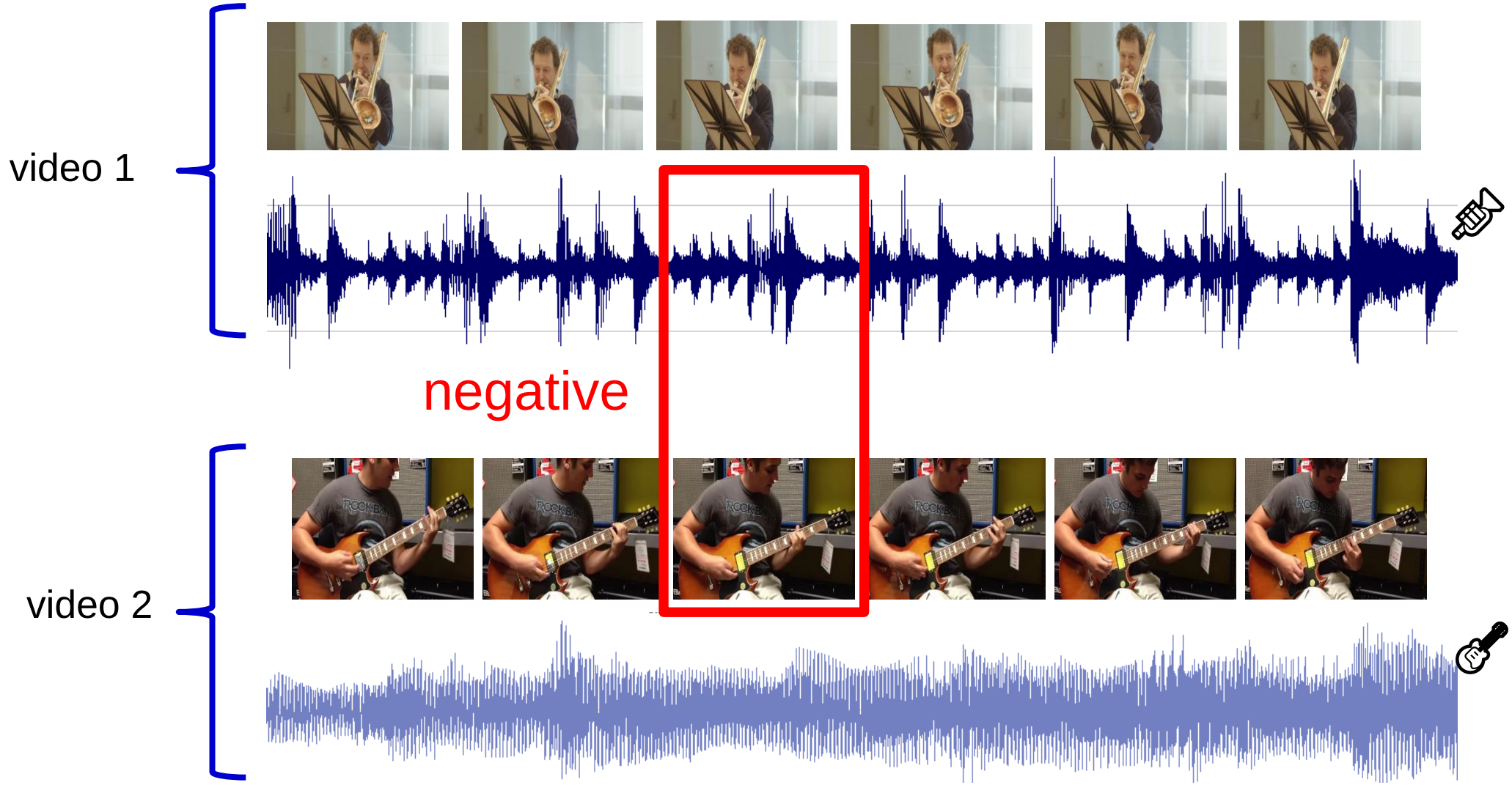
Advantage of Self-Supervised: no human annotation required!

Obtaining positives and negatives



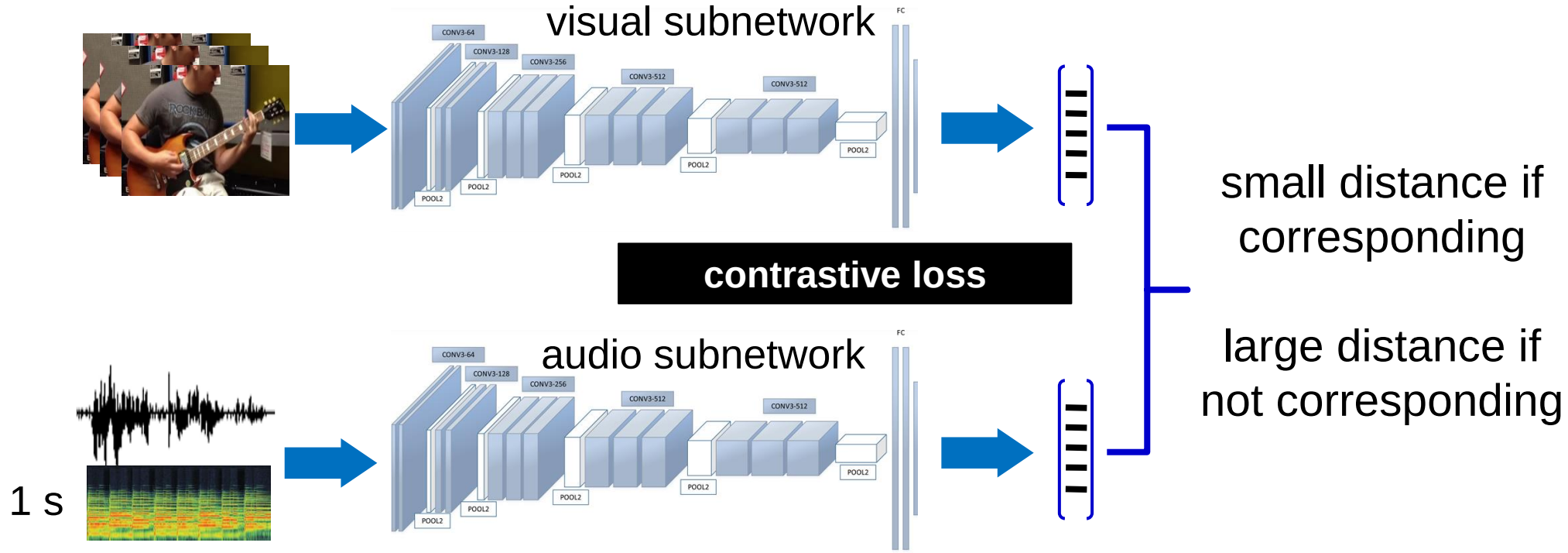
Advantage of Self-Supervised: no human annotation required!

Obtaining positives and negatives



Advantage of Self-Supervised: no human annotation required!

Learning from Correspondence

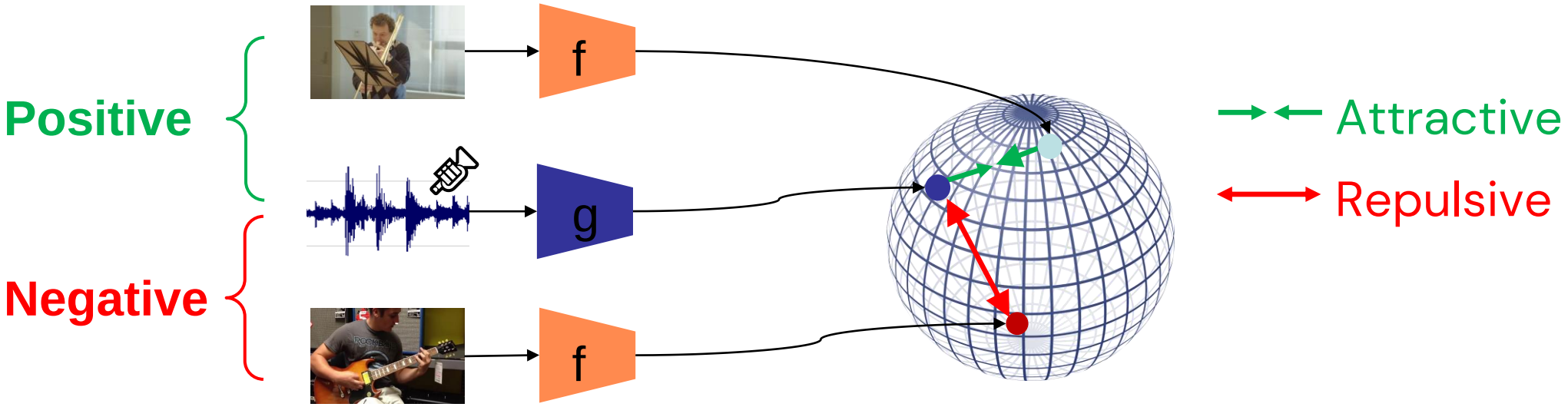


The network is trained from scratch with contrastive loss to:

- Minimise distance between positive pairs (same video clip)
- Maximise distance between negative pairs (different video clips)

on hundreds of hours of video

Multi-Modal Self-supervised Contrastive Learning



Joint audio-video embedding

Multi-Modal Contrastive Objective

Noise Contrastive Estimation (NCE) loss

Video

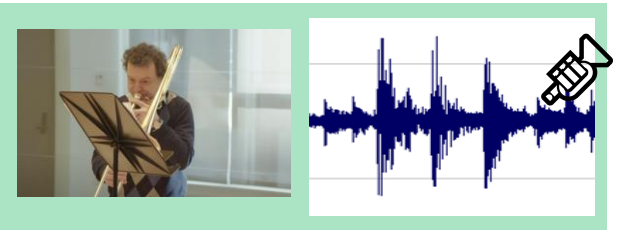
Audio

$$\max_{f,g} \log \left(\frac{\exp(f(x)^\top g(y))}{\exp(f(x)^\top g(y)) + \sum_{(x',y')\in\mathcal{N}} \exp(f(x')^\top g(y'))} \right)$$

Video network

Audio network

Positive Video-Audio pair



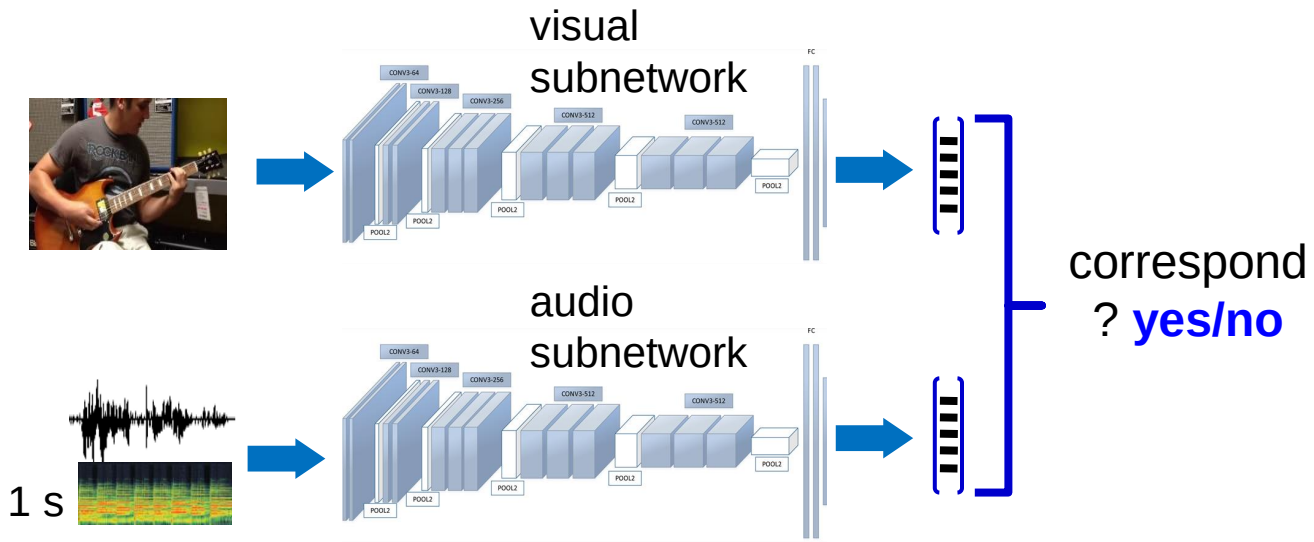
Negative Video-Audio pairs

Softmax over positive and multiple negatives

Use audio and visual features

What can be learnt by watching and listening to videos?

- Good representations
  - Visual features
  - Audio features
- Intra- and cross-modal retrieval
  - Aligned audio and visual embeddings
- “What is making the sound?”
  - Learn to localize objects that sound



“Objects that Sound”, Arandjelović and Zisserman, ICCV 2017 & ECCV 2018

Audio-visual self-supervision using a Noise Contrastive Estimation (NCE) loss

UCF-101 Top-1 Acc % (classifying human actions in video clips)

- Supervised (Kinetics 400): 95.0
- Self-supervised (IG65M): 95.2

Surpasses performance of strong supervision (training with class labels) on Kinetics-400

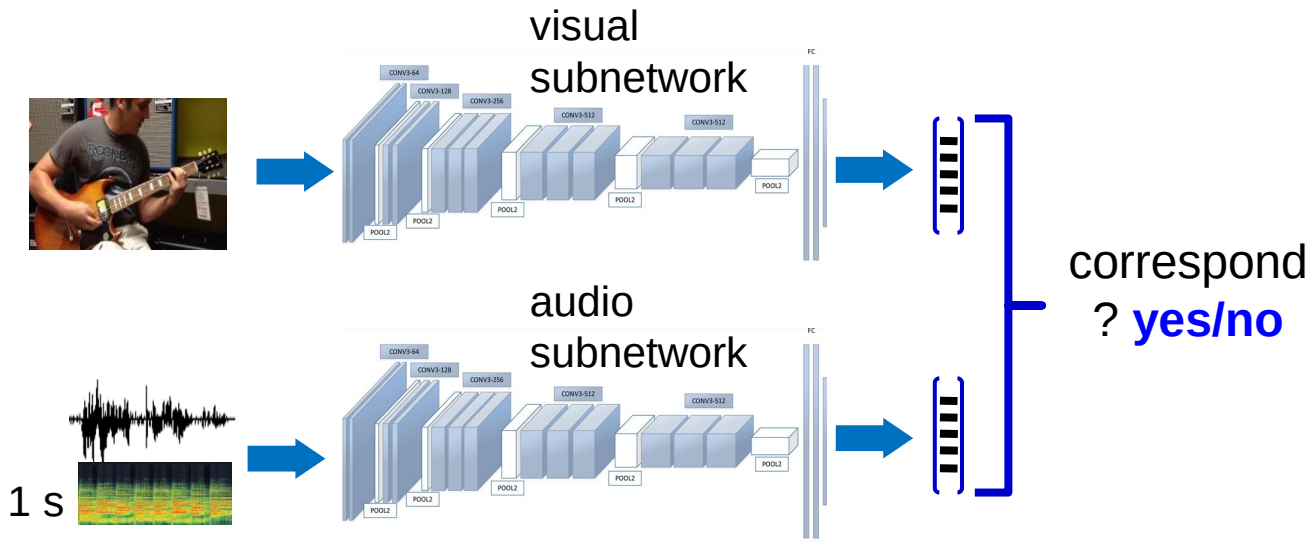
“Multi-modal Self-Supervision from Generalized Data Transformations”,

Mandela Patrick, Yuki M. Asano, Polina Kuznetsova, Ruth Fong, João F. Henriques, Geoffrey Zweig, Andrea Vedaldi, ICCV 2021

Use audio and visual features

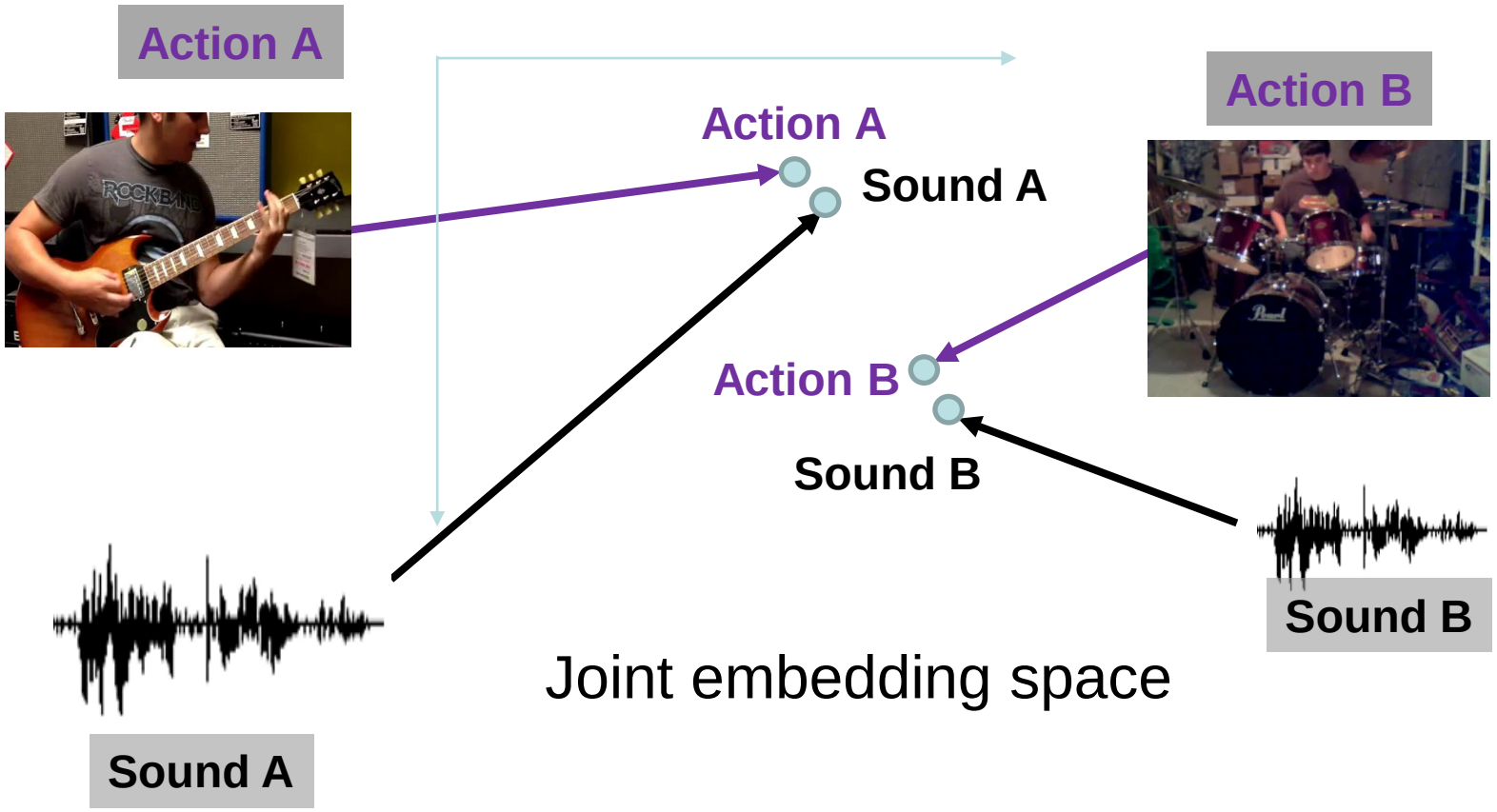
What can be learnt by watching and listening to videos?

