THE UNIVERSITY OF TEXAS AT DALLAS

Vision + X

CS 6384 Computer Vision

Professor Yapeng Tian

Department of Computer Science

Some slides borrowed from Prof. Yu Xiang

Image Classification

ImageNet dataset

- Training: 1.2 million images
- Testing and validation: 150,000 images
- 1000 categories

n02119789: kit fox, Vulpes macrotis
n02100735: English setter
n02096294: Australian terrier
n02066245: grey whale, gray whale, devilfish, Eschrichtius gibbosus, Eschrichtius robustus
n02509815: lesser panda, red panda, panda, bear cat, cat bear, Ailurus fulgens
n02124075: Egyptian cat
n02417914: ibex, Capra ibex
n02123394: Persian cat
n02125311: cougar, puma, catamount, mountain lion, painter, panther, Felis concolor
n02423022: gazelle

https://image-net.org/challenges/LSVRC/2012/index.php

Vision + Language

Image captioning

Object grounding

Visual question answering

Representation learning with images and languages

Text-to-Image Generation

…

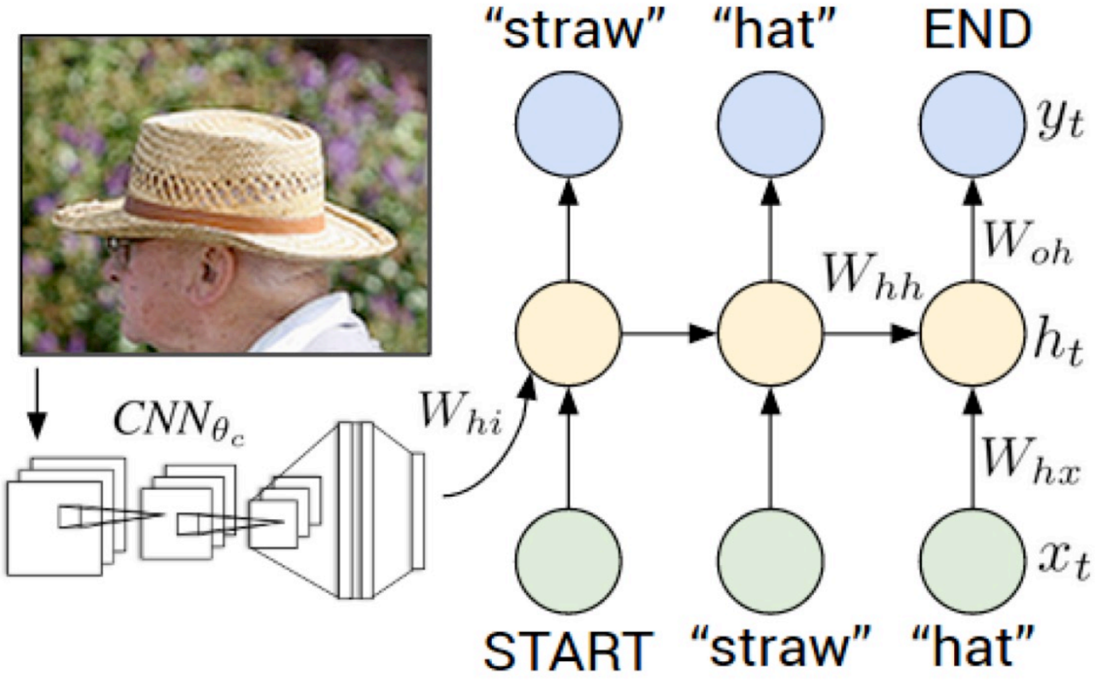
Image Captioning

Automatically generate texture descriptions of images

the person is riding a surfboard in the ocean

https://www.tensorflow.org/tutorials/text/image_captioning

Image Captioning with RNNs



- Image embedding

$$b_v = W_{hi}[CNN_{\theta_c}(I)]$$

- Hidden state at time t

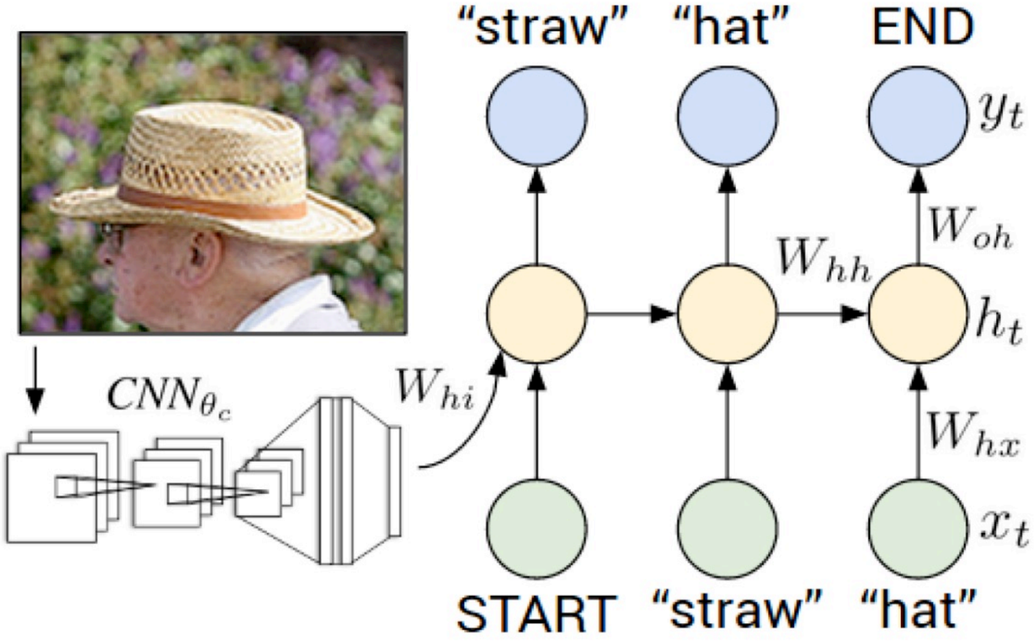
$$h_t = f(W_{hx}x_t + W_{hh}h_{t-1} + b_h + \mathbb{1}(t=1) \odot b_v)$$

Parameters

- Word embedding $x_t = W_w \mathbb{I}_t$
- Output $y_t = softmax(W_{oh}h_t + b_o)$

Deep Visual-Semantic Alignments for Generating Image Descriptions. Karpathy & Fei-fei, CVPR, 2015