- Good representations
  - Visual features
  - Audio features
- Intra- and cross-modal retrieval
  - Aligned audio and visual embeddings
- “What is making the sound?”
  - Learn to localize objects that sound



“Objects that Sound”, Arandjelović and Zisserman, ICCV 2017 & ECCV 2018

Joint audio-video embedding



Query on audio, retrieve video

Search in 200k video clips of AudioSet

Query audio

Query on audio, retrieve video

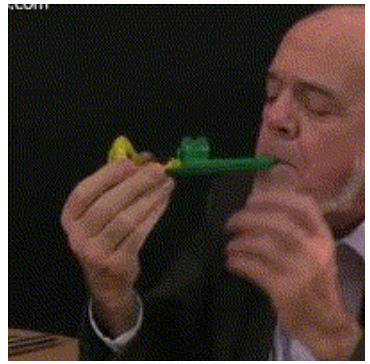
Search in 200k video clips of AudioSet

Query audio

Rank 1

Rank 2

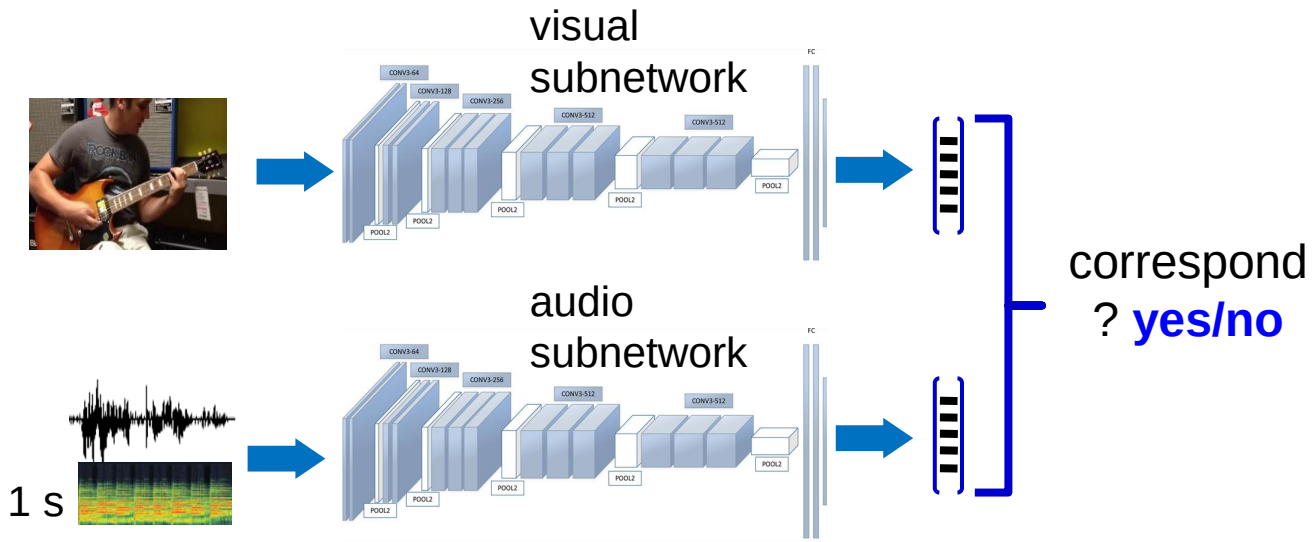
Rank 3



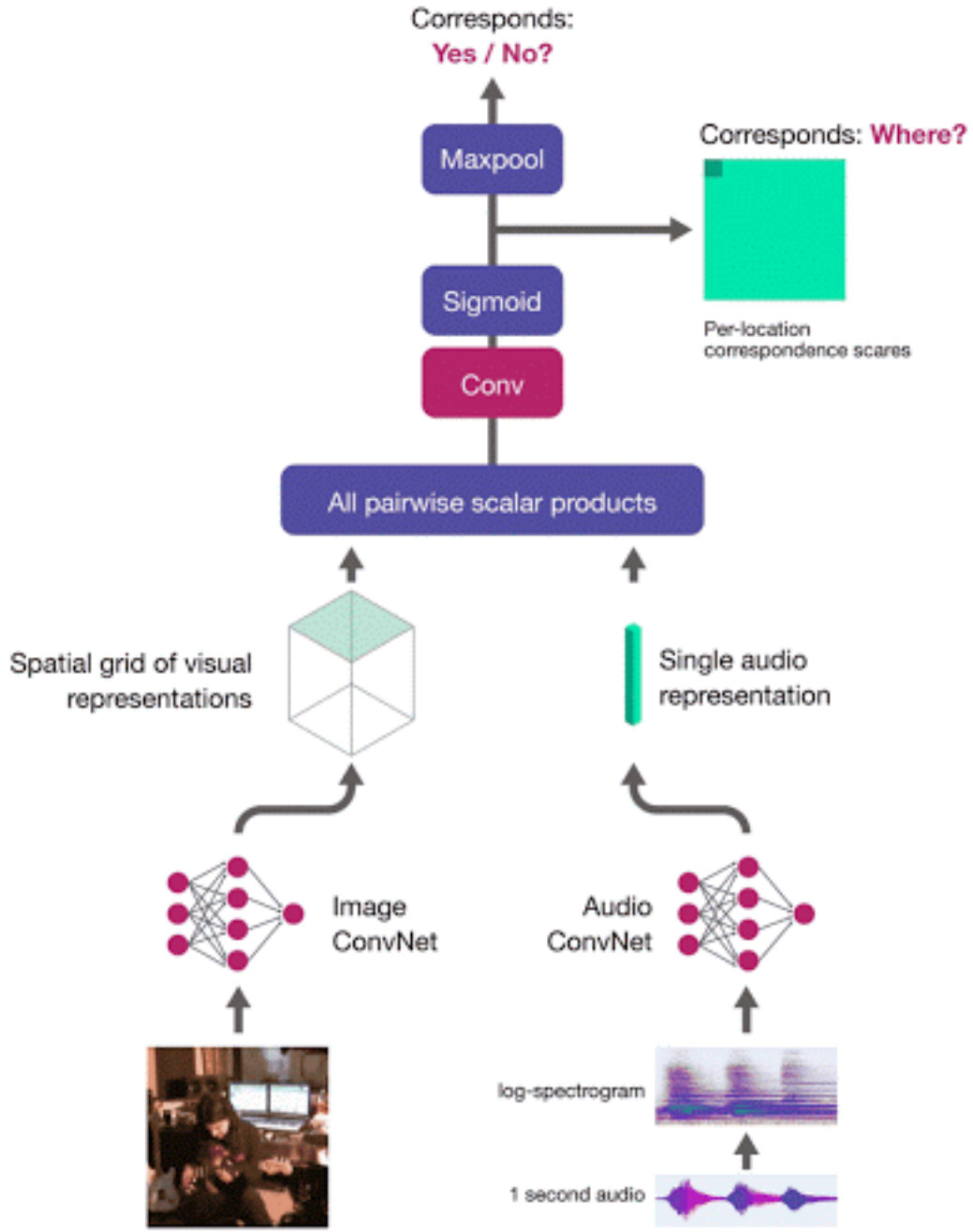
Use audio and visual features

What can be learnt by watching and listening to videos?

- Good representations
  - Visual features
  - Audio features
- Intra- and cross-modal retrieval
  - Aligned audio and visual embeddings
- “What is making the sound?”
  - Learn to localize objects that sound



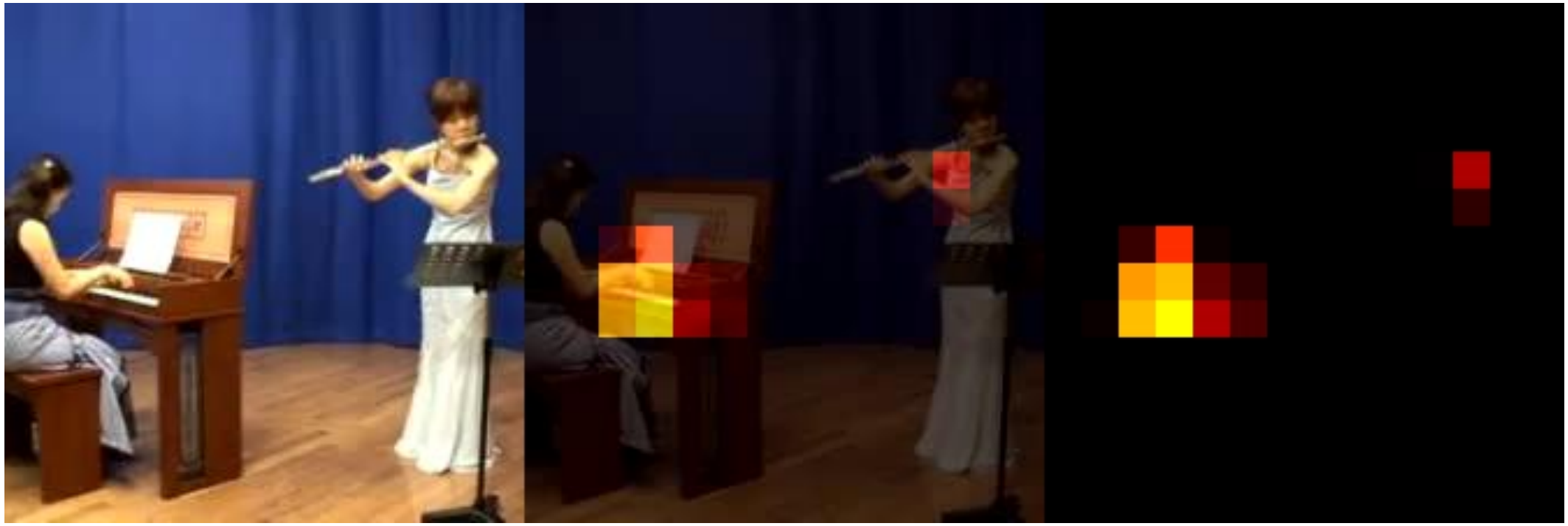
“Objects that Sound”, Arandjelović and Zisserman, ICCV 2017 & ECCV 2018



audio attention vector

Objects that Sound: object localization

Input: audio and video frame



frame frame+heatmap heatmap

- Frame by frame
- No motion information
- No memory
- No smoothing

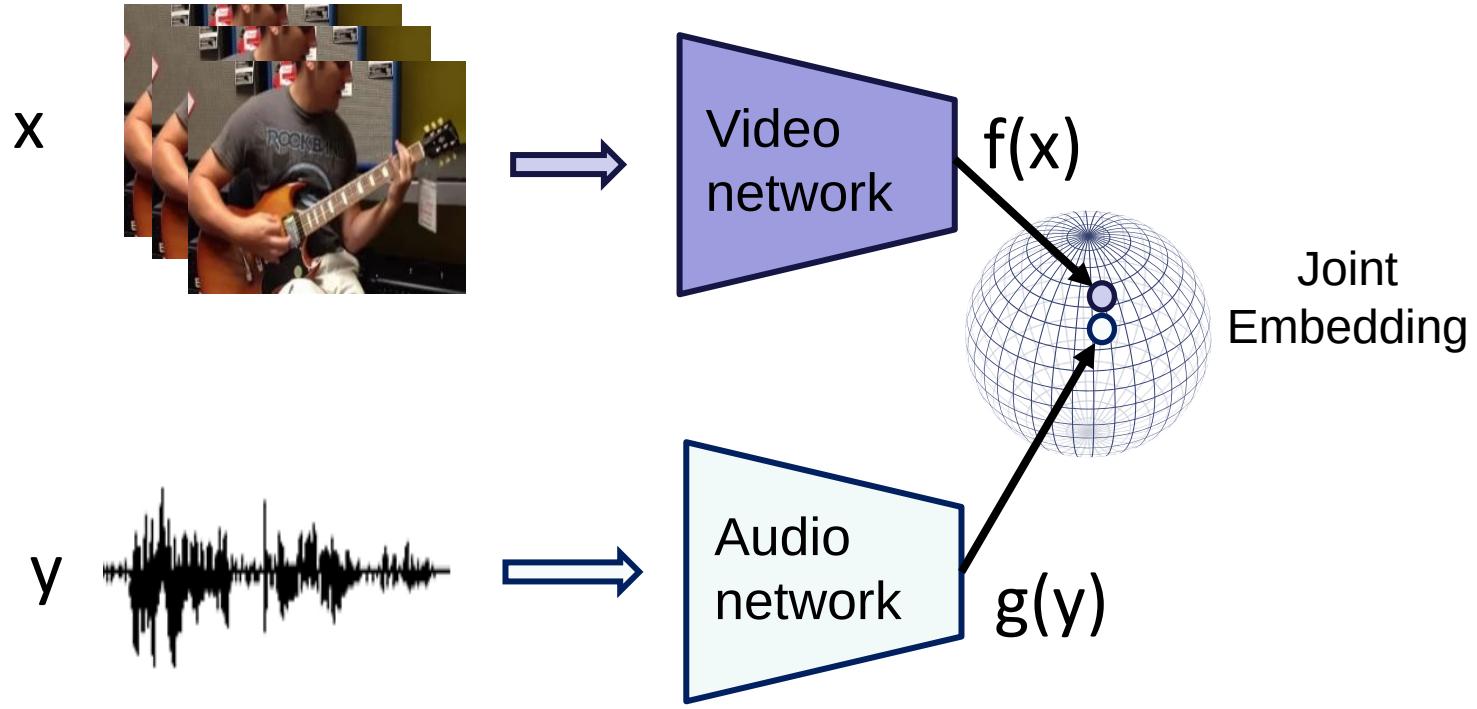
“Objects that Sound”, Arandjelović and Zisserman, ICCV 2017 & ECCV 2018

Part I Summary: learn a joint audio-video embedding

Architecture: Dual Encoder

Separate networks for video and audio encoding

Score similarity, e.g. by $f(x)^\top g(y)$



Contrastive training using a Noise Contrastive Estimation (NCE) loss

Natural Question Break

Part II

Learning Object-Centric Representations using Visual Self-Supervision

Object-centric representations

Learn a structured representation where moving objects are represented explicitly:

- Object discovery
- Object segmentation
- Object tracking
- Object effects

Infant development:

- 3 months+ Object permanence
- Object recognition
- 4-6 months: Cause and effect
- 8-9 months: hidden objects

Self-Supervised Video Understanding Tasks

1. Video moving object discovery and segmentation

2. Object correlated effects

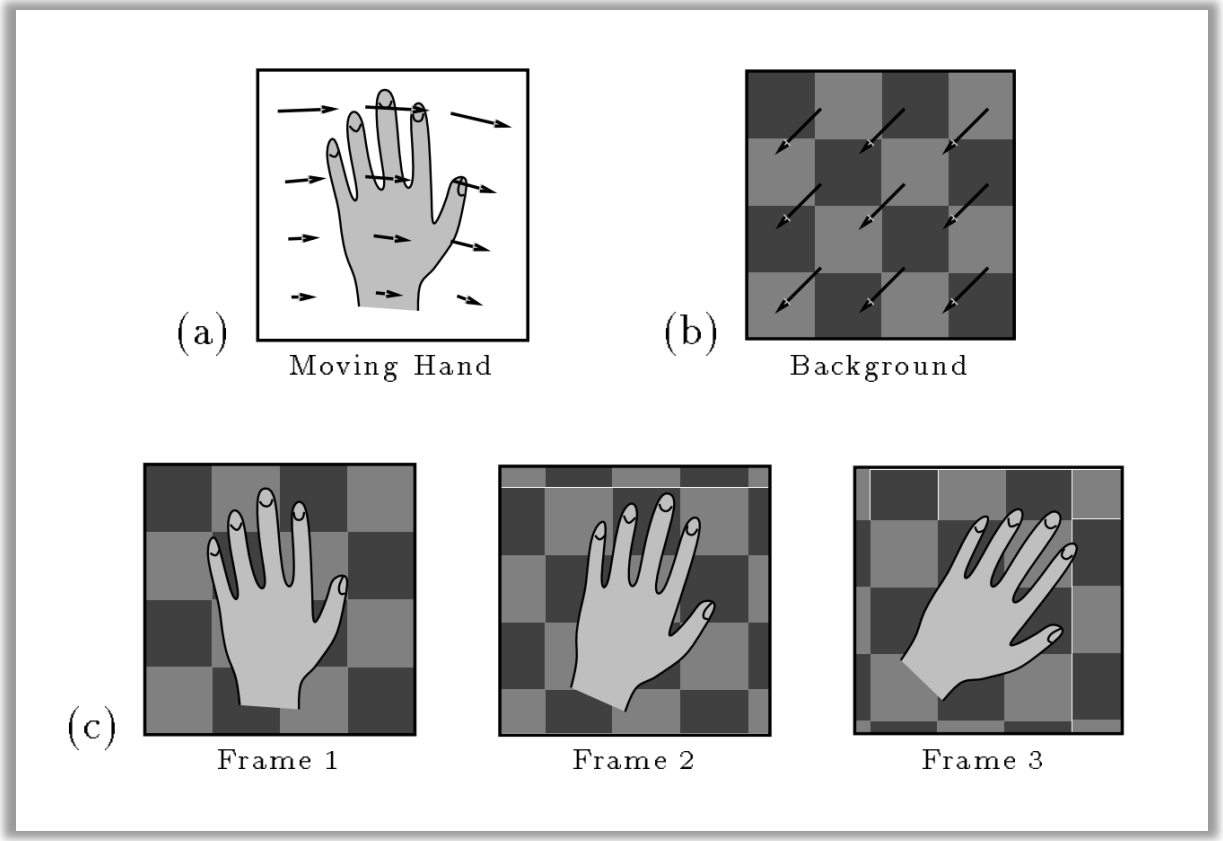
Both tasks build on a layered video representation …

Representing a Video using Layers

"Representing Moving Images with Layers"

John Y. A. Wang and Edward H. Adelson

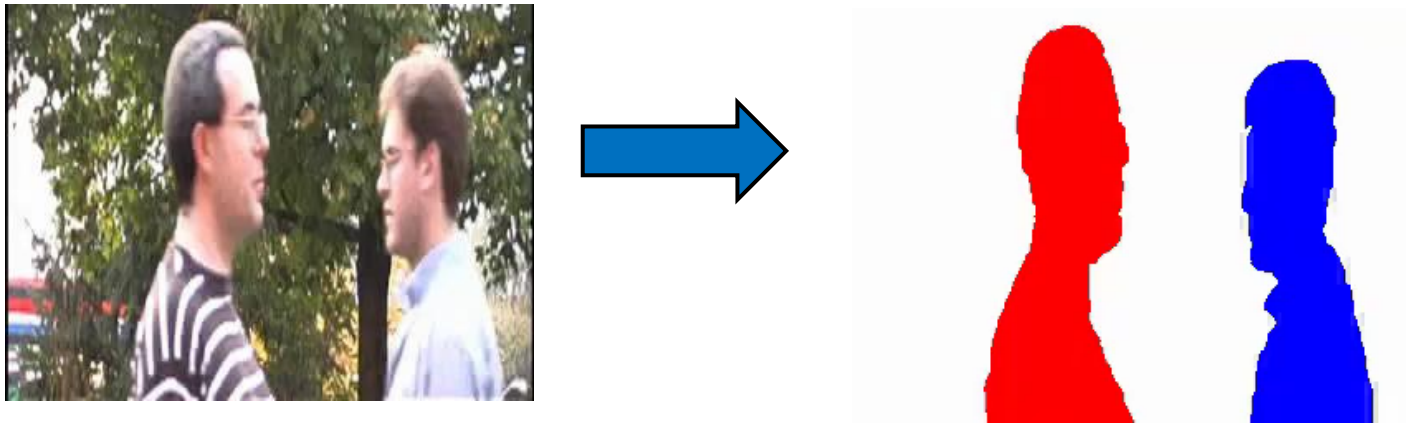
Trans. on Image Processing, 1994



Further applications:

- Layer extraction from multiple images containing reflections and transparency
  Richard Szeliski, Shai Avidan, P. Anandan, CVPR 2000

Representing a Video using Layers



“Learning flexible sprites in video layers”, Jojic & Frey, CVPR 2001

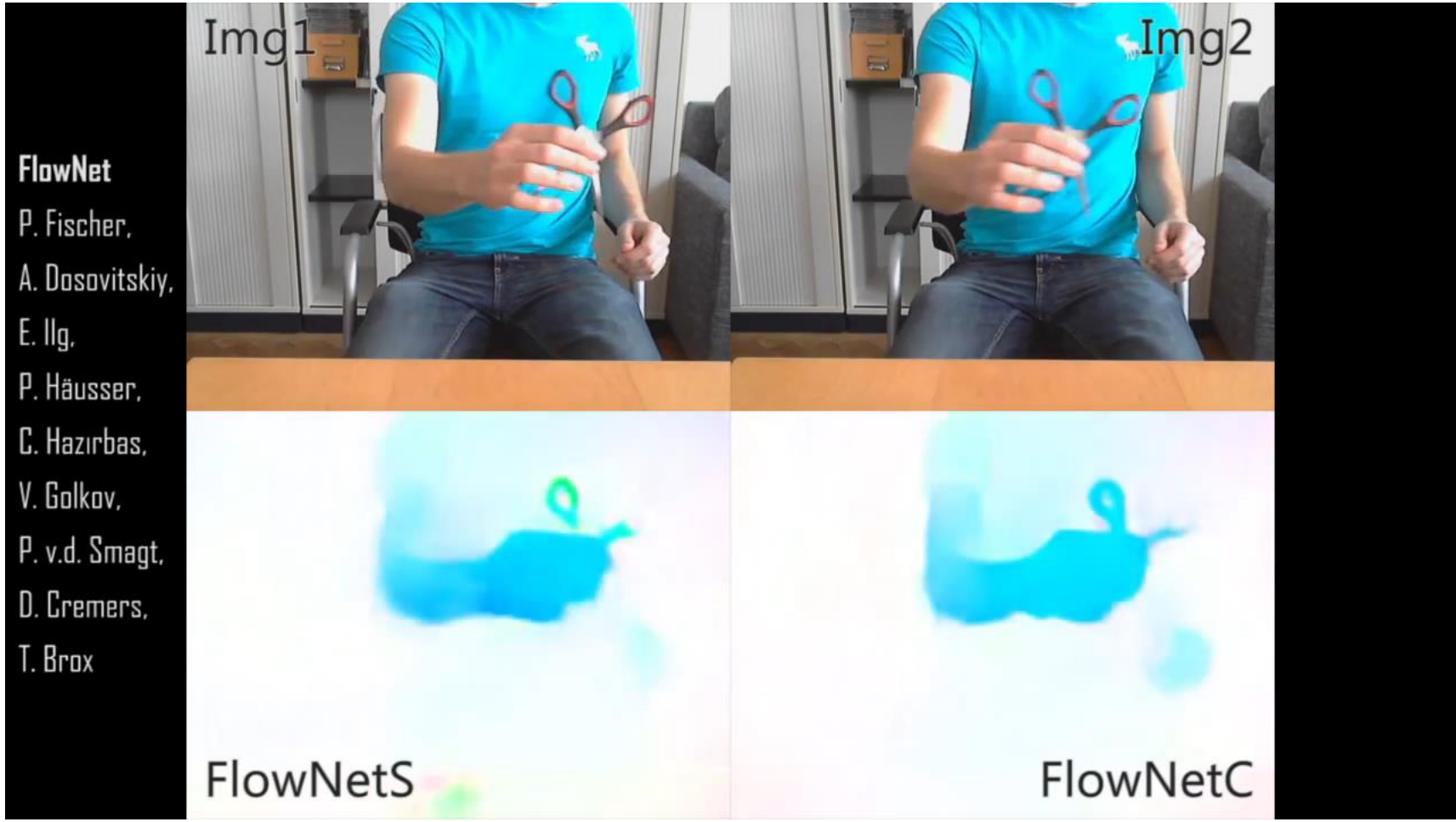
Self-supervised learning for video

Video, beyond images, naturally …

- extends and develops sequentially in time,
- has multiple modalities (audio stream),
- has motion (optical flow stream)

Representing motion using optical flow

- Throws away “nuisance factors” like appearance of clothes and skin
- Helps with foreground/background segmentation



Self-Supervised Video Object Segmentation by Motion Grouping

Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, Weidi Xie

ICCV 2021

Objective: video object discovery and segmentation

- Given a video sequence, the task is to segment the primary object in the video, usually the most ”salient” one.

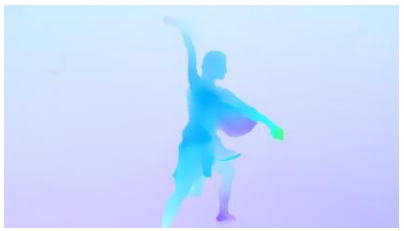






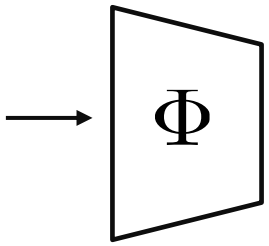

Key idea – represent motion (flow) using layers, rather than RGB

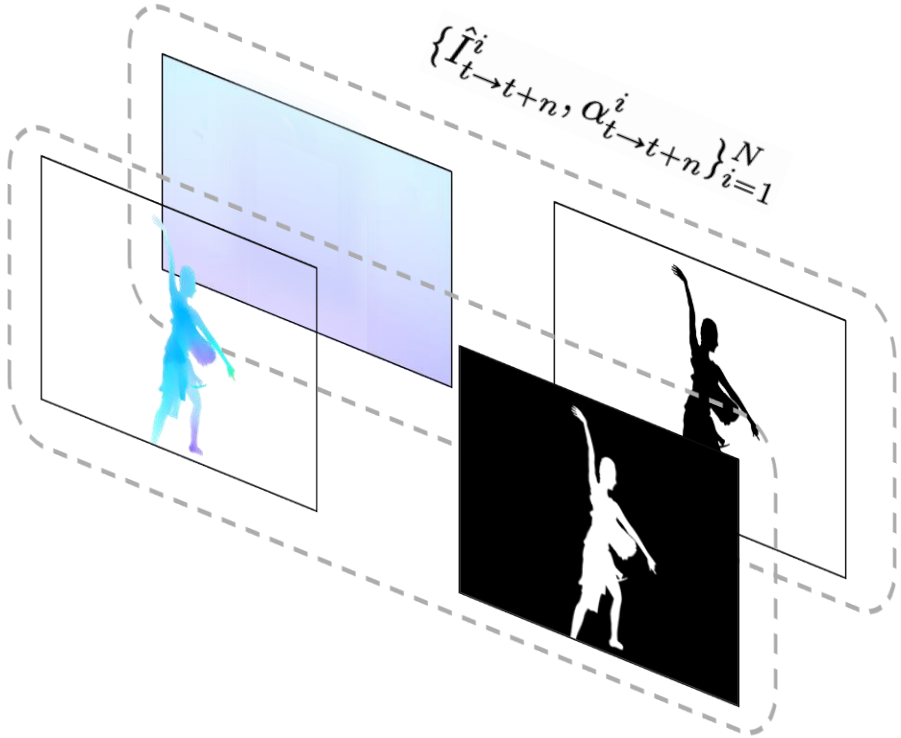
- Autoencoder with layer representation, decomposing the input flow to layers.
- Each layer only needs to represent homogeneous motion field, discontinuities appear between layers.
- Alpha channels (for linear composition) are the segmentation mask.



Optical Flow

$I_{t \to t+n}$





Reconstruction

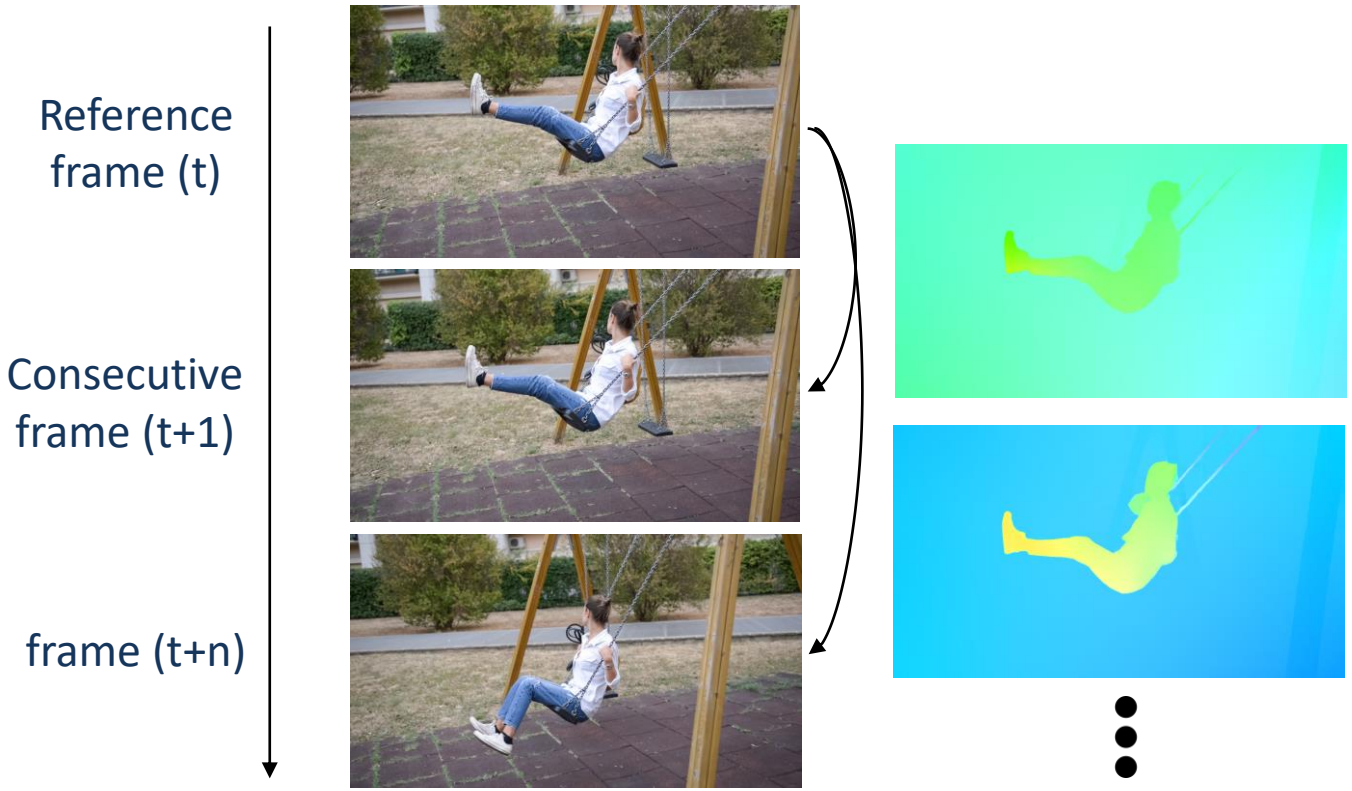
$$\hat{I}_{t \to t+n} = \sum_{i=1}^{N} \alpha^i_{t \to t+n} \hat{I}^i_{t \to t+n}$$

$$\mathcal{L}_{recon} = \frac{1}{\Omega} \sum_{p \in \Omega} |I_{t \to t+n}(p) - \hat{I}_{t \to t+n}(p)|^2$$

J. Y. A Wang and E. H. Adelson. “Representing Moving Images with Layers”. In: TIP, 1994.

Temporal Consistency

- Challenge: objects that undergo similar motion to camera, will not be clearly visible in flow field.
- Compute flows with different frame gaps, and force consistency on the inferred alpha channels.

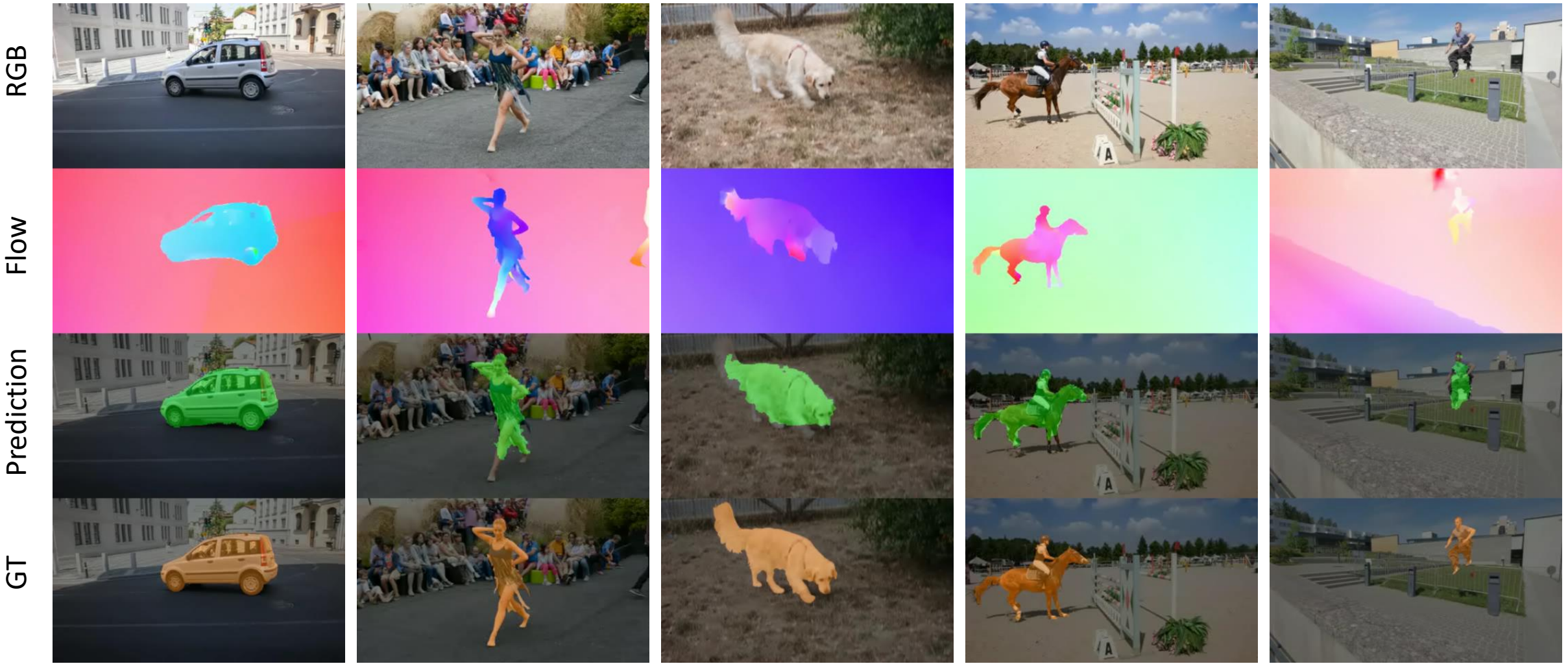


$$\mathcal{L}_{cons} = \frac{1}{\Omega}\min\Big(\sum_{p\in\Omega} |\alpha^1_{t\to t+n_1}(p) - \alpha^1_{t\to t+n_2}(p)|^2, \sum_{p\in\Omega} |\alpha^1_{t\to t+n_1}(p) - \alpha^0_{t\to t+n_2}(p)|^2\Big)$$

Consistency Loss (L2 between masks). Layers numbering is arbitrary, so loss is **permutation invariant**.

Qualitative Results – DAVIS2016

All predictions are **only** based on optical flow, RGB is used here for demonstration purpose.

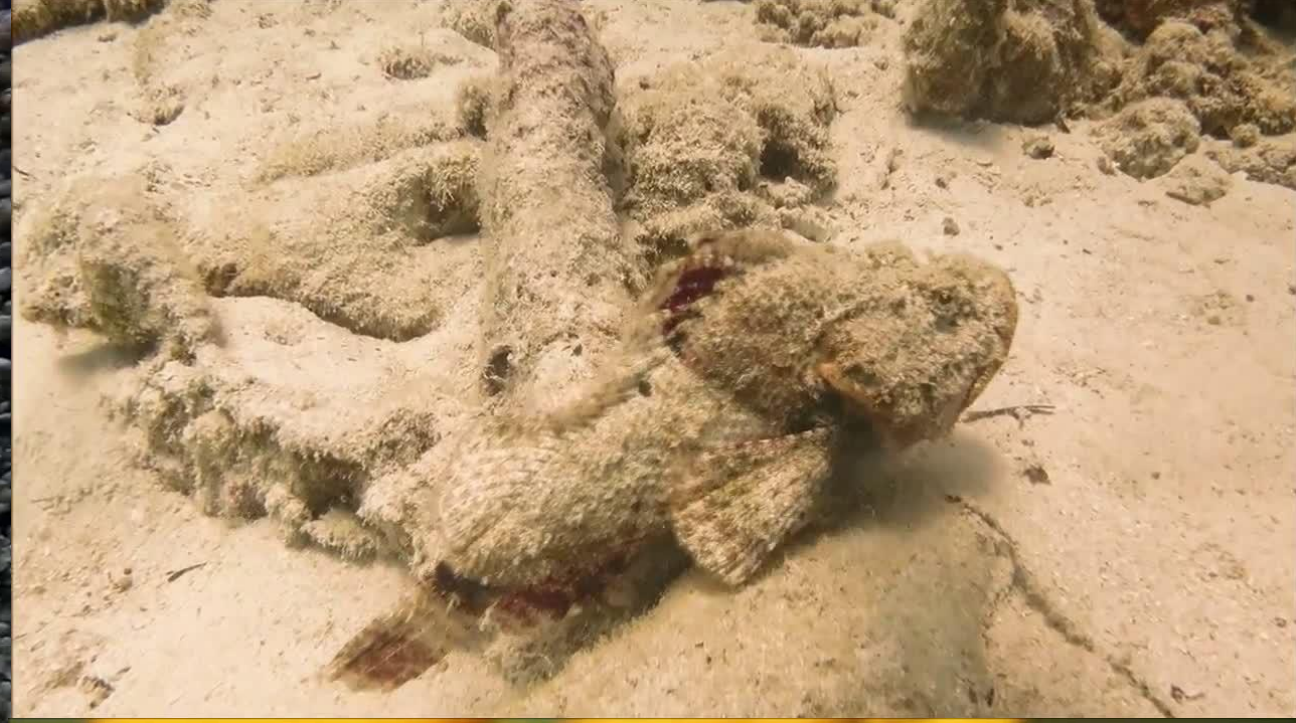


141 Videos, 37K frames
The largest video dataset for camouflaged animals discovery

Annotations
We provide both bounding box annotations and motion type labels

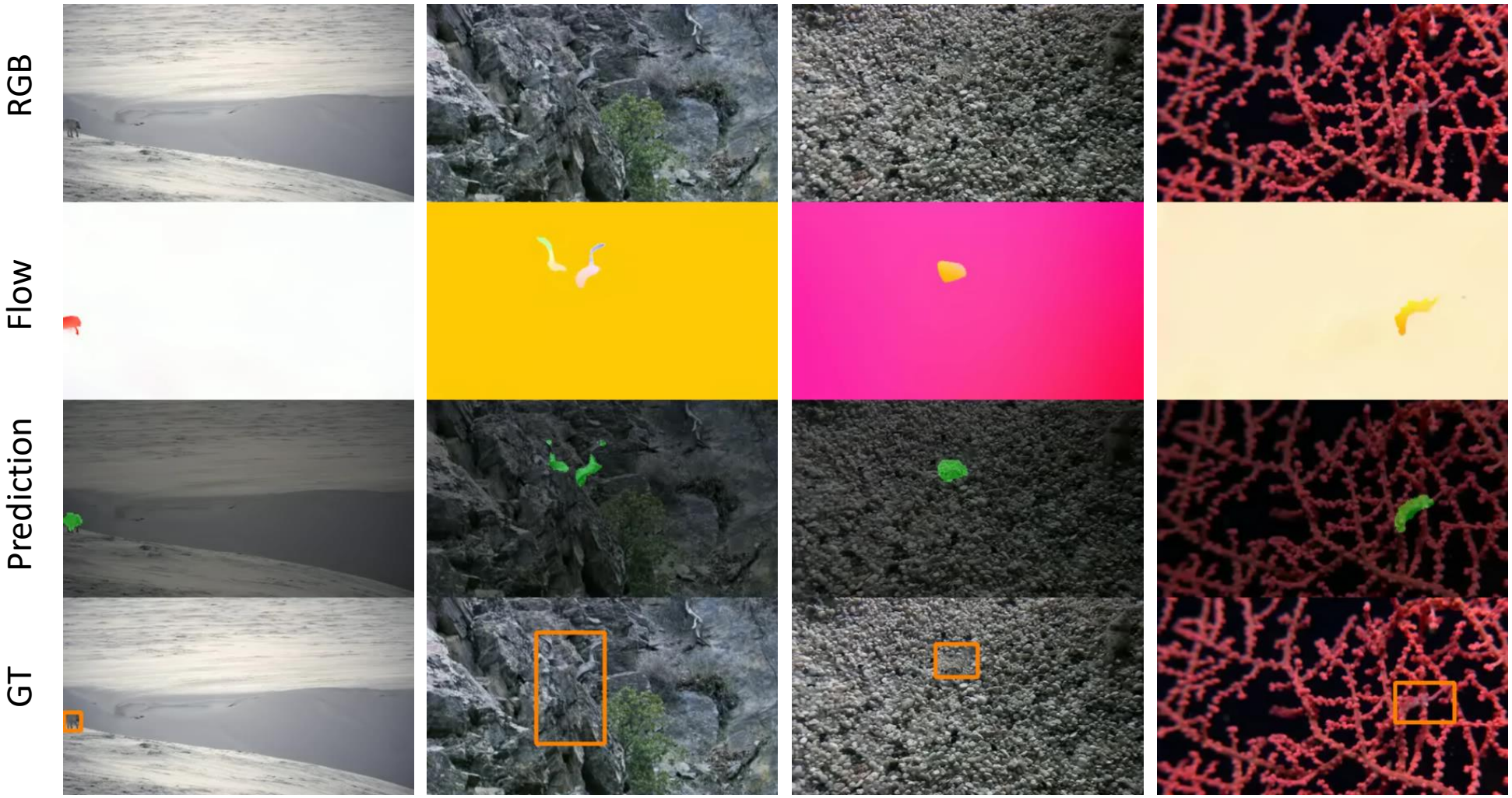
Highly Challenging
67 categories of animals which have mastered camouflage

Hala Lamdouar, Charig Yang, Weidi Xie, Andrew Zisserman
ACCV 2020



Qualitative Results - MoCA

All predictions are **only** based on optical flow, RGB is used here for demonstration purpose.