Image Captioning with RNNs

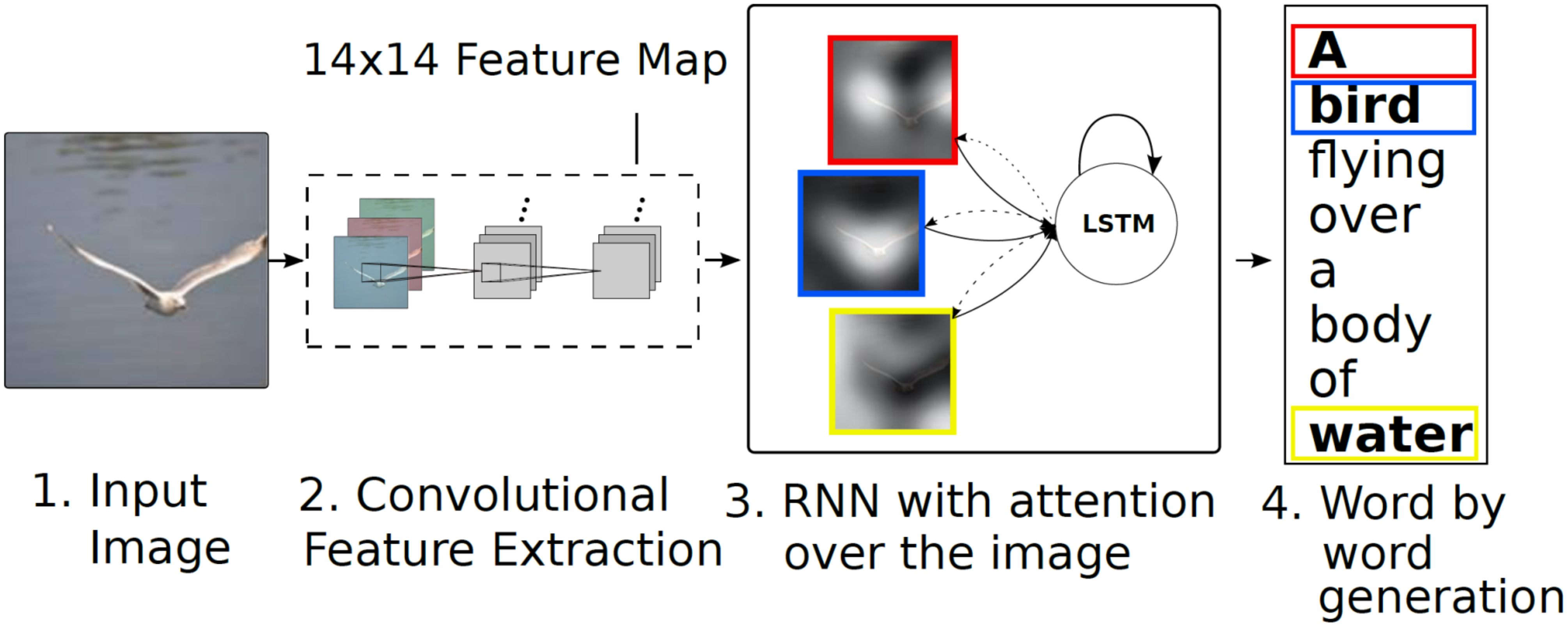


man in black shirt is playing guitar.

construction worker in orange safety vest is working on road.

Deep Visual-Semantic Alignments for Generating Image Descriptions. Karpathy & Fei-fei, CVPR, 2015

Image Captioning with Attentions



Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. Xu et al., PMLR, 2015.

Image Captioning with Attentions

Dataset	Model	BLEU				
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Flickr8k	Google NIC(Vinyals et al., 2014)$^{\dagger\Sigma}$	63	41	27	—	—
	Log Bilinear (Kiros et al., 2014a)$^{\circ}$	65.6	42.4	27.7	17.7	17.31
	Soft-Attention	**67**	44.8	29.9	19.5	18.93
	Hard-Attention	**67**	**45.7**	**31.4**	**21.3**	**20.30**
Flickr30k	Google NIC$^{\dagger\circ\Sigma}$	66.3	42.3	27.7	18.3	—
	Log Bilinear	60.0	38	25.4	17.1	16.88
	Soft-Attention	66.7	43.4	28.8	19.1	**18.49**
	Hard-Attention	**66.9**	**43.9**	**29.6**	**19.9**	18.46
COCO	CMU/MS Research (Chen & Zitnick, 2014)^{a}	—	—	—	—	20.41
	MS Research (Fang et al., 2014)$^{\dagger a}$	—	—	—	—	20.71
	BRNN (Karpathy & Li, 2014)$^{\circ}$	64.2	45.1	30.4	20.3	—
	Google NIC$^{\dagger\circ\Sigma}$	66.6	46.1	32.9	24.6	—
	Log Bilinear$^{\circ}$	70.8	48.9	34.4	24.3	20.03
	Soft-Attention	70.7	49.2	34.4	24.3	**23.90**
	Hard-Attention	**71.8**	**50.4**	**35.7**	**25.0**	23.04

BLEU (BiLingual Evaluation Understudy) **METEOR (Metric for Evaluation of Translation with Explicit ORdering)**

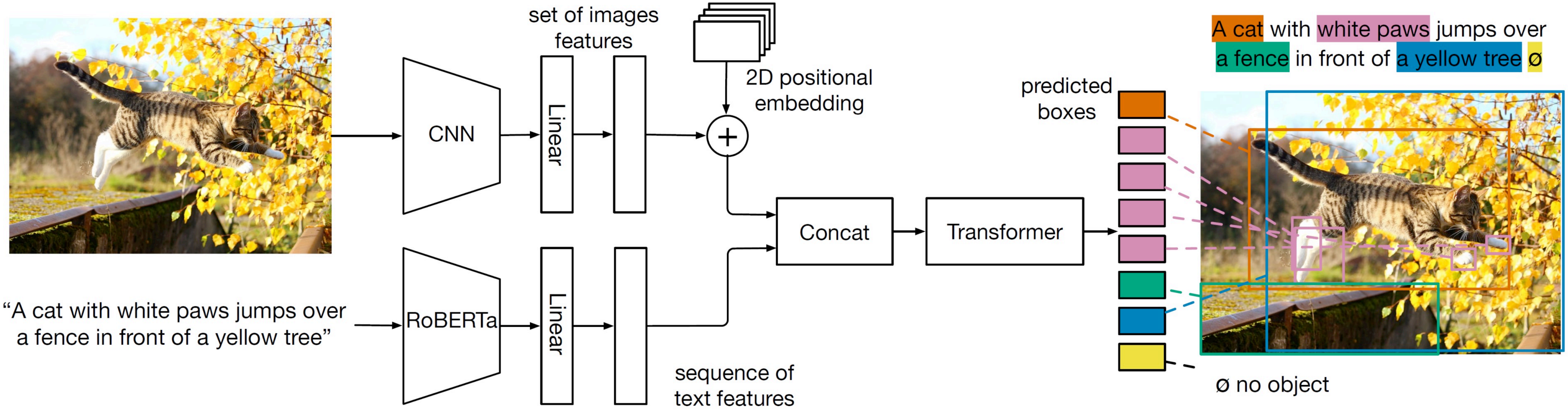
Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. Xu et al., PMLR, 2015.

Object Grounding

A man with **pierced ears** is wearing **glasses** and **an orange hat**.
A man with **glasses** is wearing **a beer can crotched hat**.
A man with **gauges** and **glasses** is wearing **a Blitz hat**.
A man in **an orange hat** starring at **something**.
A man wears **an orange hat** and **glasses**.

Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. Plummer et al., ICCV, 2015.