Omnimatte: Associating Objects and their Effects in Video

Erika Lu, Forrester Cole, Tali Dekel, Andrew Zisserman,
William T. Freeman, Michael Rubinstein

CVPR 2021

Motivation

- An object’s effects extends over the scene: shadows, reflections …

Motivation

- An object’s effects extends over the scene: shadows, reflections …

Objective

- Given probe masks for each object (* generated automatically using Mask R-CNN)
- Generate layers that include both the object and its effects

Input Video

Input Mask 1*

Input Mask 2*

Output Background Layer

Output Layer 1

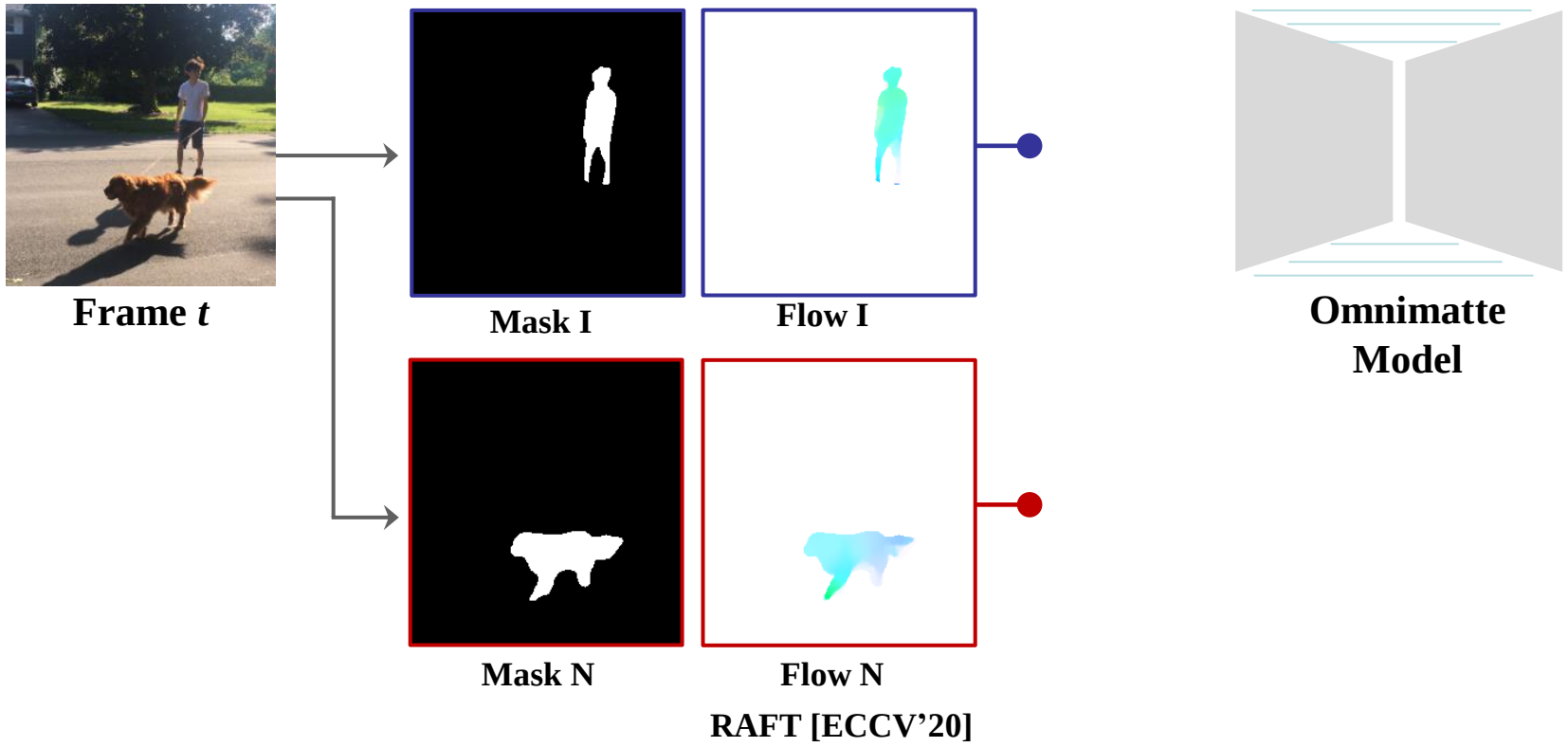
Output Layer 2

“Omnimatte”

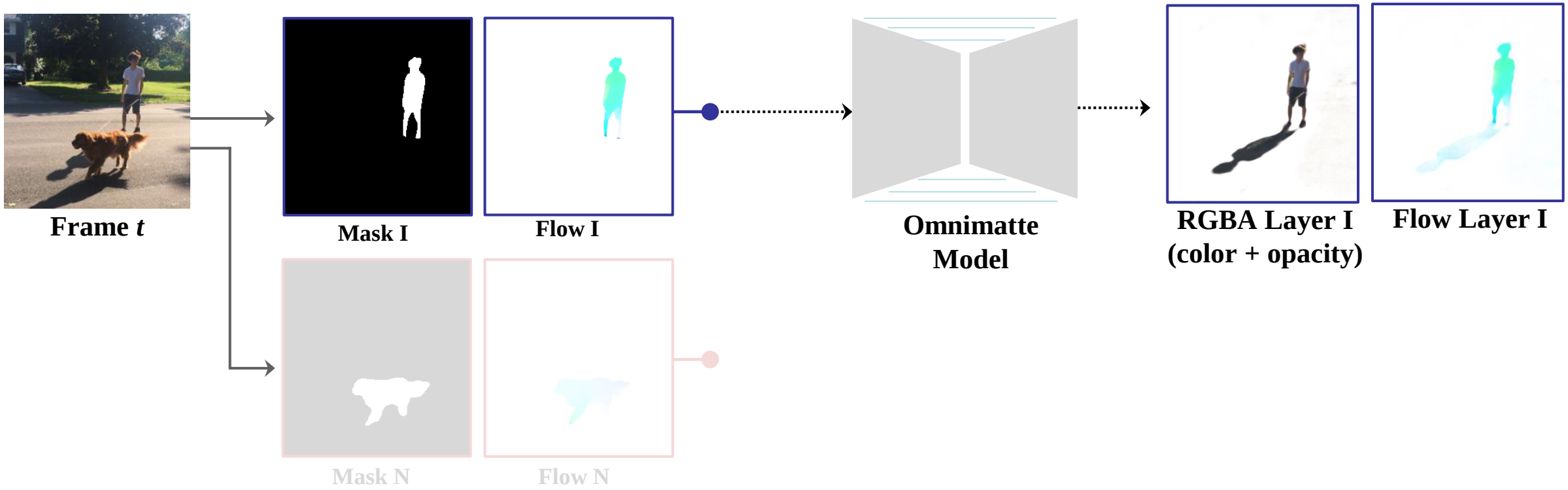
A matte containing an object of interest and all its correlated effects (e.g. shadows)

RGBA

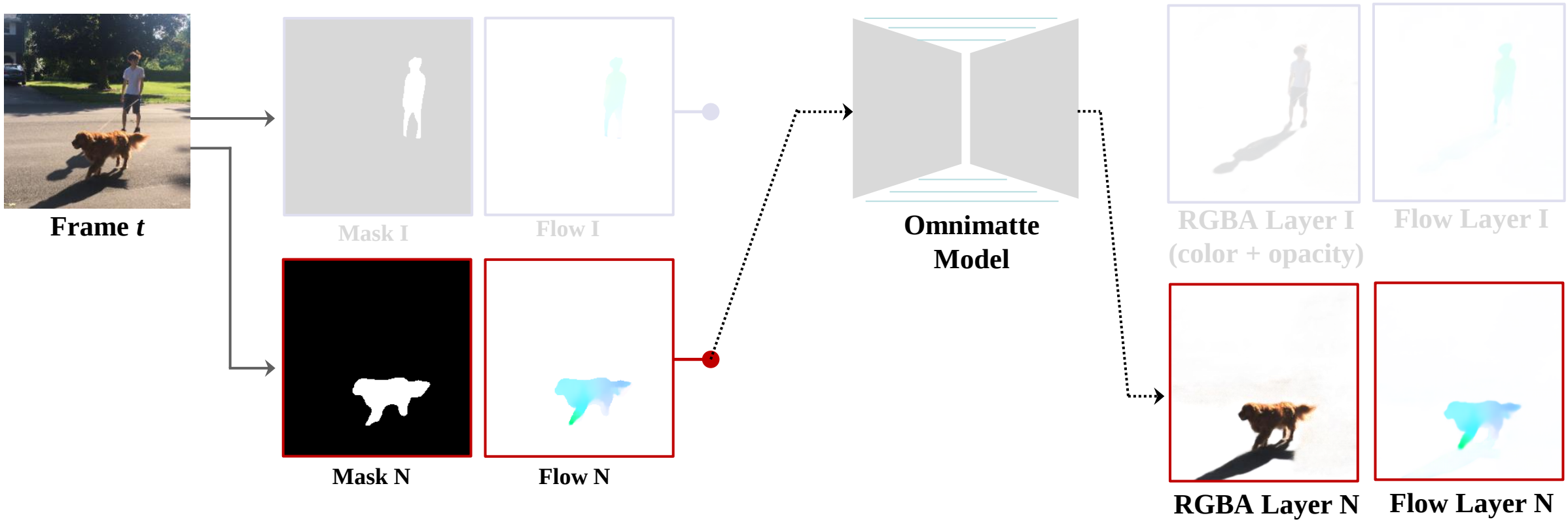
Omnimatte Method



Omnimatte Method



Omnimatte Method



Omnimatte Method

Fixed Noise Input

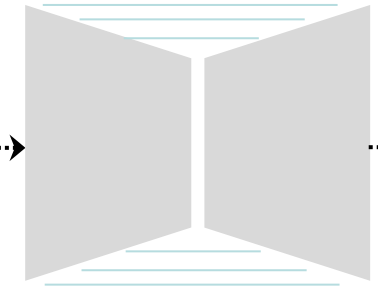
Frame *t*

Mask I

Flow I

Mask N

Flow N



Omnimatte Model

Background Layer

RGBA Layer I (color + opacity)

Flow Layer I

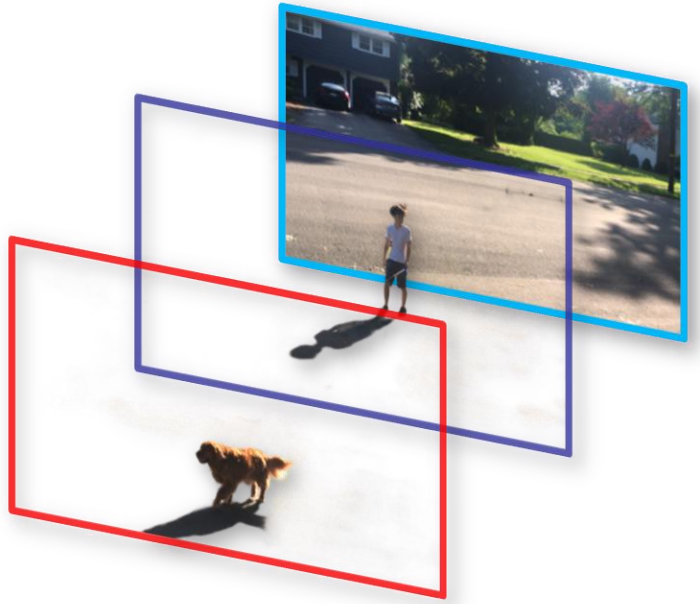
RGBA Layer N

Flow Layer N

Losses – Reconstruction

$$\underbrace{E_{\text{recon}}(\mathcal{L}_t, I_t)}$$

Reconstruction loss



Back-to-front compositing ⬇

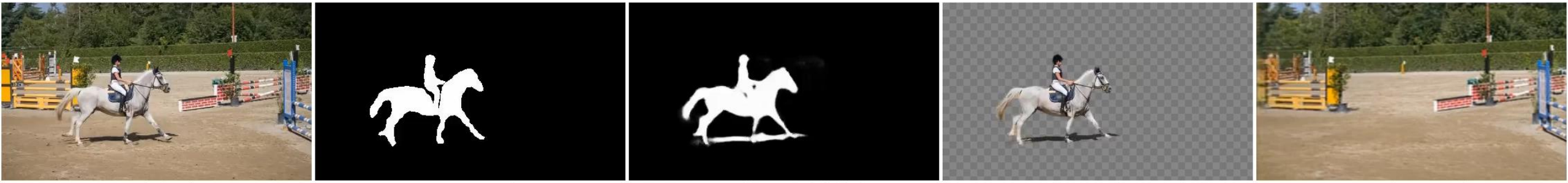
Reconstruction (Frame t)

Omnimatte Results

DAVIS 2017 dataset.
Masks generated using
STM [ICCV'19]

Original | Input mask | Omnimatte (alpha) | Omnimatte (RGBA) | Background

Original | Input mask | Omnimatte (alpha) | Omnimatte (RGBA) | Background



Original | Input mask | Omnimatte (alpha) | Omnimatte (RGBA) | Background

Omnimatte Reflections

Our Input

Our Result

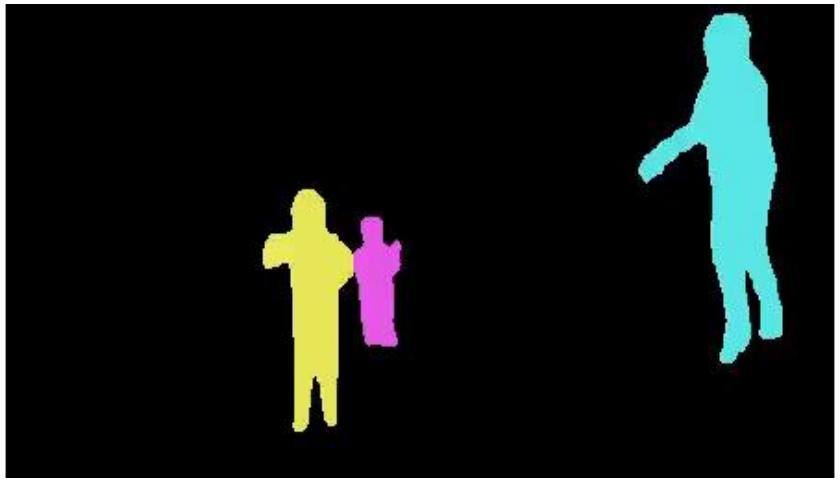
Our Input

Our Result

Omnimatte Deformations

Our omnimatte layers visualized

Input video



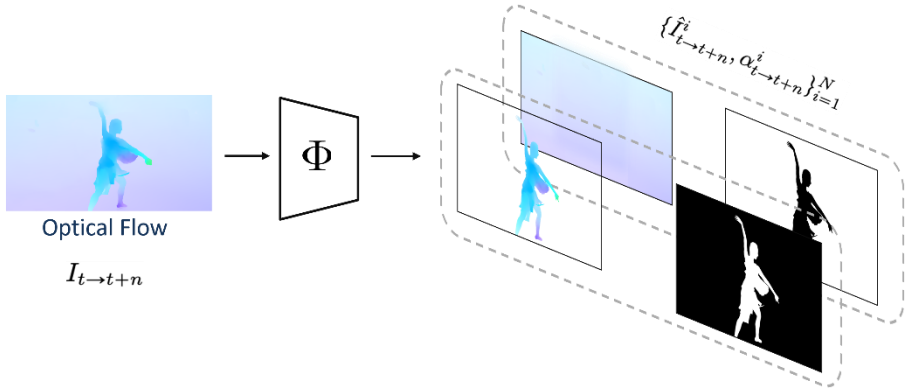
Input segments

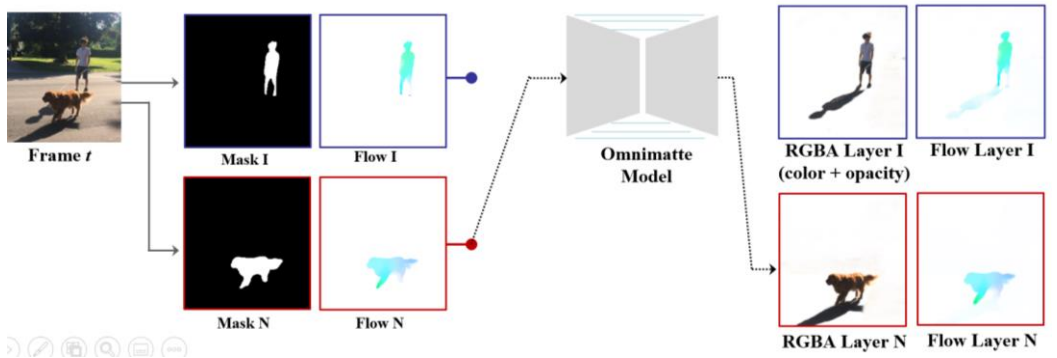
Omnimatte layer visualization

Object removal result – only render one layer and background

Part II Summary

- Learn a layered video object-centric representation
- Self-supervised video object discovery and segmentation
- New self-supervised task: Ominimatte – segment an object and its effects
  - See videos at https://omnimatte.github.io/





Natural Question Break

Part III

Learning Video Representations using Weak Supervision from Text

The benefits of language for visual descriptions …

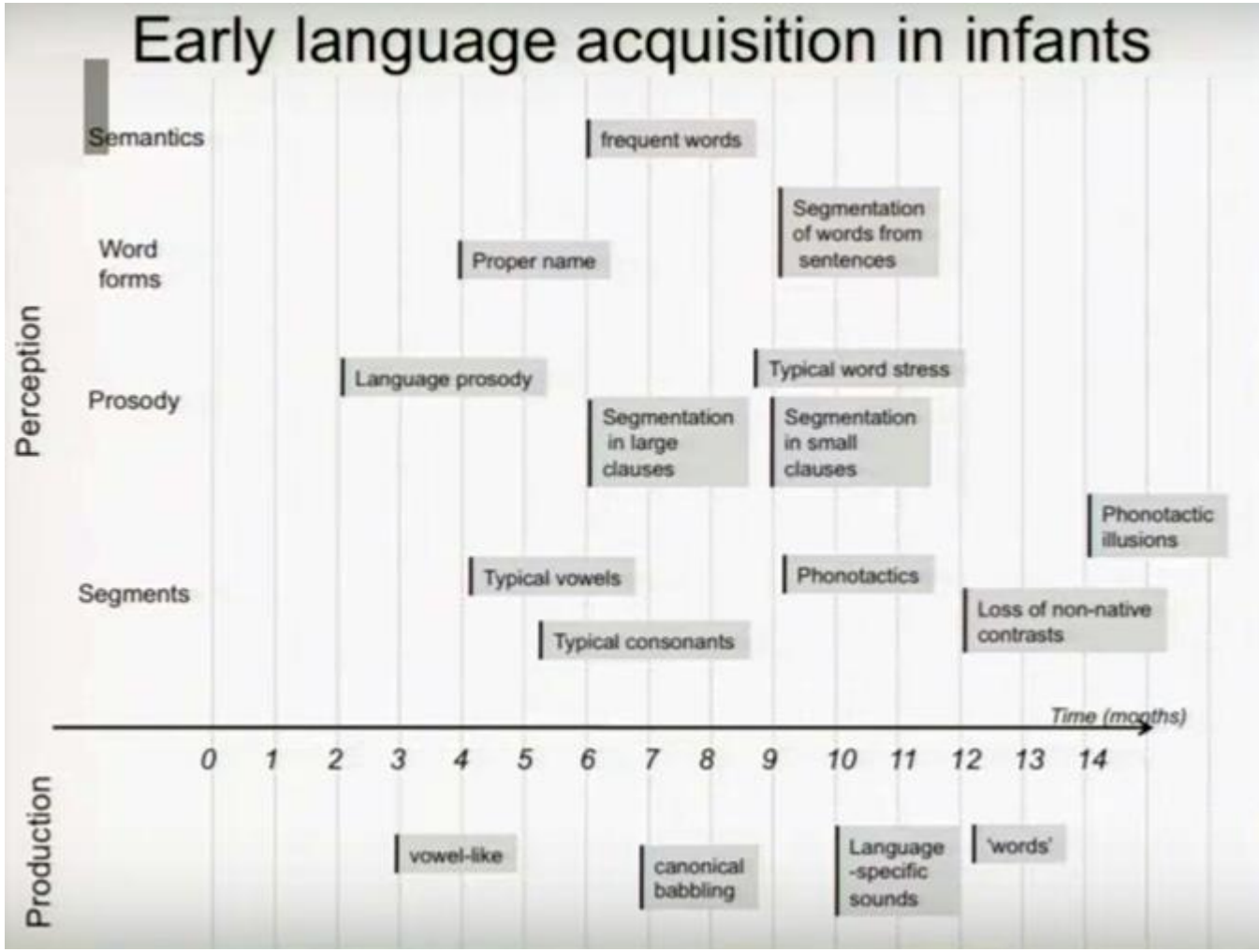
- Gives names to things: objects and actions
- It is compositional (so infinite possibilities for sentences)
- It can be used to communicate with a machine (both ways)

“The child looks and recognizes before it can speak”

John Berger, Ways of Seeing

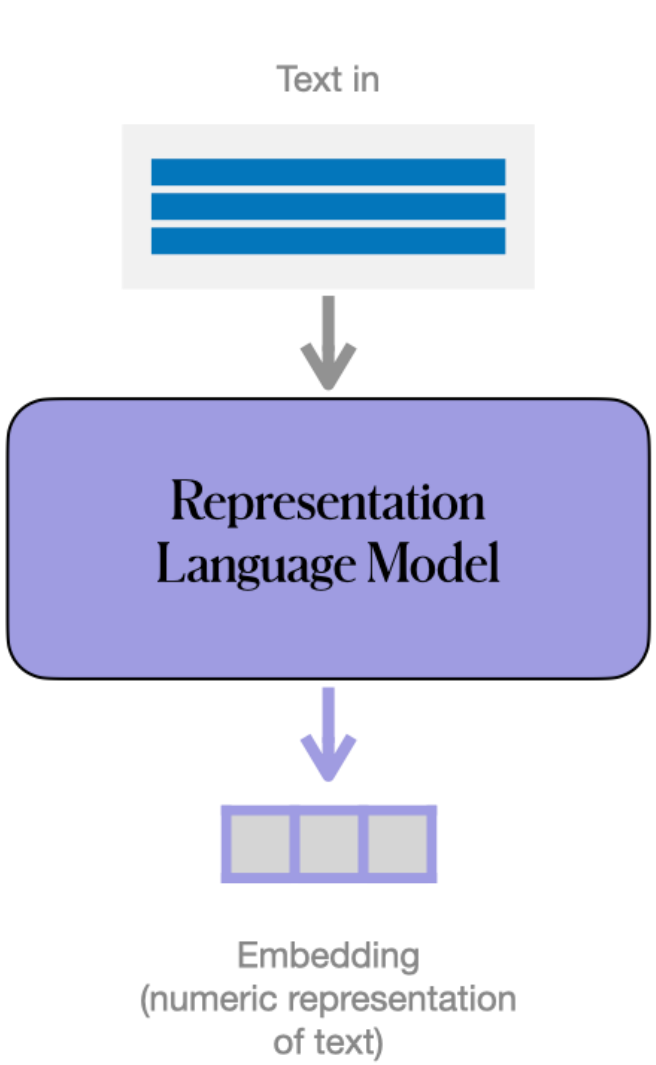
Infant development language skills:

- Understanding (perception) begins earlier than production (talking)
- 10-12 months:
  - First words
  - Gestures for communicating