Object Grounding



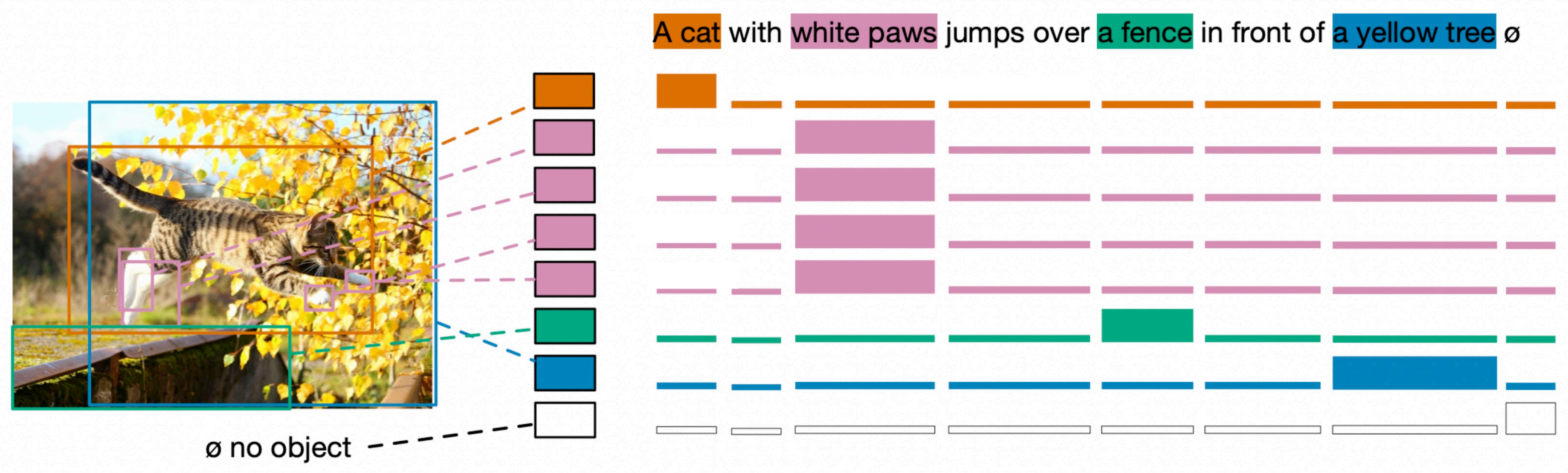
MDETR - Modulated Detection for End-to-End Multi-Modal Understanding. Kamath et al., 2021

Object Grounding

Soft token prediction

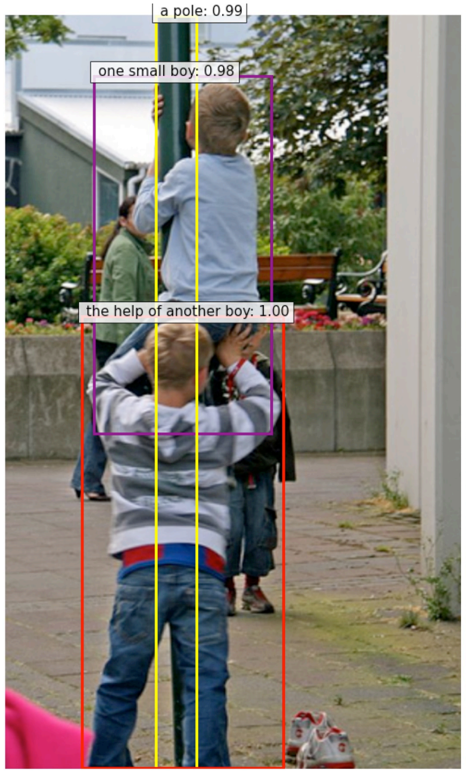
- For each detected bounding, predict a probability distribution over the tokens in the input phase

maximum number of tokens: 256

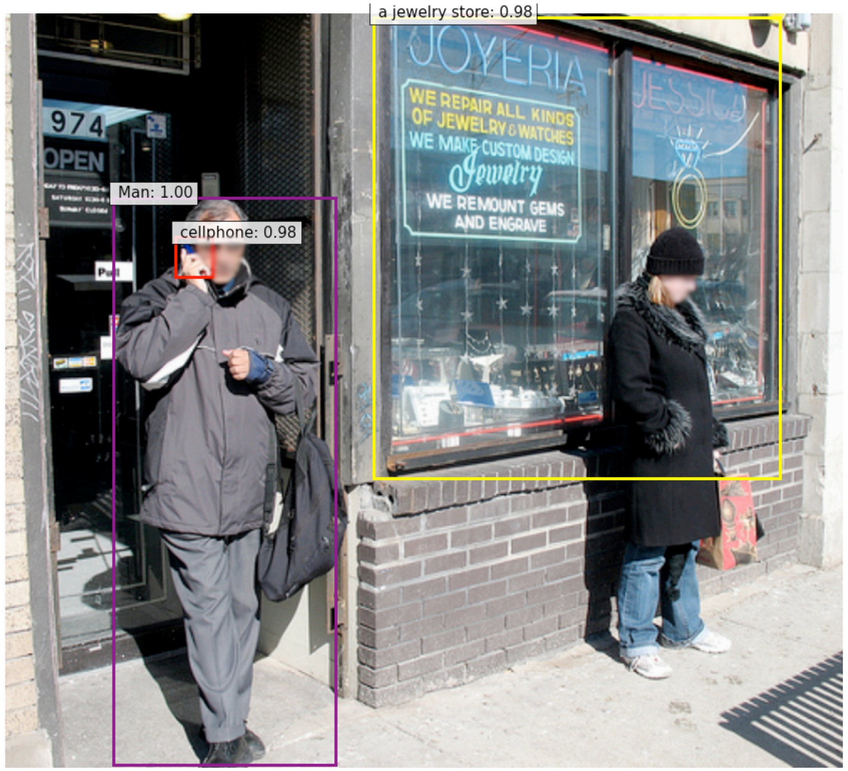


MDETR - Modulated Detection for End-to-End Multi-Modal Understanding. Kamath et al., 2021

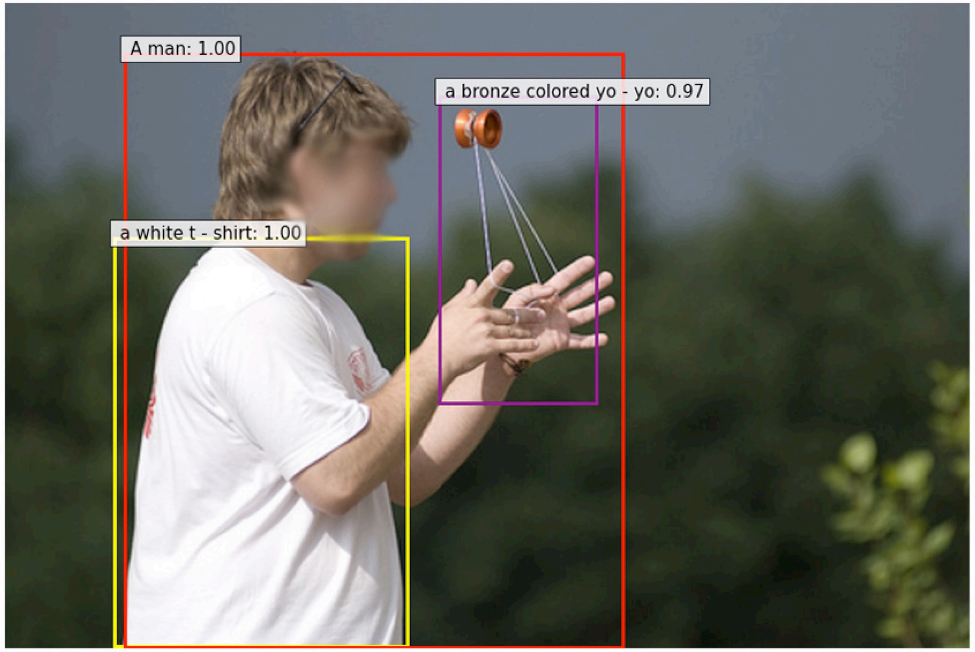
Object Grounding



(a) "one small boy climbing a pole with the help of another boy on the ground"



(b) "A man talking on his cellphone next to a jewelry store"



(c) "A man in a white t-shirt does a trick with a bronze colored yo-yo"

MDETR - Modulated Detection for End-to-End Multi-Modal Understanding. Kamath et al., 2021

Visual Question Answering

What color are her eyes?
What is the mustache made of?

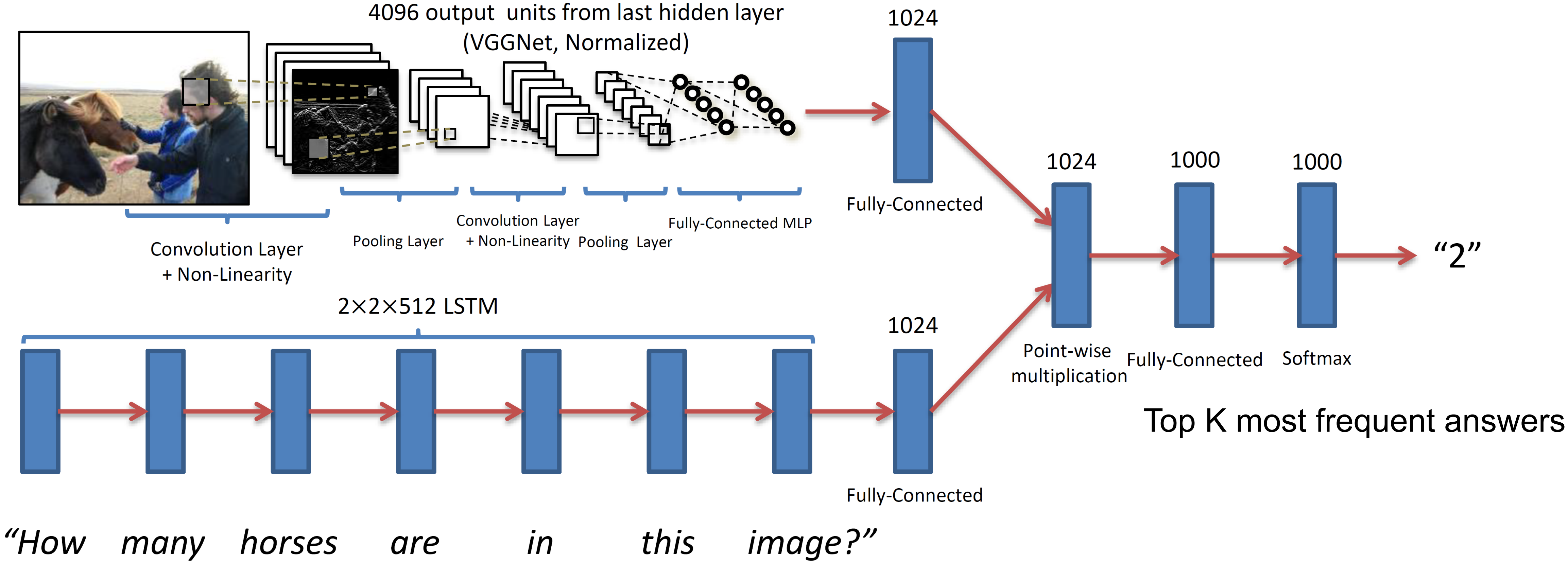
How many slices of pizza are there?
Is this a vegetarian pizza?

- Input
  - An image
  - A free-form, open-ended, natural language question
- Output
  - Case 1: open-ended answer
  - Case 2: multiple-choice task

$$\text{accuracy} = \min\left(\frac{\text{\# humans that provided that answer}}{3}, 1\right)$$