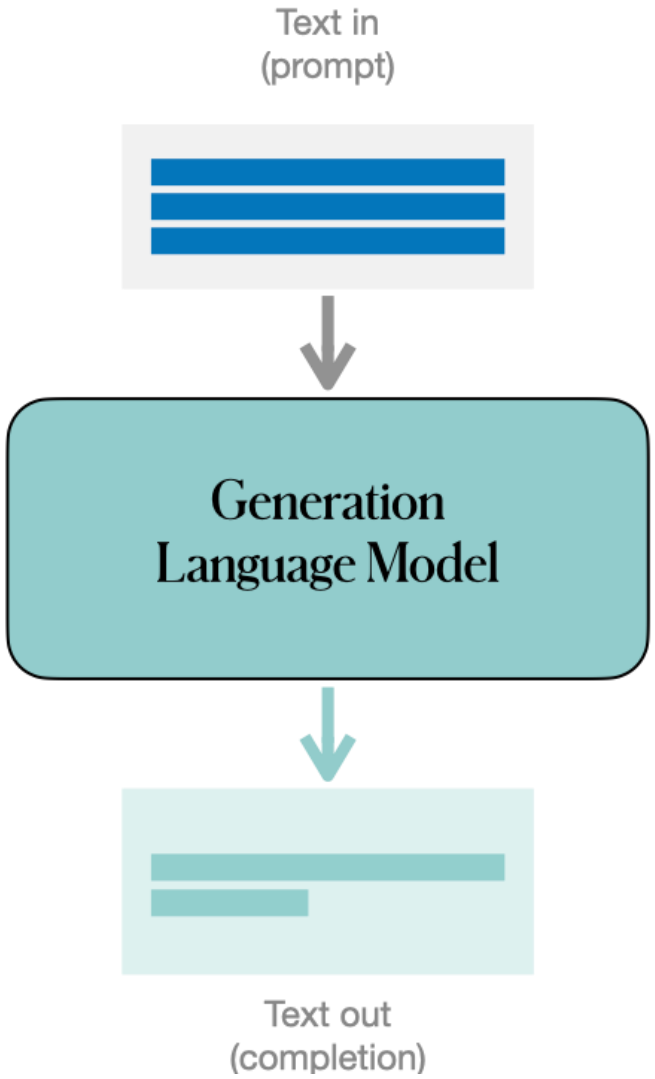


Emmanuel Dupoux 2019

Two types of language models



e.g. BERT



e.g. GPT

Slide credit: Jay Alammar https://docs.cohere.ai/intro-to-llms/

Representation Language Model

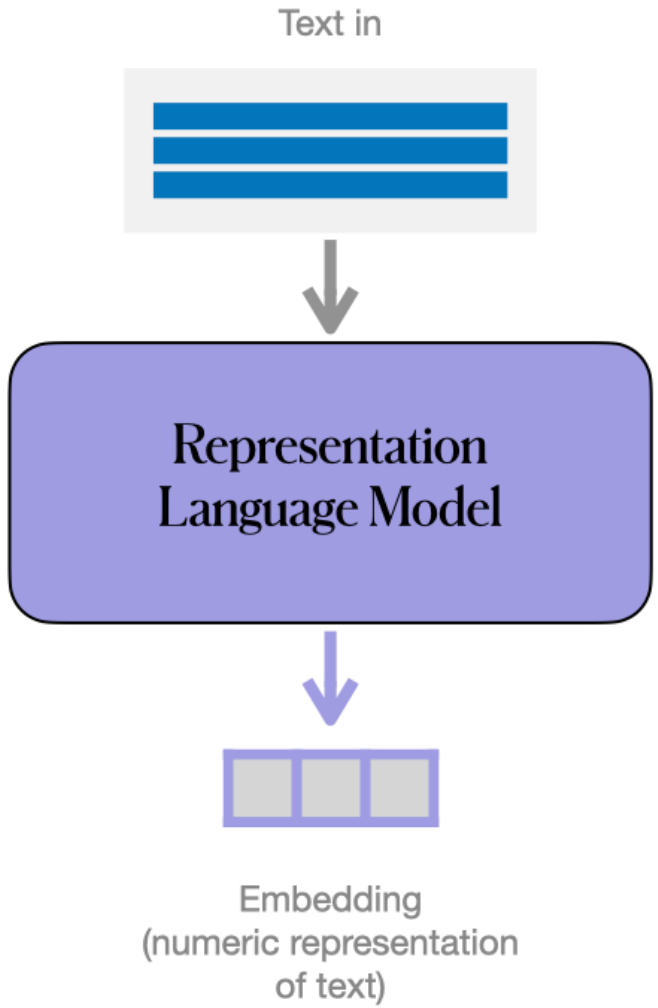
Used for learning a sentence representation

Bidirectional, non-causal

Architecture: Transformer encoder,

e.g. Bidirectional Encoder Representations from Transformers (BERT)

Training: Masked language modelling



e.g. BERT

the chef cooked the meal

Generative Language Model

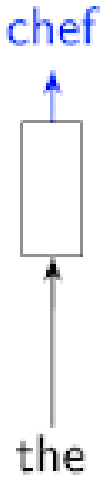
Used for generating text

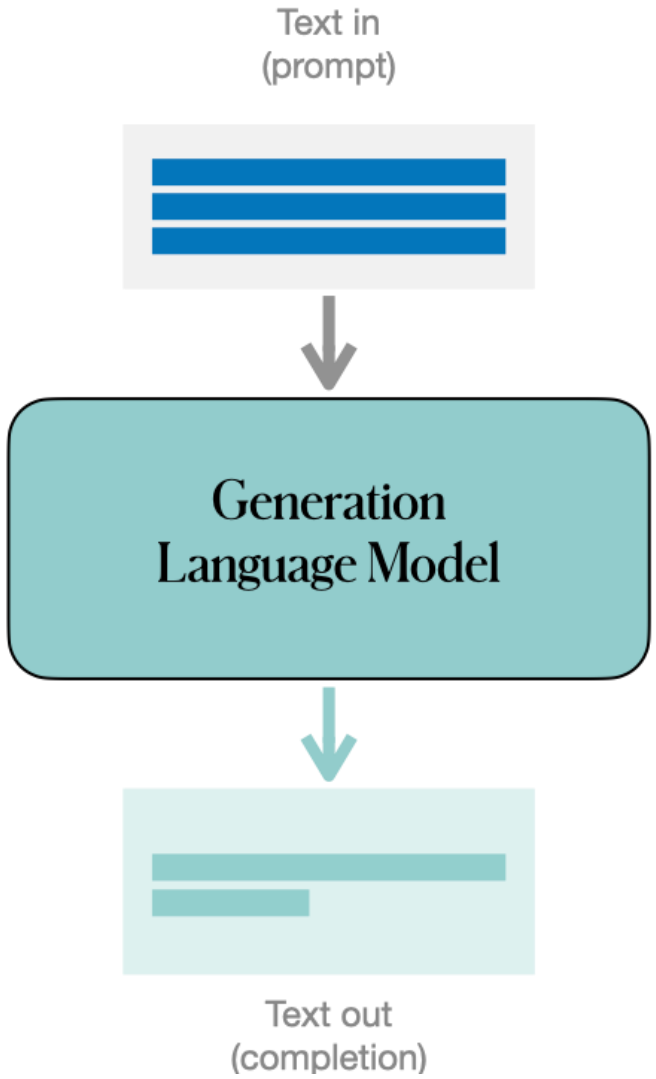
Auto-regressive, uni-directional, causal

Architecture: Transformer decoder,

e.g. Generative Pretrained Transformer (GPT)

Training: next word prediction





e.g. GPT

1. Representation Language Model

Learn a joint text-video embedding

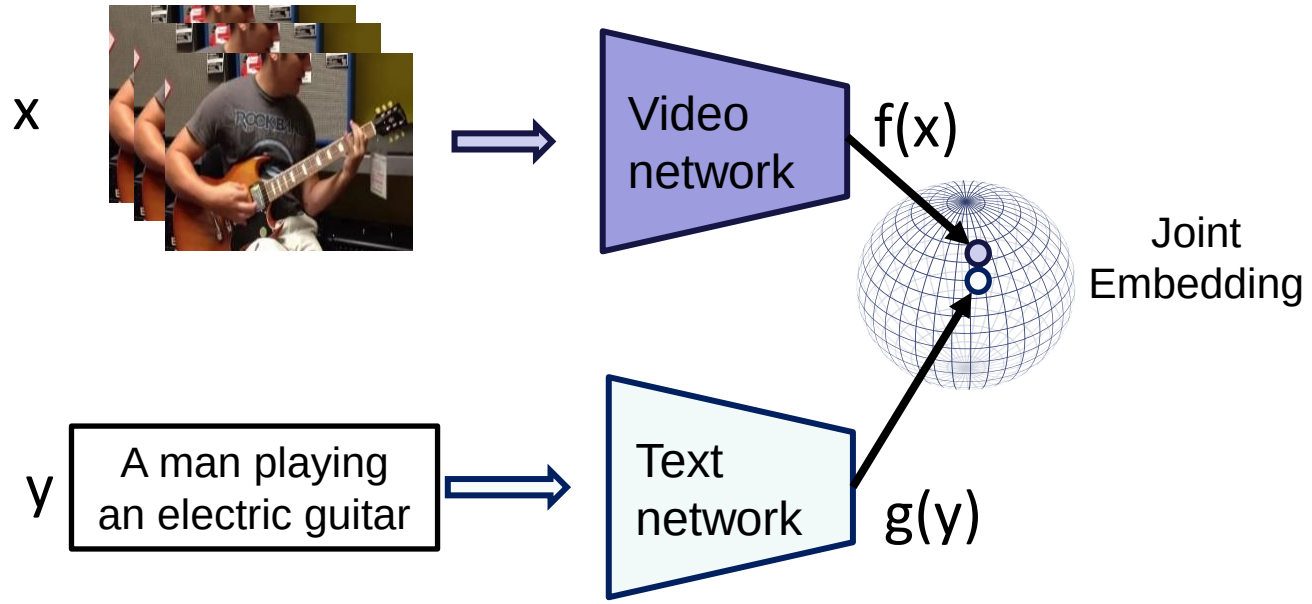
Architecture: Dual Encoder

Separate networks for video and text encoding

Score similarity, e.g. by $f(x)^\top g(y)$

Contrastive loss function for training

Dataset for weak supervision?



This is an "old" idea: DeVise, *Frome et al.*, ***NeurIPS2013*** and ***WSABIE***, *Weston et al.* ***IJCAI 2011***.

Conceptual Captions Image-Caption Dataset

3M image-text pairs from Alt-text HTML attribute of web images

- Unique tokens (vocabulary): 51,201

- Tokens per caption:

Mean	StdDev	Median
10.3	4.5	9.0

Piyush Sharma, Nan Ding, Sebastian Goodman, Radu Soricut, "Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning" ACL 2018

WebVid-2M Video-Caption Dataset

2.5M video-text pairs from stock footage websites

“Runners feet in a sneakers close up. realistic three dimensional animation.”

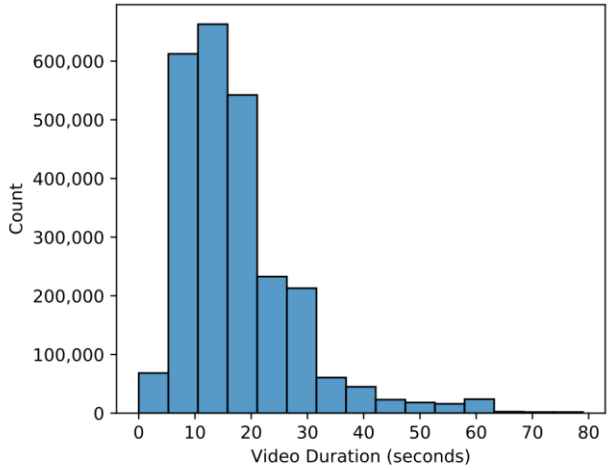
“Female cop talking on walkie talkie, responding emergency call, crime prevention”

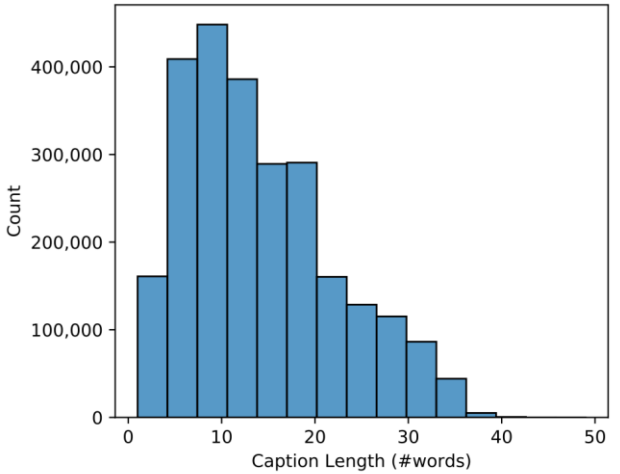
“Billiards, concentrated young woman playing in club”

“Lonely beautiful woman sitting on the tent looking outside. wind on the hair and camping on the beach near the colors of water and shore. freedom and alternative tiny house for traveler lady drinking”

“Kherson, ukraine - 20 may 2016: open, free, rock music festival crowd partying at a rock concert. hands up, people, fans cheering clapping applauding in kherson, ukraine - 20 may 2016. band performing”

“Cabeza de toro, punta cana/ dominican republic - feb 20, 2020: 4k drone flight over coral reef with manta"





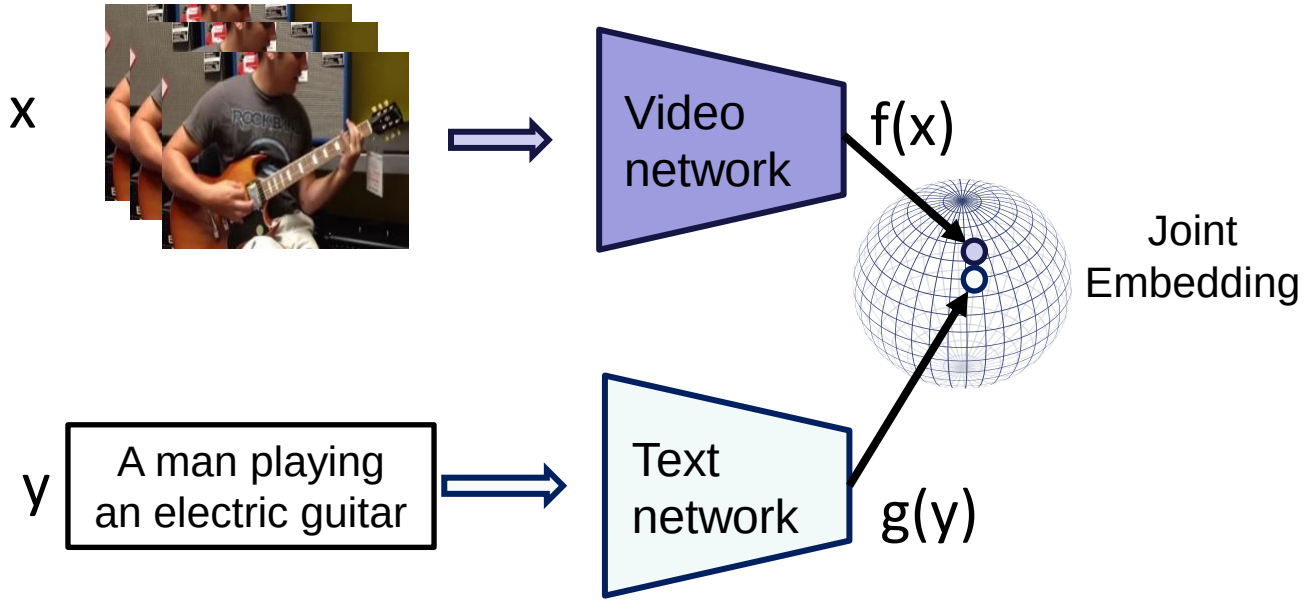
Captions written manually to encourage use of video clips

- less noisy than narrated instructional videos

Objective: learn a joint text-video embedding

Architecture: Dual Encoder

Separate networks for video and text encoding



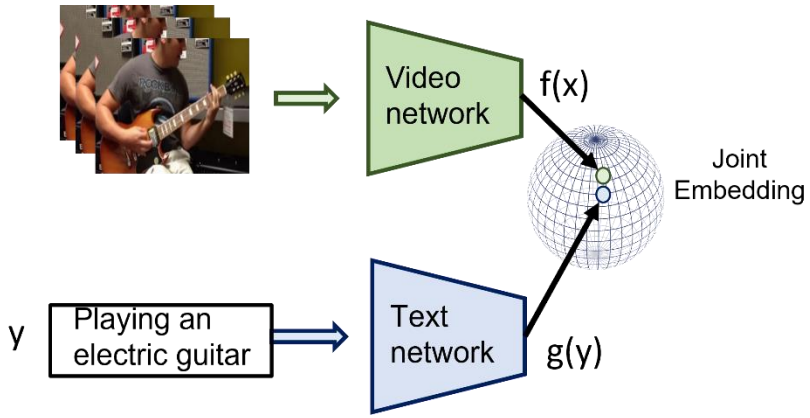
Score similarity, e.g. by $f(x)^\top g(y)$

Train end-to-end with a **contrastive loss** on:

- CC 3M
- WebVid-2M

Learn from both images and videos with captions

- Transformer based architecture that can naturally ingest images and videos
  - Visual encoder that accepts a variable-length sequence
  - Treating images as 1-frame videos**, *frozen in time***
  - End-to-end-training (for images and video) from pixels
- Train on
  - Conceptual Captions 3M (images with captions)
  - WebVid-2M (video clips with captions)

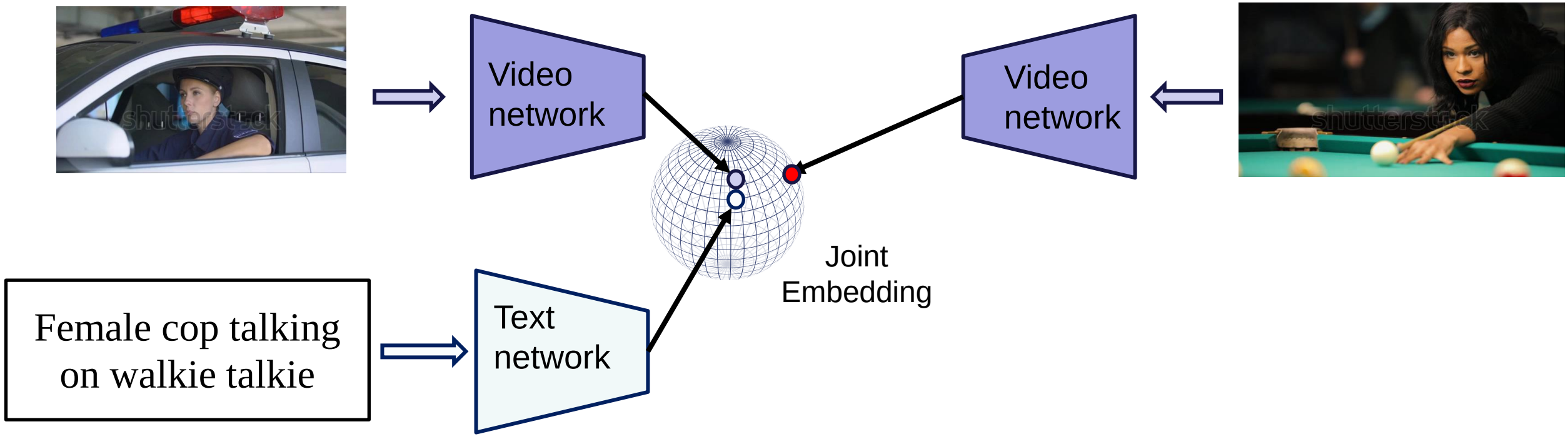


"Frozen in Time: a Joint Video and Image Encoder for End-to-End Retrieval",
Max Bain, Arsha Nagrani, Gül Varol, Andrew Zisserman, ICCV 2021

Application: Text-to-video retrieval

- Given a (large scale) library of video clips
- And a text query that describes the target clip
- Retrieve the clip corresponding to the description

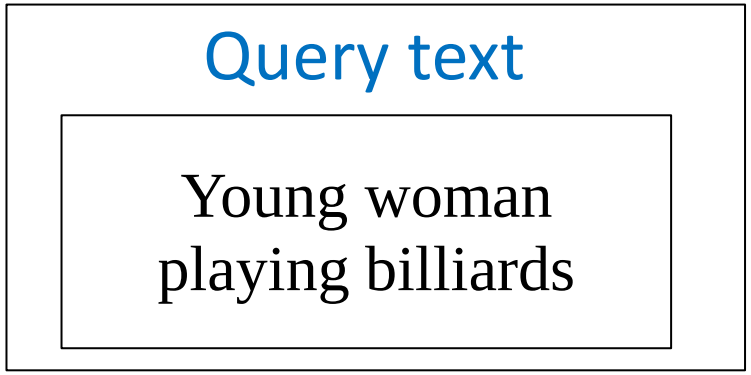
Using the joint text-video embedding for retrieval



Using the joint text-video embedding for retrieval

Online: encode query text and search for match

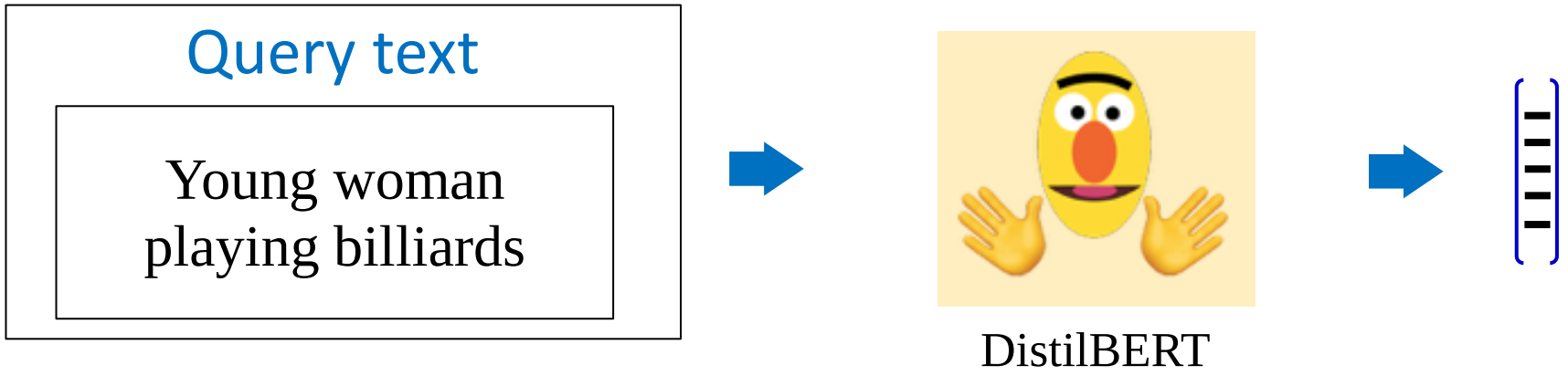
- use approximate nearest neighbours (ANN) for fast search