VQA: Visual Question Answering. Agrawal et al., ICCV, 2015

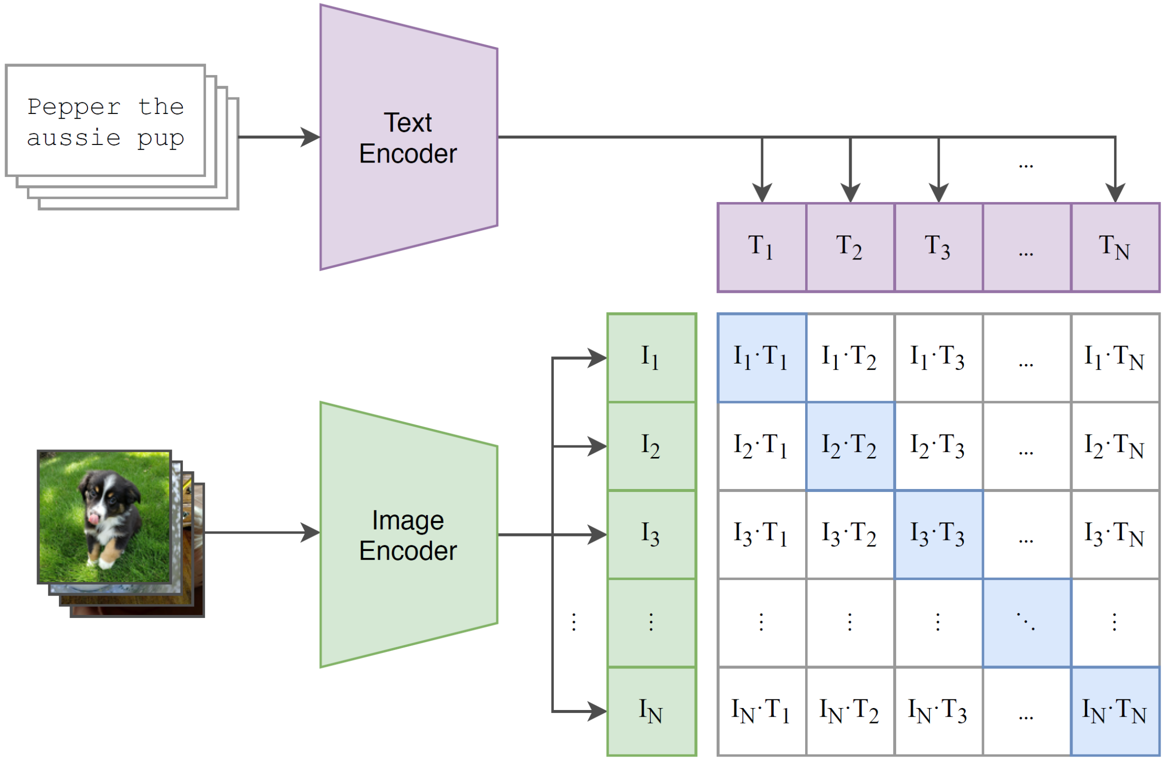
Visual Question Answering



VQA: Visual Question Answering. Agrawal et al., ICCV, 2015

CLIP: Contrastive Language-Image Pre-Training

Contrastive pre-training: representation learning

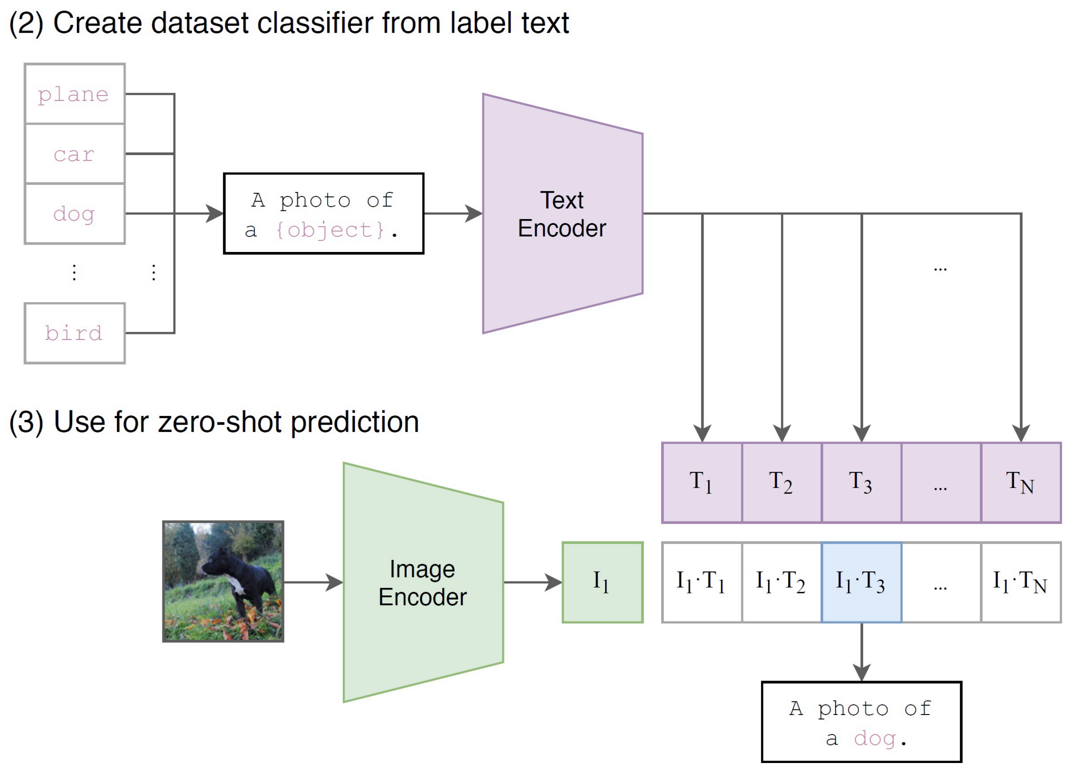


- 400 million (image, text) pairs from Internet

Learning Transferable Visual Models From Natural Language Supervision. Radford, et al., 2021

CLIP: Contrastive Language-Image Pre-Training

Zero-shot classification (no training on target datasets)



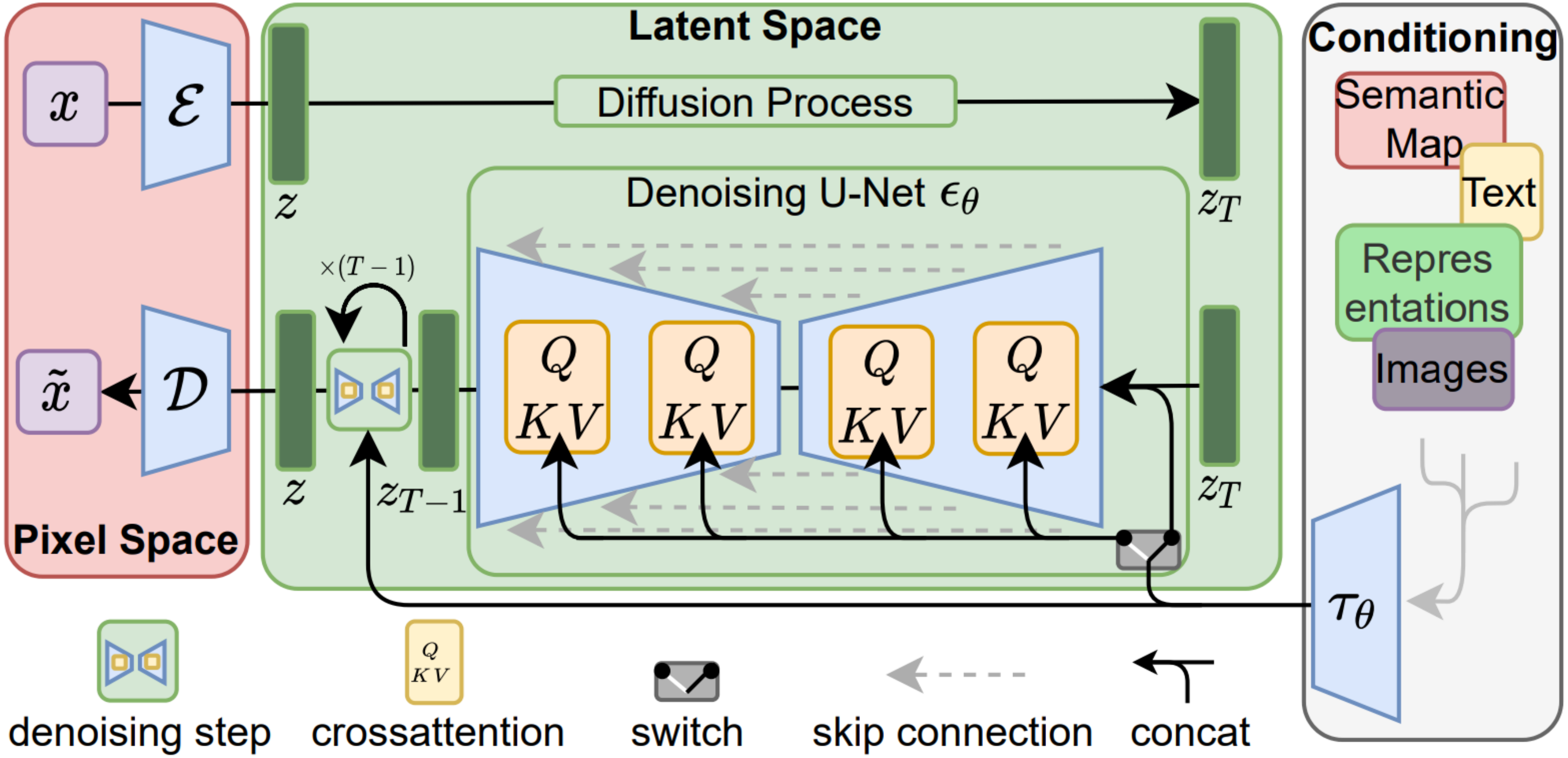
Learning Transferable Visual Models From Natural Language Supervision. Radford, et al., 2021

Text2Image

'A street sign that reads "Latent Diffusion" ' | 'A zombie in the style of Picasso' | 'An image of an animal half mouse half octopus' | 'An illustration of a slightly conscious neural network' | 'A painting of a squirrel eating a burger' | 'A watercolor painting of a chair that looks like an octopus' | 'A shirt with the inscription: "I love generative models!" '

High-Resolution Image Synthesis with Latent Diffusion Models. Rombach et al., CVPR, 2022.