Using the joint text-video embedding for retrieval

Online: encode query text and search for match

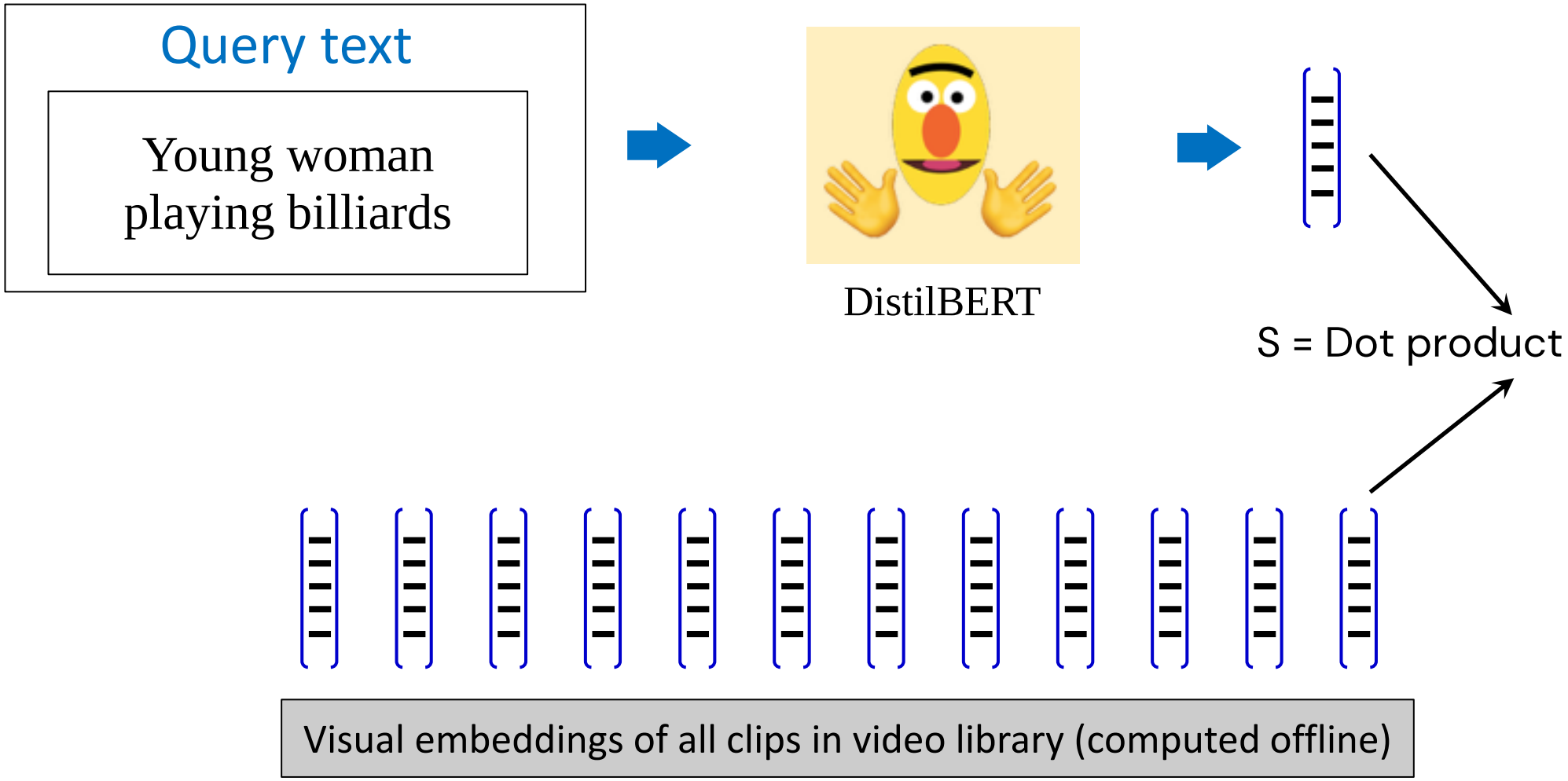
- use approximate nearest neighbours (ANN) for fast search



Using the joint text-video embedding for retrieval

Online: encode query text and search for match

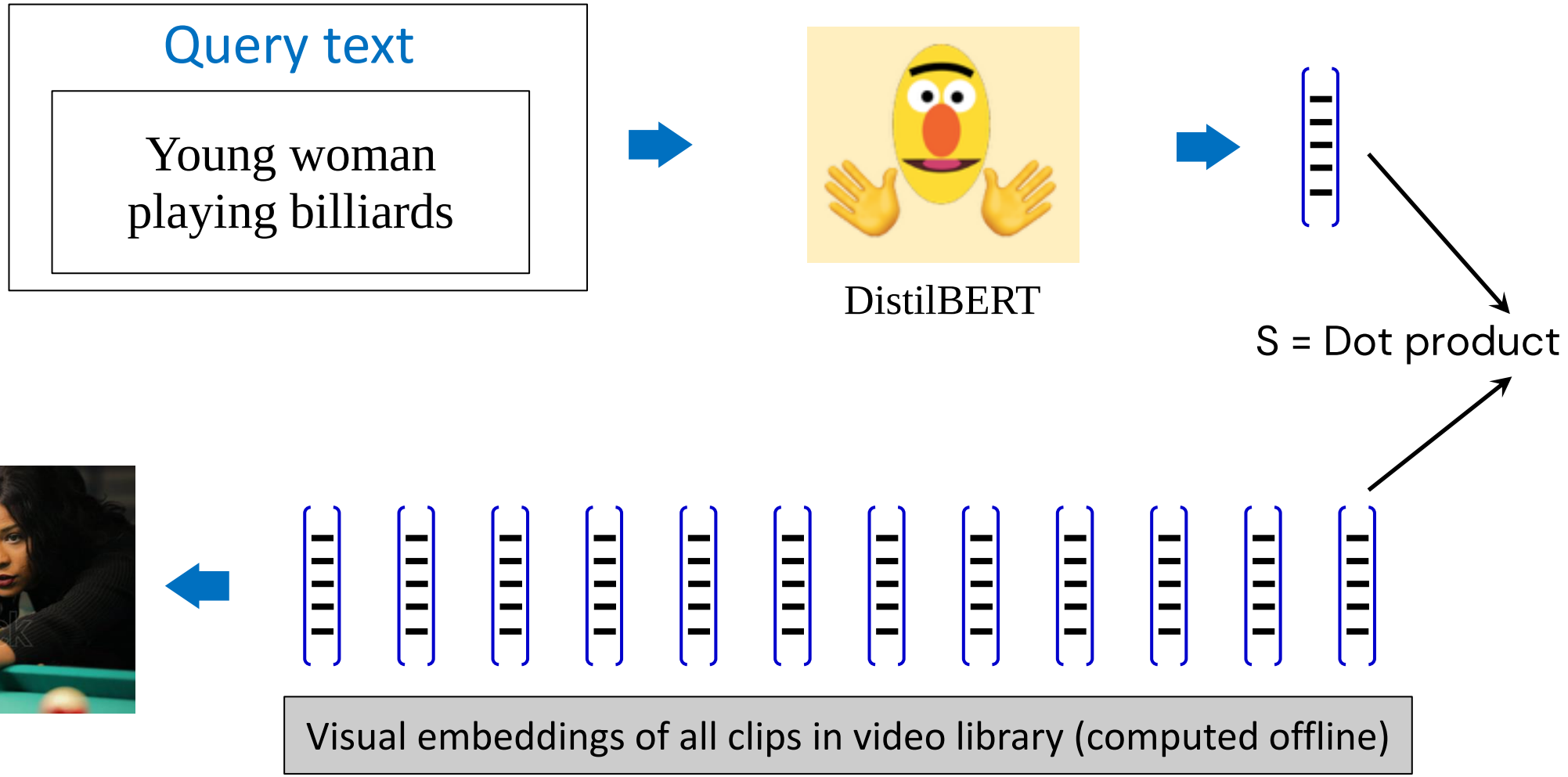
- use approximate nearest neighbours (ANN) for fast search



Using the joint text-video embedding for retrieval

Online: encode query text and search for match

- use approximate nearest neighbours (ANN) for fast search



Real-Time Video Search Demo

Enter your search term... max display: 8

Paper

M. Bain, A. Nagrani, G. Varol, A. Zisserman.
Frozen in Time: A Joint Video and Image Encoder for End to End Paper.
ICCV, 2021.
(hosted on ArXiv)

[Bibtex]

Demo at: https://meru.robots.ox.ac.uk/frozen-in-time/

2. Model using a Generative Language Model

Flamingo: a Visual Language Model for Few-Shot Learning

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, Karen Simonyan, ArXiv 2022

DeepMind

What is Flamingo?

Output: Free-form text

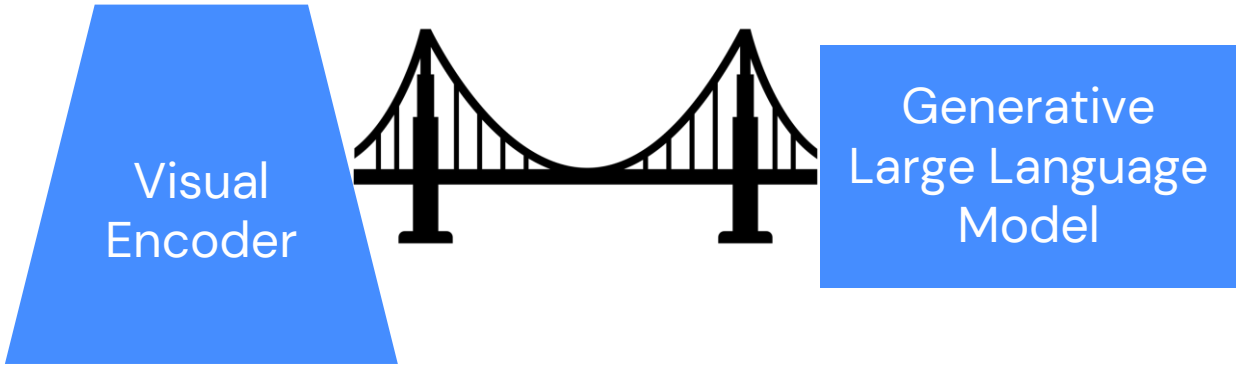
A portrait of Salvador Dali with a robot head.

Flamingo Model

Input: Text and visual data interleaved

Output: A propaganda poster depicting a cat dressed as French emperor Napoleon holding a piece of cheese.

Overview of architecture



The perception

The “reasoning part” and “knowledge source”

The Visual Encoder and the Generative Large Language Model (LLM) are pre-trained and frozen.

LLM, like GPT-3, has few-shot learning capability

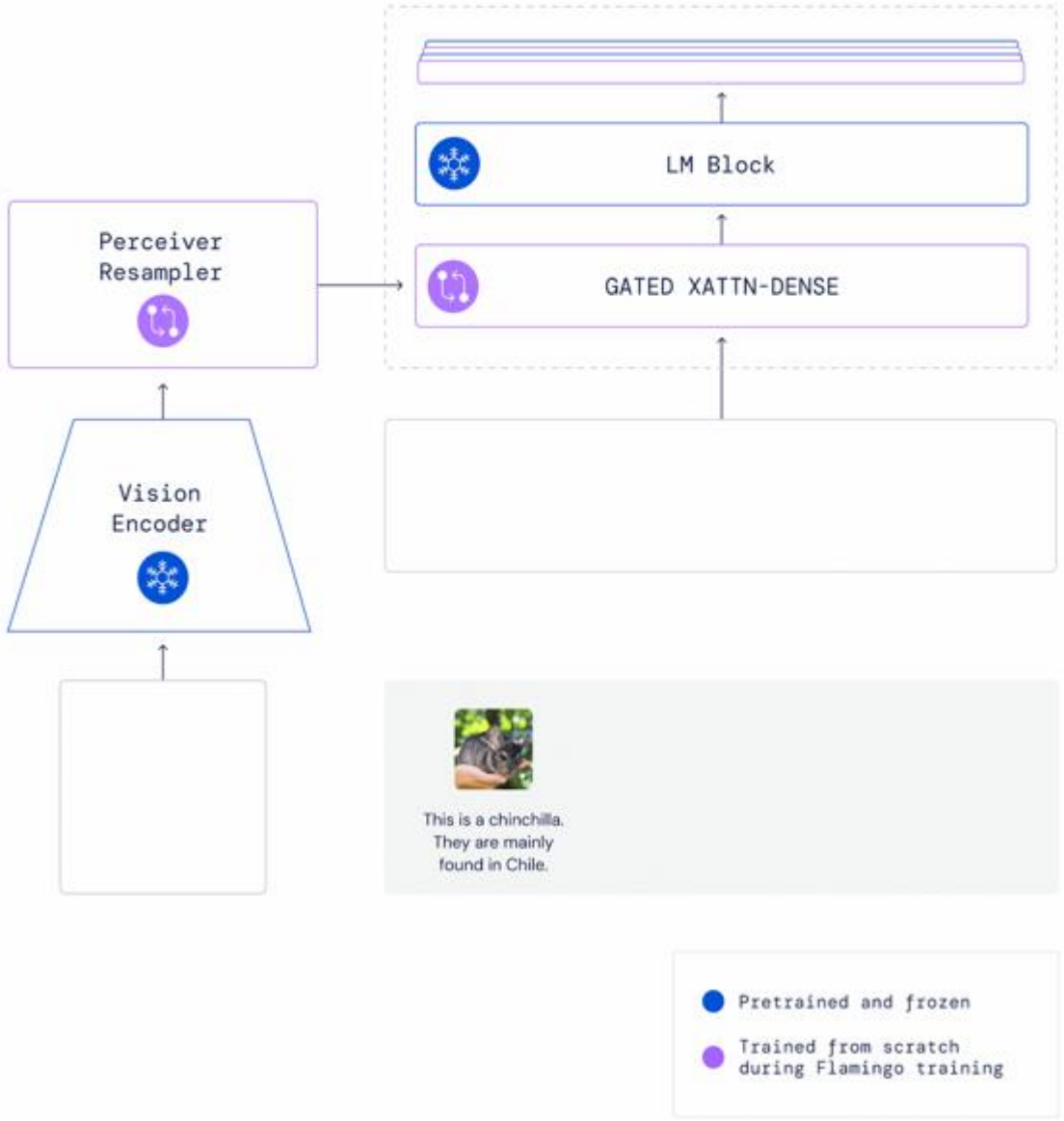
Flamingo model:

- Visually-conditioned auto-regressive text generation model
- **Input**: a sequence of words interleaved with images and/or videos
- **Output**: Free-form text