Stable Diffusion



High-Resolution Image Synthesis with Latent Diffusion Models. Rombach et al., CVPR, 2022.

Summary

Vision + language tasks

- Image captioning
- Object/phase grounding
- Visual question answering
- Image-text retrieval
- Text2Image
- …

Representation learning (Pre-training)

- Learning image-text representations from large numbers (image, text) pairs
- Fine-turning for downstream tasks


What are in the video?
A group of singing birds
pine cone
bird
tree

Human: Multisensory Perception

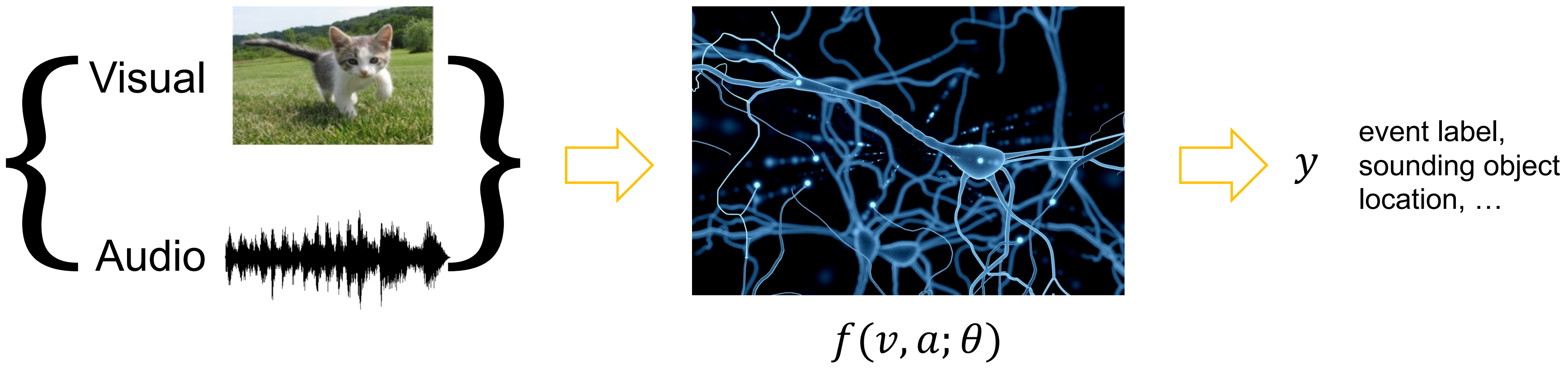
- We live in a multisensory world
- What we see can help us listen, what we hear can help us see
- Humans unconsciously integrate information from different modalities in daily perception experience


the McGurk Effect [McGurk and MacDonald, 1976]

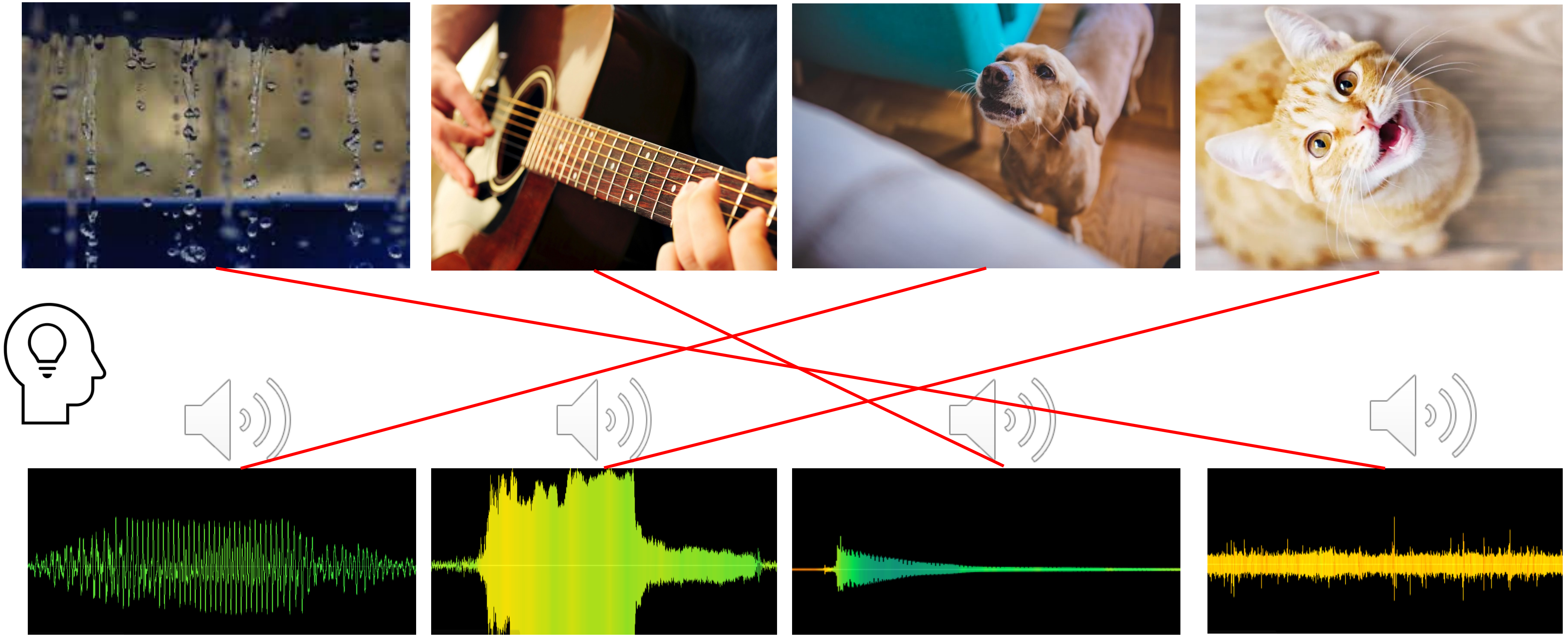
Video Credit: https://www.youtube.com/watch?v=2k8fHR9jKVM

Computational Multisensory Perception

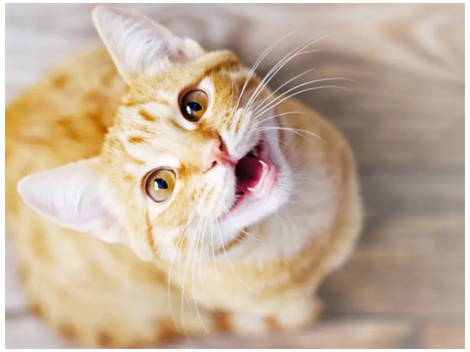
- Learn functions (e.g., neural networks) to model and understand auditory and visual inputs

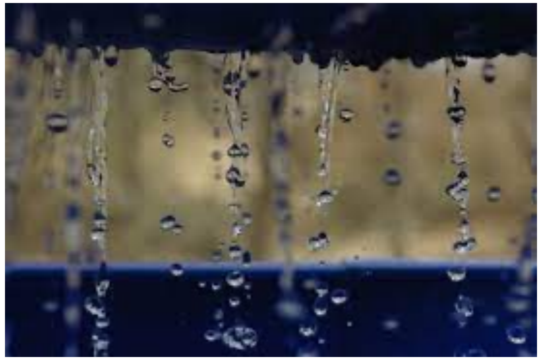


Audio-Visual Matching Puzzle

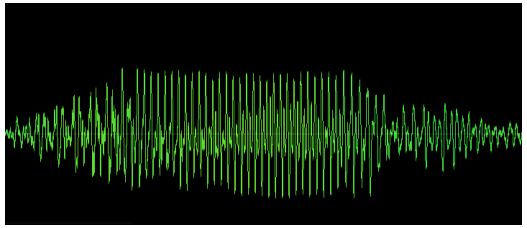


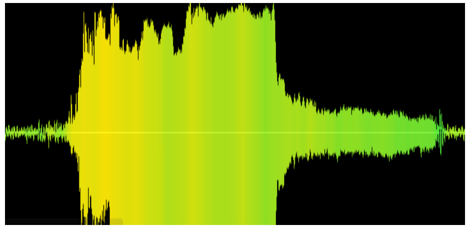
Data Prior: Natural Semantic Correspondence

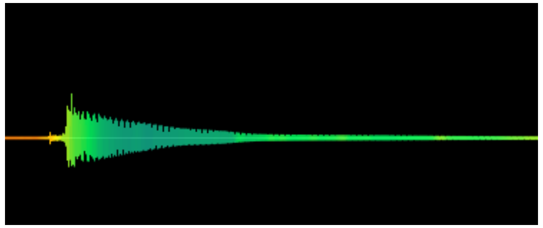


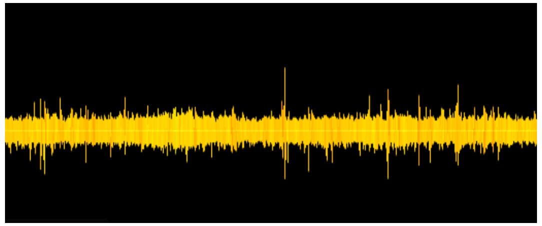


Woof **Meow** **Guitar sound** **Drizzle**









Both sound and sight carry semantic information

Data Prior: Natural Temporal Synchronization

The two modalities carry temporally aligned content.

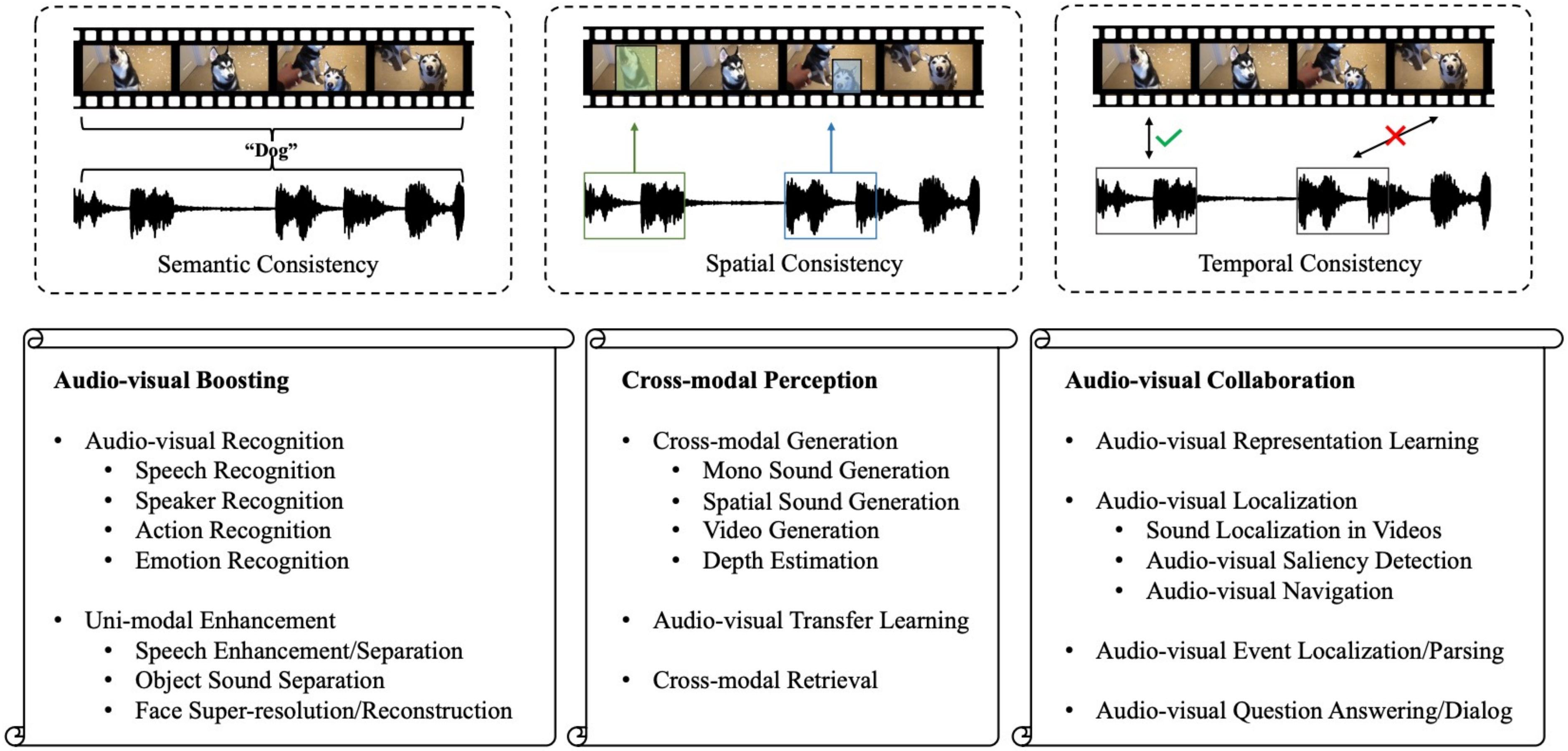
https://www.youtube.com/watch?v=2k8fHR9jKVM

Data Prior: Natural Spatial Correspondence

Spatial audio can indicate sound source locations

Morgado et al. 2018

Vision + Audio



Learning in Audio-visual Context: A Review, Analysis, and New Perspective. Wei et al., ArXiv, 2022.