Overview of architecture

Input: Text and visual data interleaved

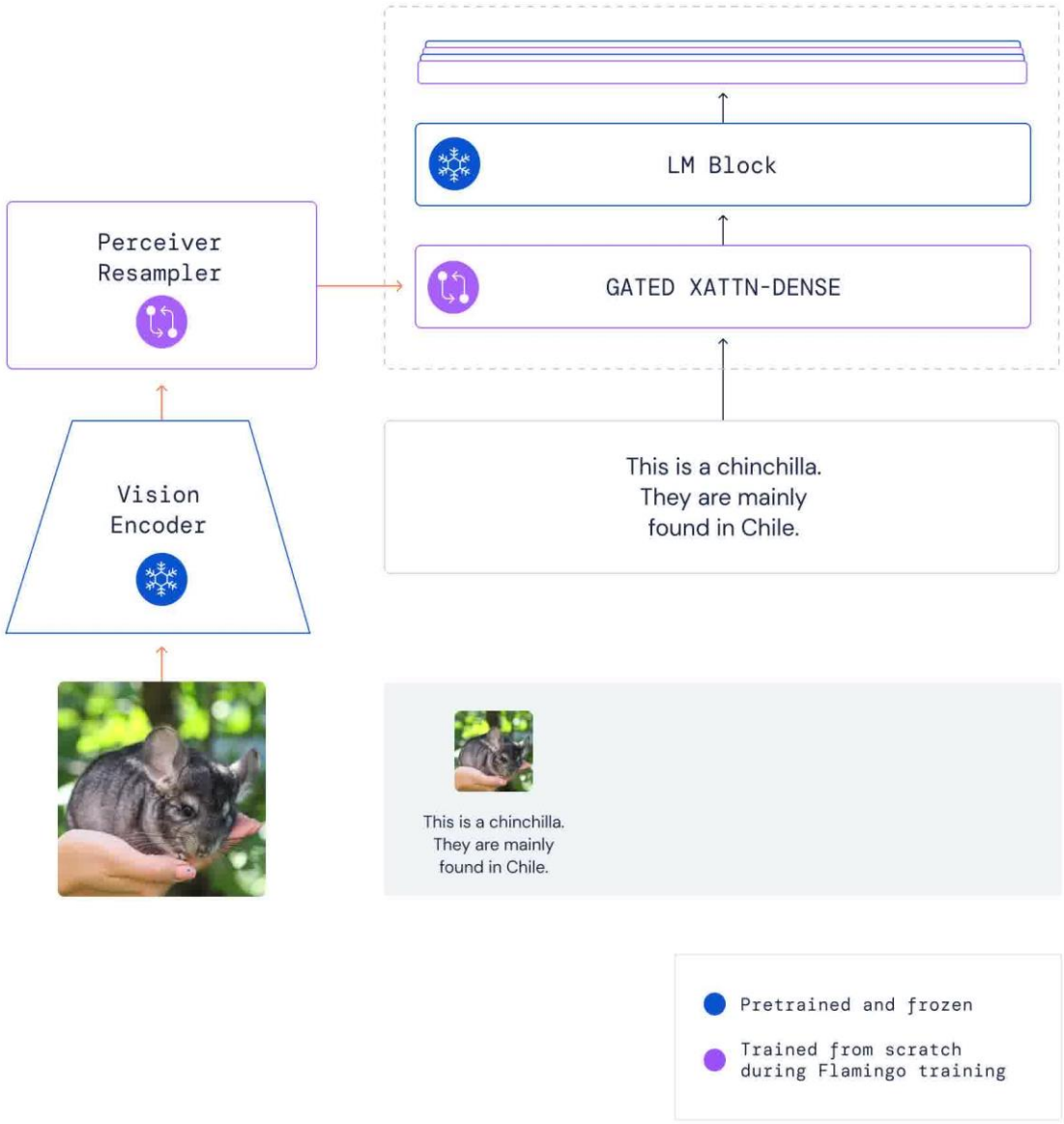
Output: Free-form text



Overview of architecture

Input: Text and visual data interleaved

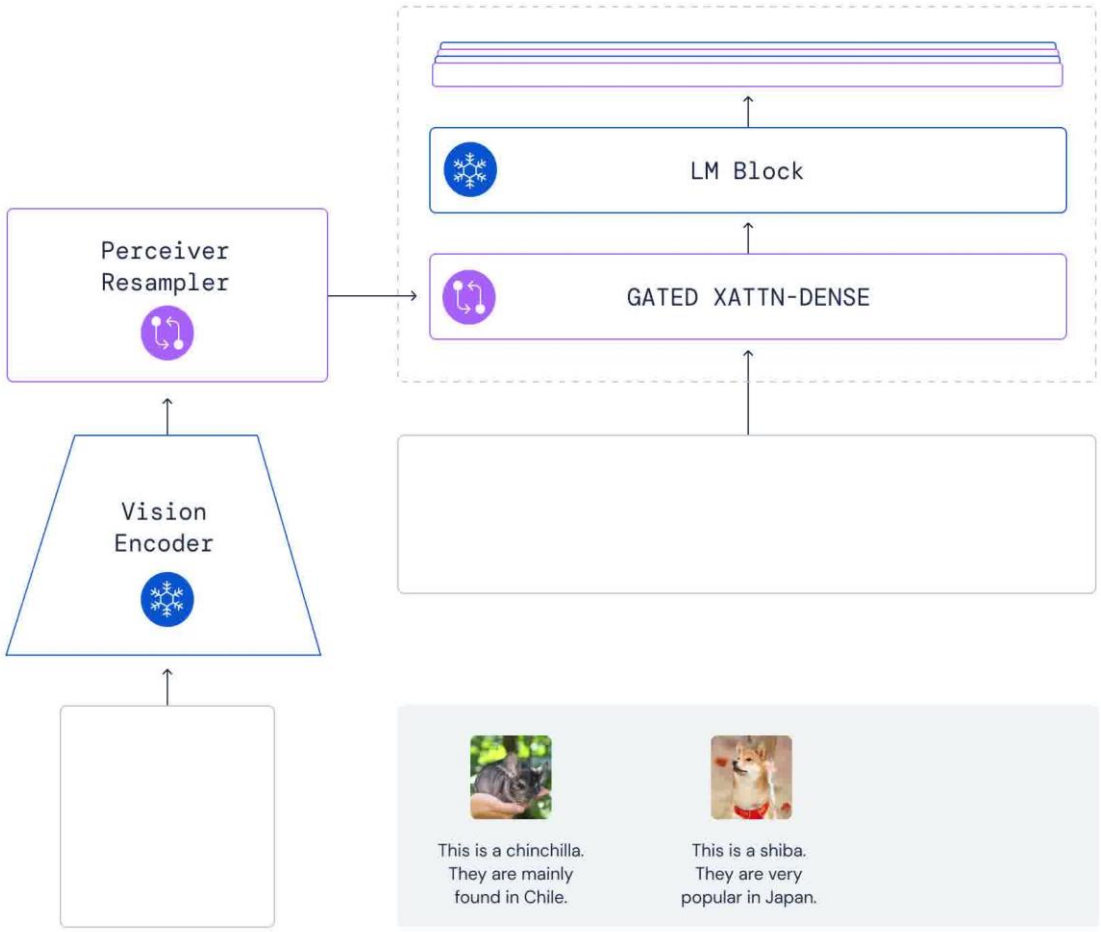
Output: Free-form text



Overview of architecture

Input: Text and visual data interleaved

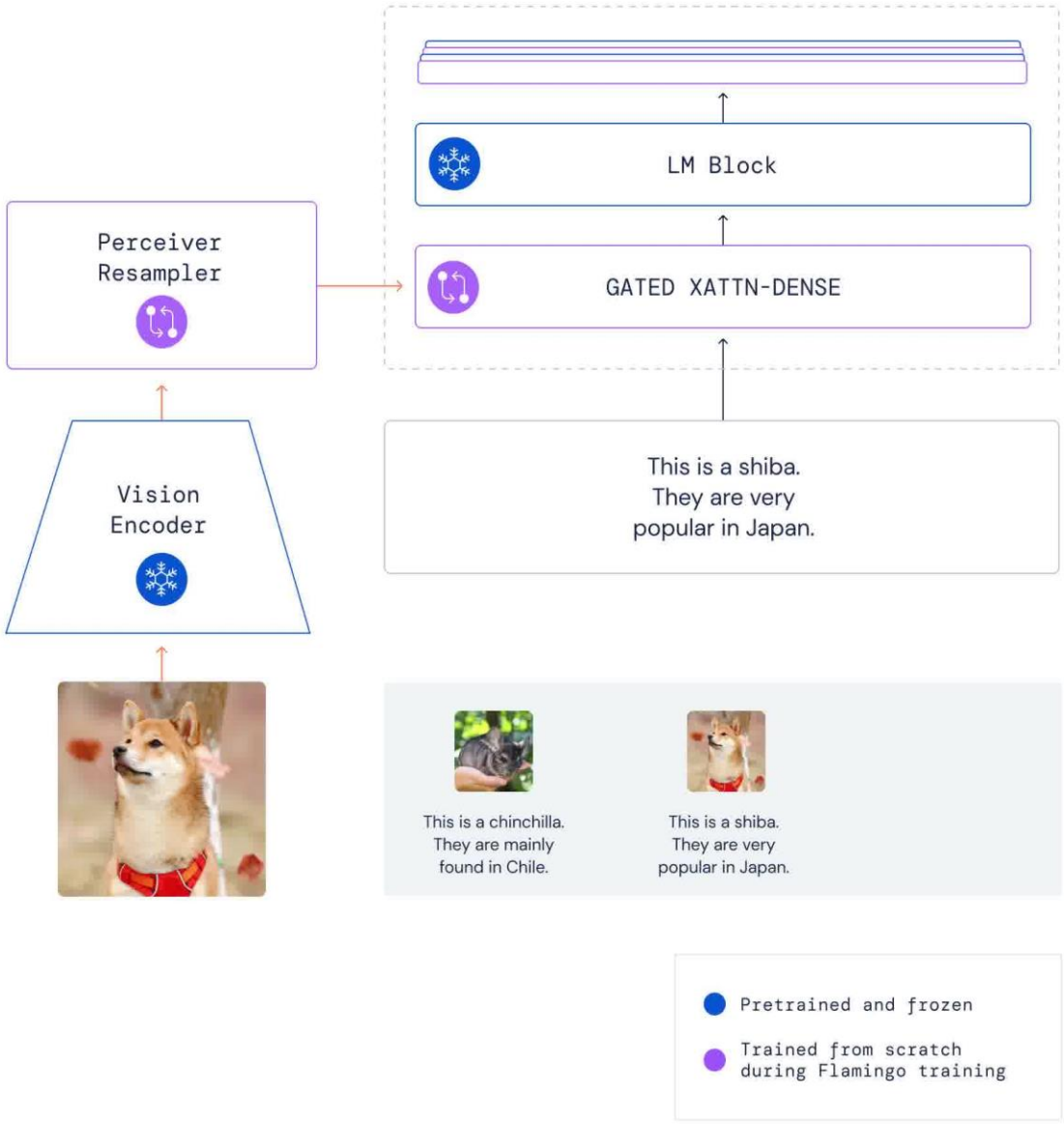
Output: Free-form text



Overview of architecture

Input: Text and visual data interleaved

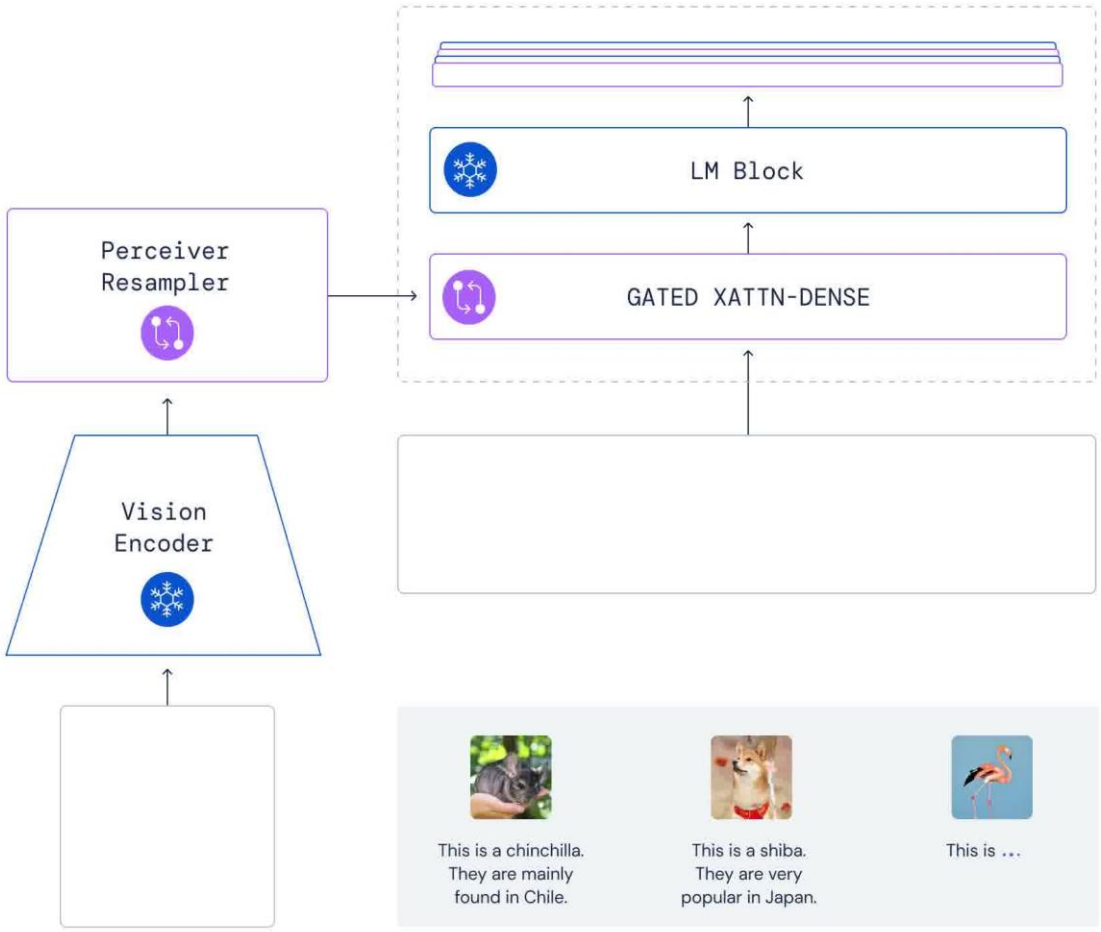
Output: Free-form text



Overview of architecture

Input: Text and visual data interleaved

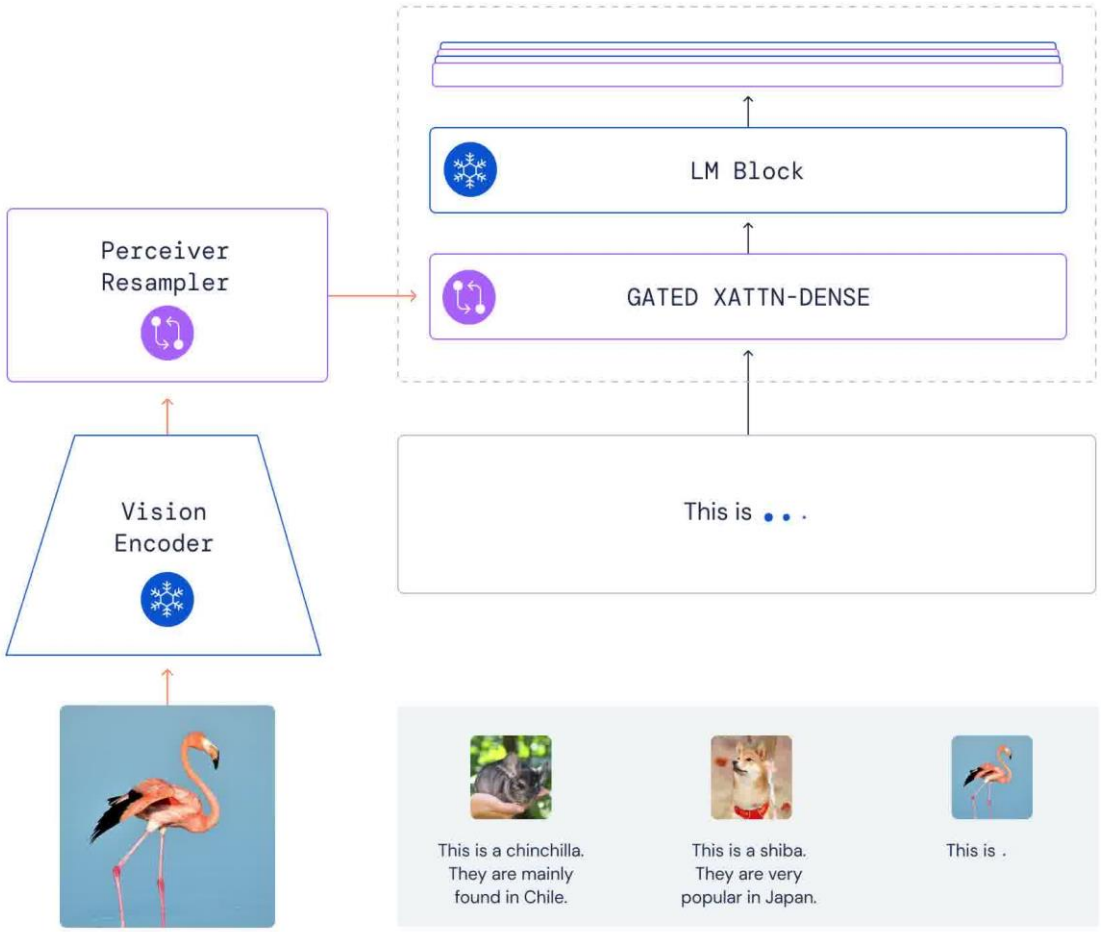
Output: Free-form text



Overview of architecture

Input: Text and visual data interleaved

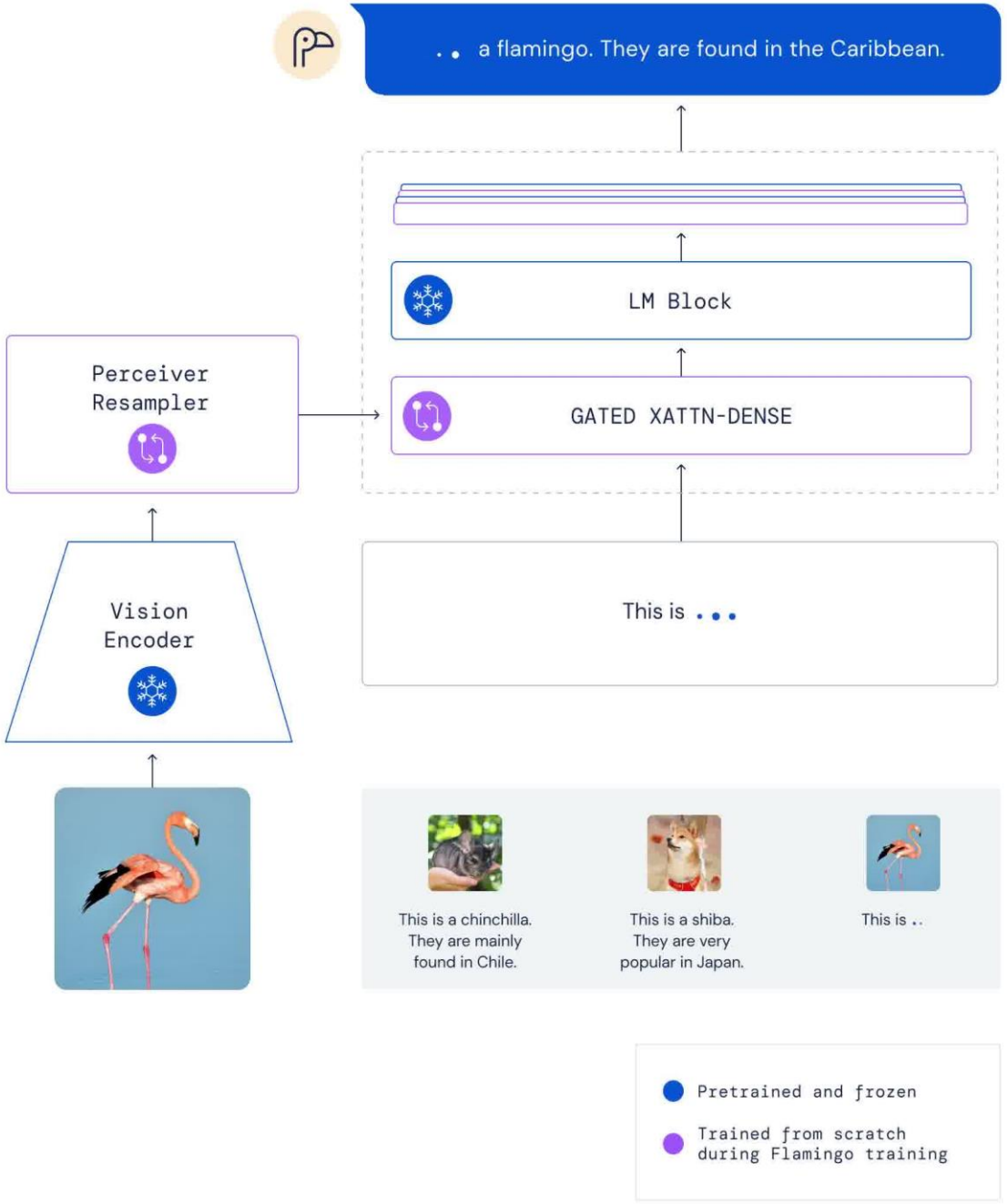
Output: Free-form text



Overview of architecture

Input: Text and visual data interleaved

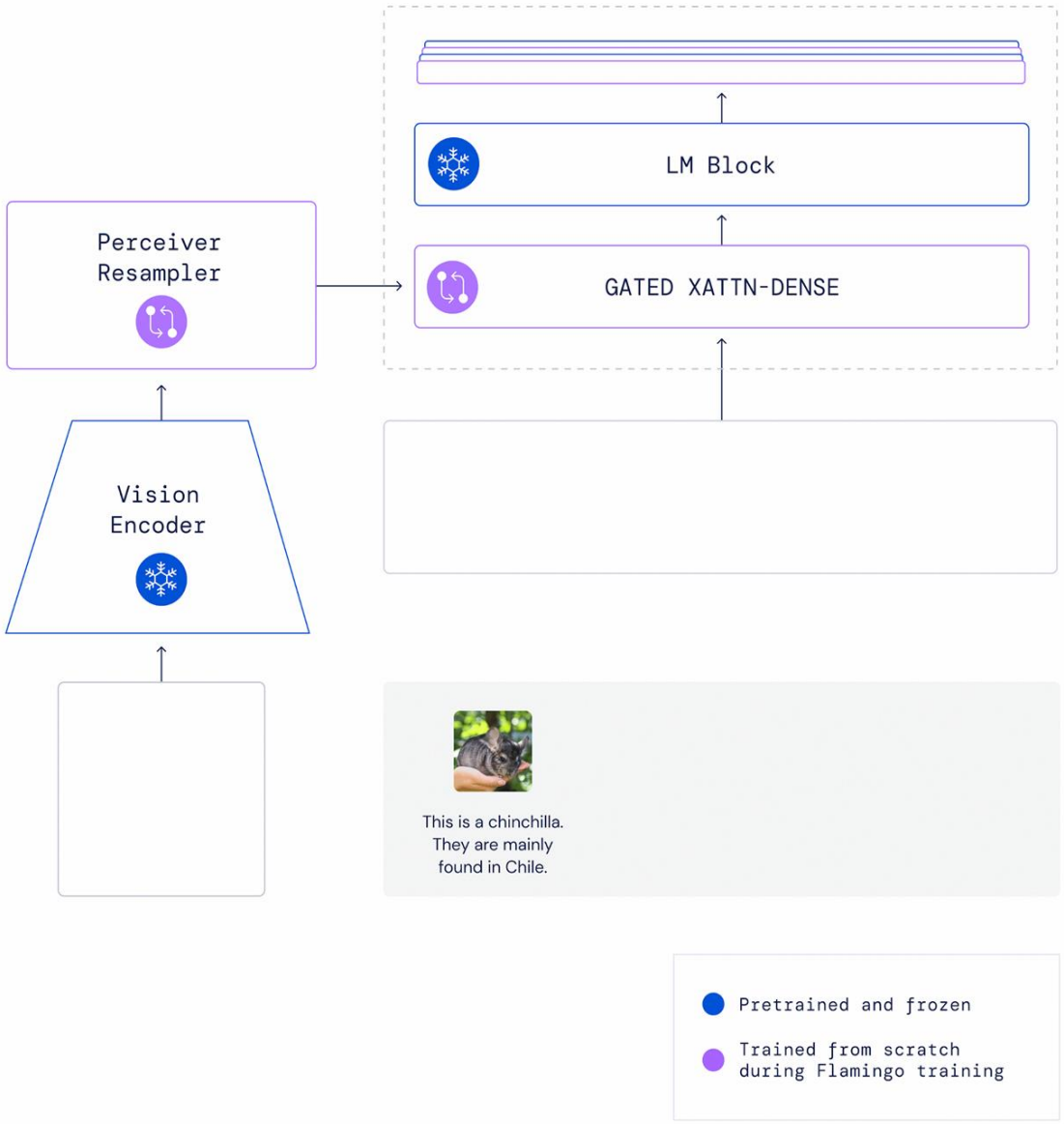
Output: Free-form text



Overview of architecture

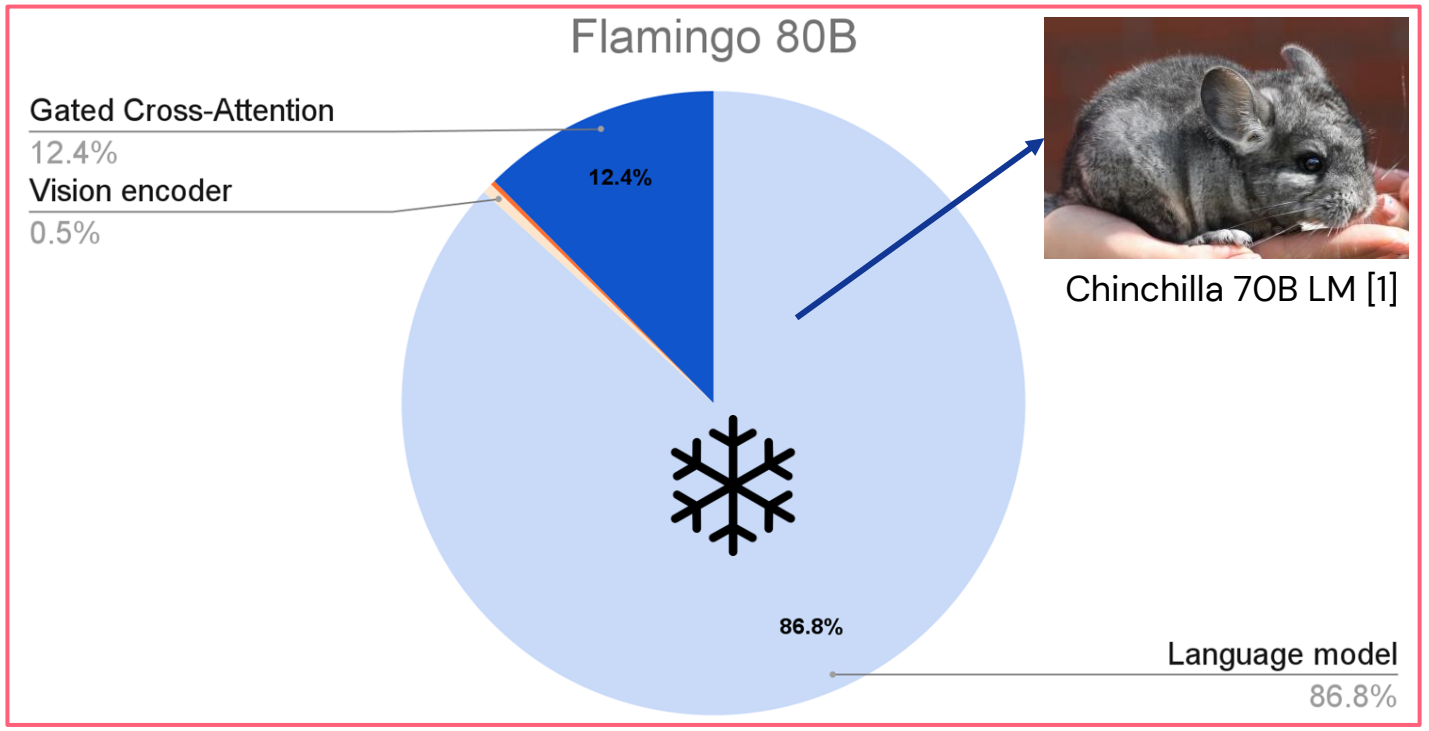
Input: Text and visual data interleaved

Output: Free-form text



The number and proportion of parameters

(*Flamingo* in short)



Chinchilla 70B LM [1]

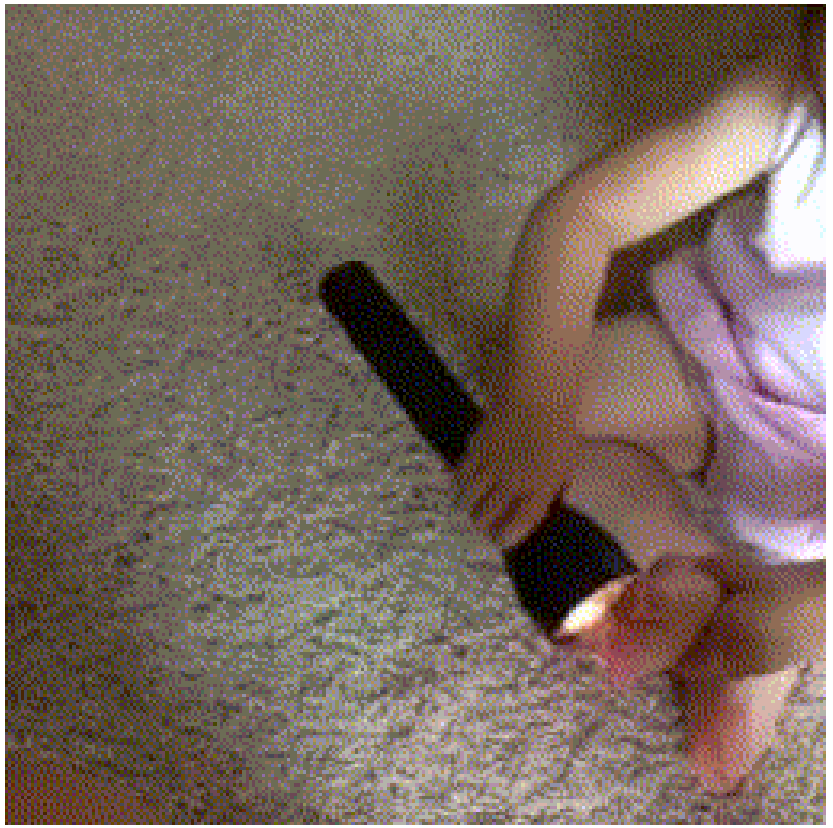
- Vision encoder (NFNet-F6) 435M
- Resampler 193M
- Language Model: 80 layers, D= 8192, # heads = 64

[1] Hoffmann et al., Training Compute-Optimal Large Language Models, 2022

Image / Video paired with Captions

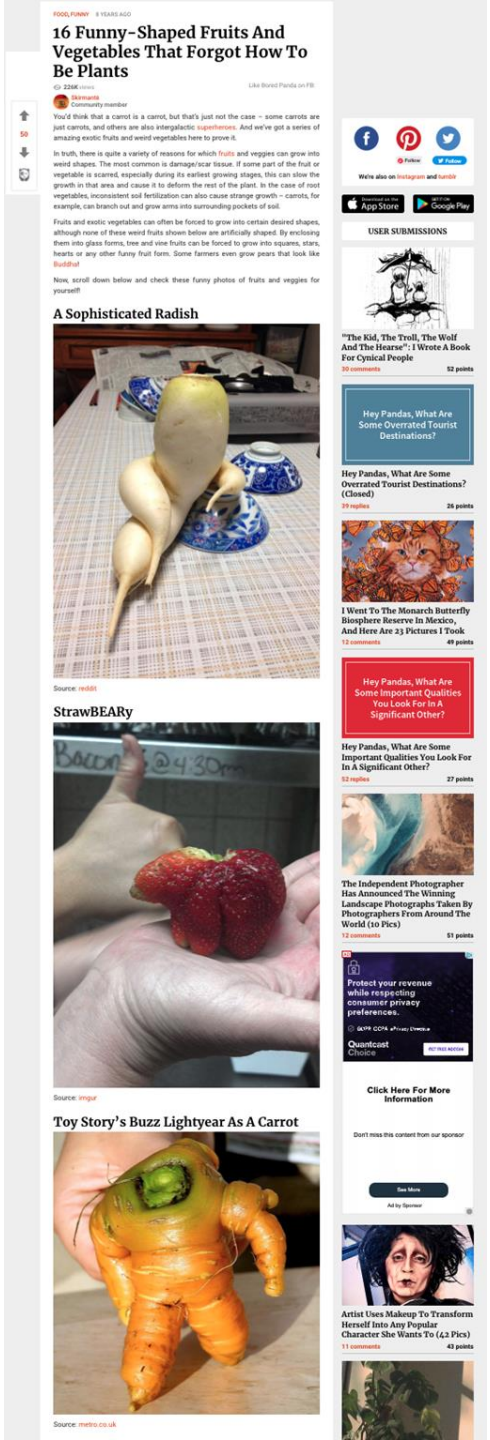
An English bulldog standing on a skateboard.

Image-Text pairs (2B examples)



A little girl playing with a flashlight.

Video-Text pairs (27M examples)



M3W: Massive MultiModal Web Dataset

44M scraped webpages with interleaved text and images.

180M images in total. (4 on average per webpage)

Processing →

16 Funny-Shaped Fruits And Vegetables That Forgot How To Be Plants

You'd think that a carrot is a carrot, but that's just not the case – some carrots are just carrots, and others are also intergalactic superheroes. [...]. Some farmers even grow pears that look like Buddha!

Now, scroll down below and check these funny photos of fruits and veggies for yourself!

A Sophisticated Radish

<IMAGE PLACEHOLDER 1>

StrawBEARy

<IMAGE PLACEHOLDER 2>

Toy Story's Buzz Lightyear As A Carrot

<IMAGE PLACEHOLDER 3>

<IMAGE PLACEHOLDER 1> →

<IMAGE PLACEHOLDER 2> →

<IMAGE PLACEHOLDER 3> →

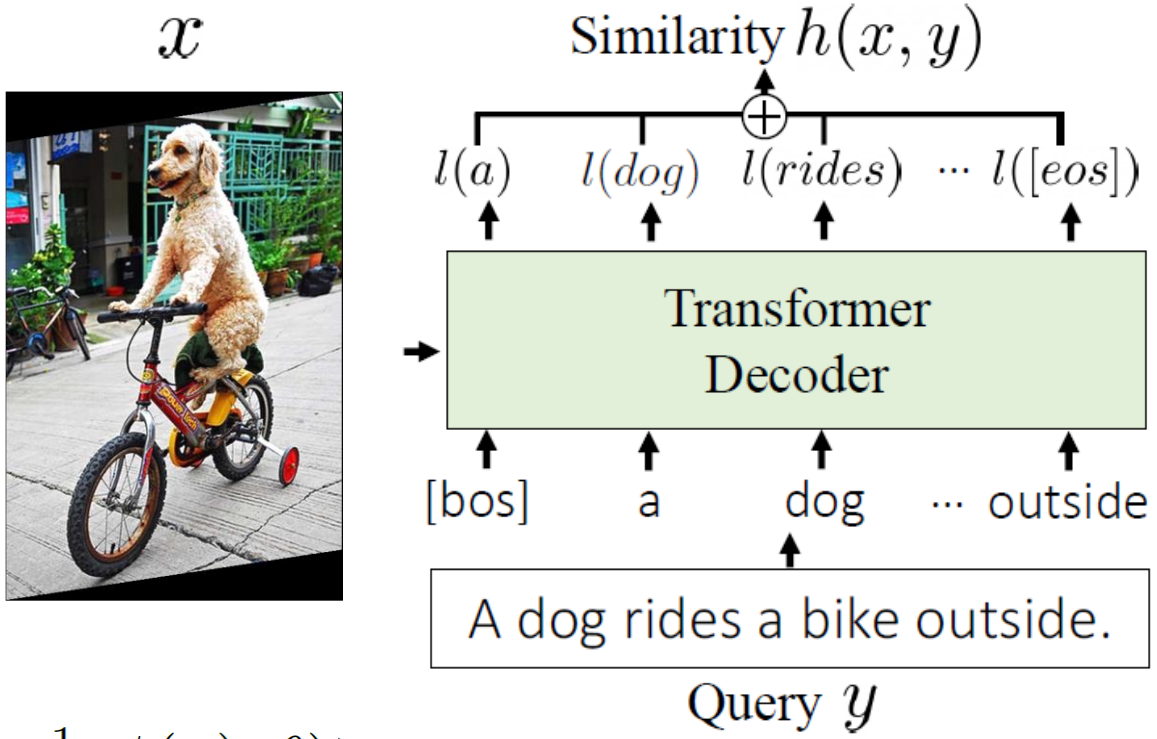
Training: captioning loss

Image x

Sentence y = [y^1, y^2, ..., y^L]

Optimize log likelihood:

$$\text{score } h(x,y) = \sum_{l=1}^{L} \log(\, p(y^l | y^{l-1}, \ldots, y^1, \phi(x); \theta))$$



Train: Perceiver Resampler, and Cross-attention (XATTN) layers

DeepMind

Qualitative examples

Video Question and Answering

Input Prompt

What happens to the man after hitting the ball? Answer:

Completion

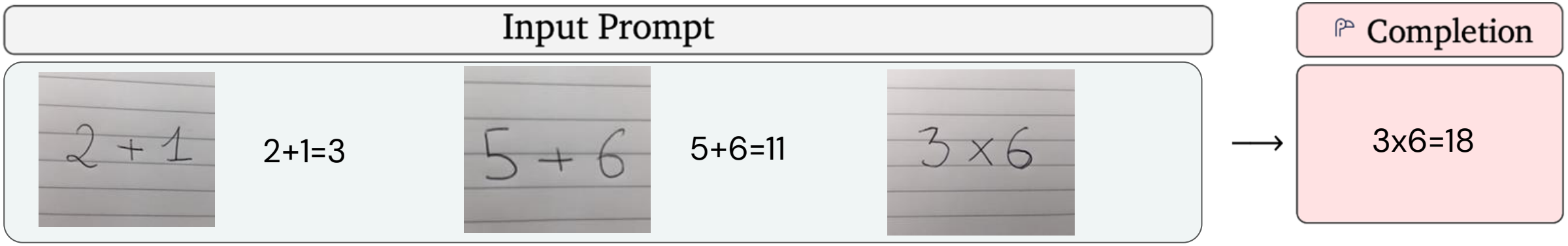
he falls down.

Few shot learning Example 1

Condition the model to solve various tasks with only a few input-output examples.

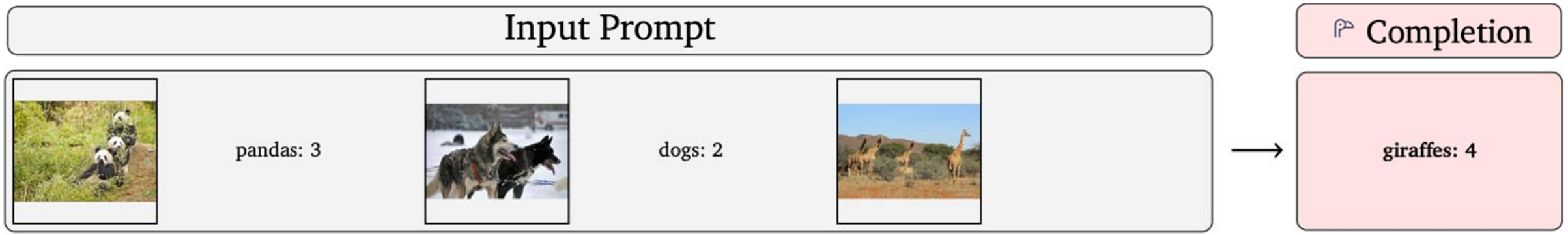
Few shot learning Example 2

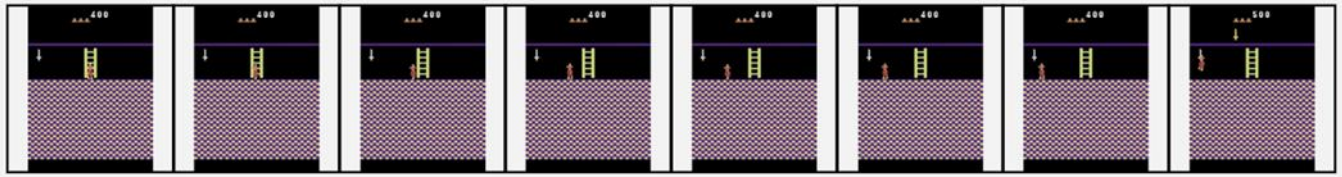
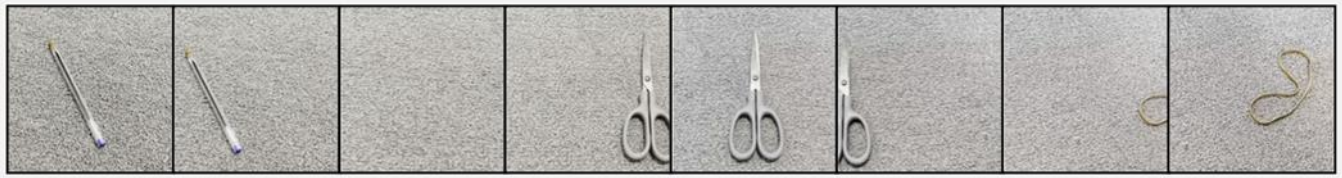
Condition the model to solve various tasks with only a few input-output examples.



Few shot learning Example 3

Condition the model to solve various tasks with only a few input-output examples.

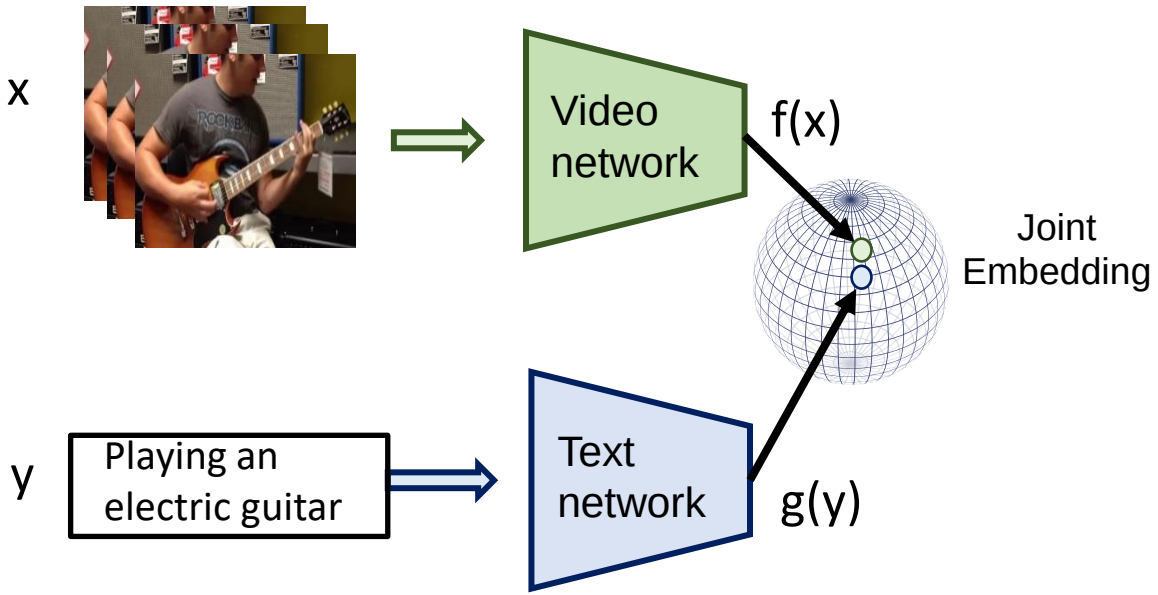


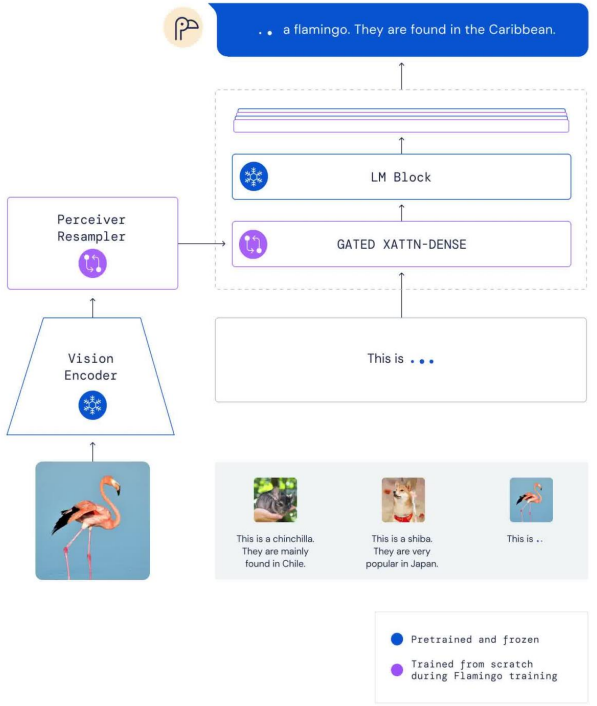
Input Prompt	Completion
 Question: What is happening here? Answer:	**The dachschund puppy is being weighed on a scale.**
 Question: What object is the avatar picking up? Answer:	**A sword.**
 Question: What is the object being lifted? Answer:	**The object is a small plastic bowl.**
 Question: What are the three objects in this video? Answer:	**A scissors, a pen, and a rubber band.**
 Question: What is written here? Answer:	**Flamingo.**

Part III Summary

Learning with weak supervision from text:

- Dual encoder using a representation language model
- Video description and few shot learning using a generative language model





Summary and future challenges

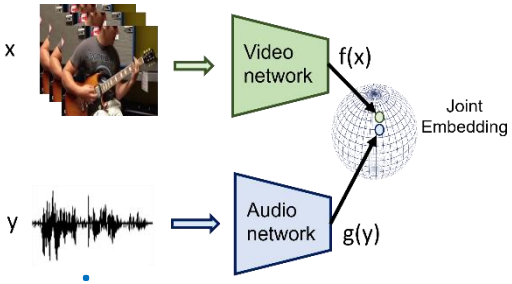
Video Understanding

- What is in the video?
  - **objects, animals, people** …
- Where is it?
  - 3D scene spatial layout
  - object shape
  - human pose …
- What is happening?
  - **human actions**
  - activities …

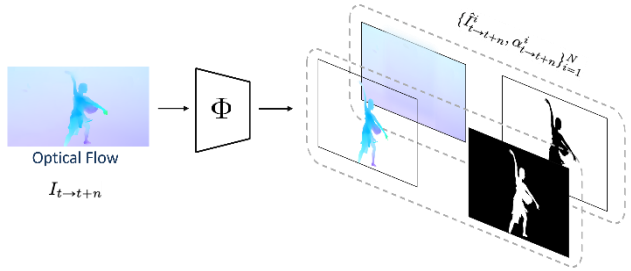
Objective: learning to recognize without explicit supervision

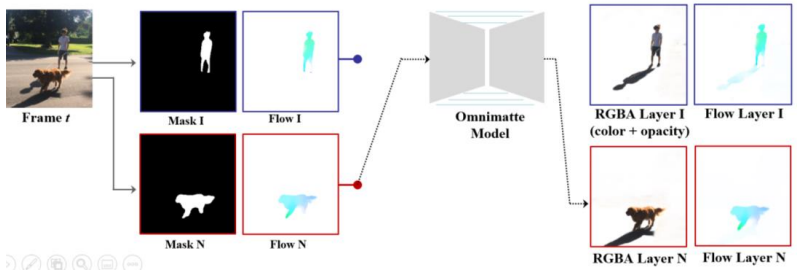
How can we learn to recognize without explicit supervision?

Part I: Using multi-modal (audio-visual) self-supervised learning

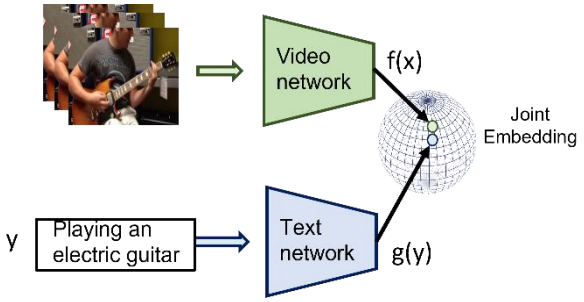


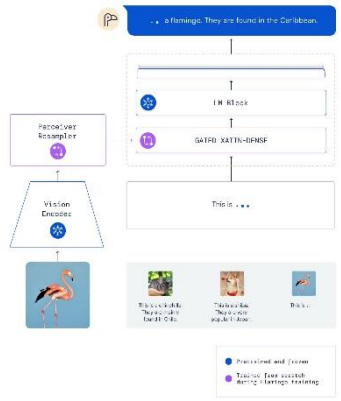
Part II: Discovering objects and their effects using self-supervised learning





Part III: Using weak supervision from videos with text





Challenges for video understanding

1. Longer term understanding (beyond 10s)
   - Need new datasets to explore this (so far have instructional & cooking videos, Ego4D)
   - Will also drive new architectures, e.g. with memory

2. Move from `what?' to `why?'
   - Why is this happening (story understanding)

3. Object-centric representations
   - Represent the objects (and animals and humans) directly

4. Multi-modality and multiple information streams
   - Learn from both audio and aligned text
   - And also: speech/ASR, scene text, multi-lingual …
   - Required to fully understand what is happening in a video