Vision + Audio

Audio-visual sound separation

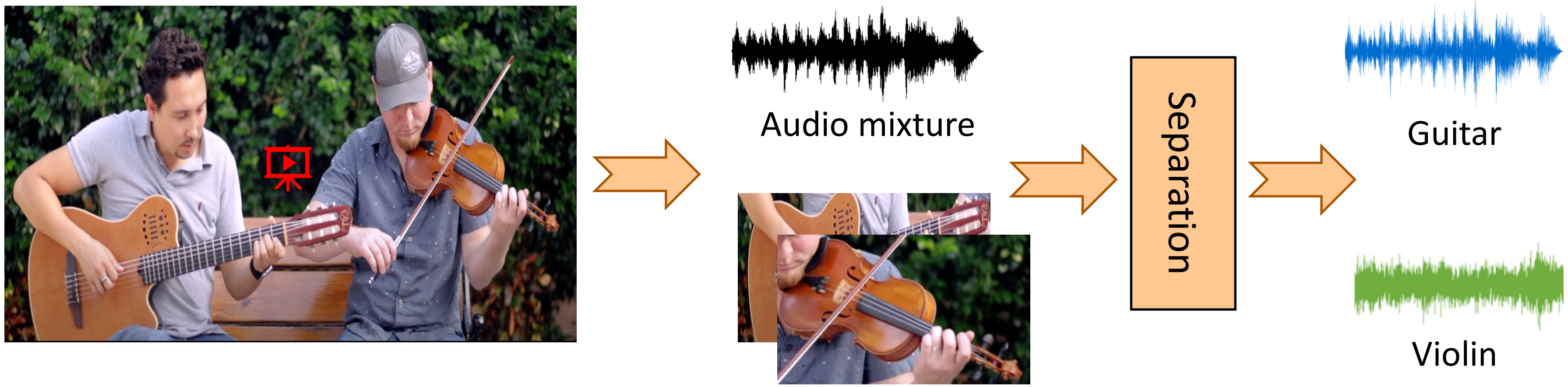
Sounding object localization

Audio-visual video parsing

Cross-model generation

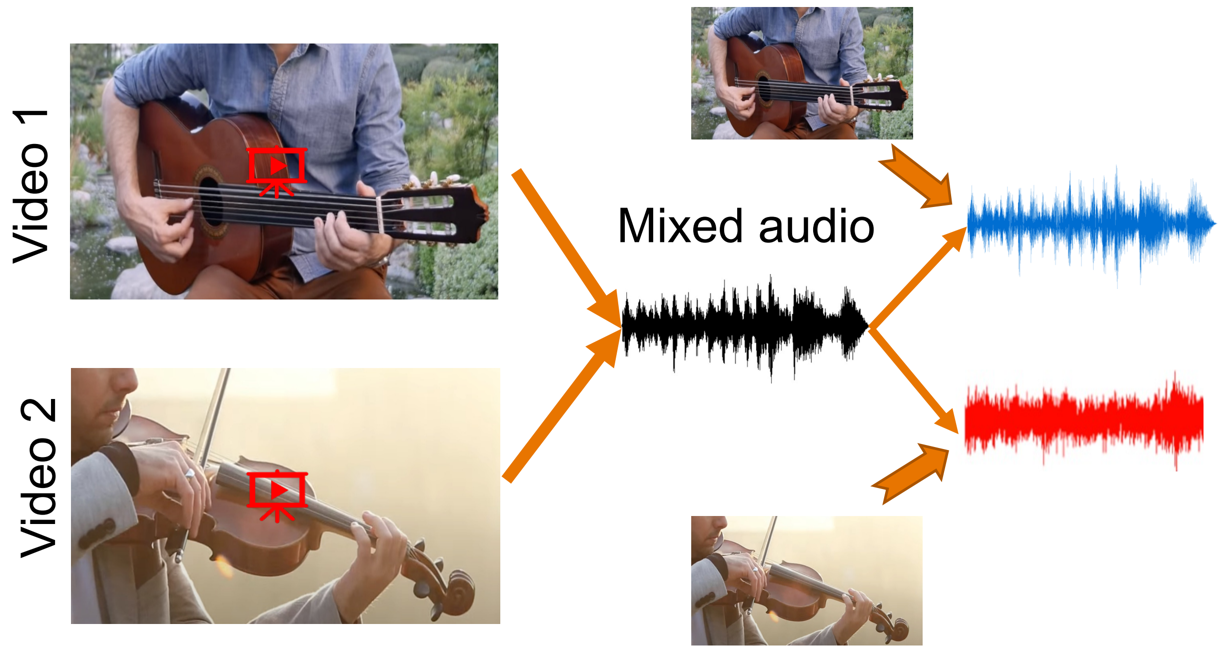
…

Audio-Visual Sound Separation



- Separate individual sounds from the audio mixture
- Incorporate visual scenes as the separation condition

Current Approaches: Mix-and-Separation

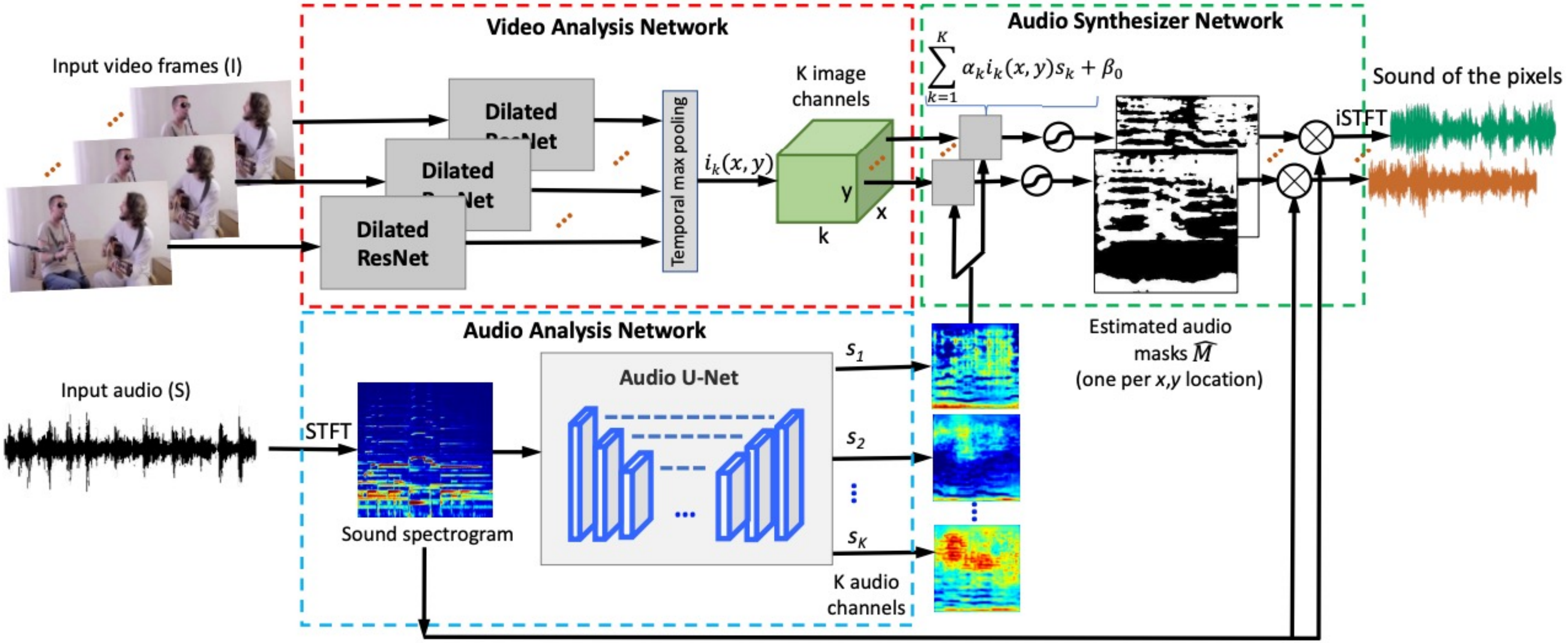


Assumptions:

- Single-source training video clips
- All visual objects are sounding

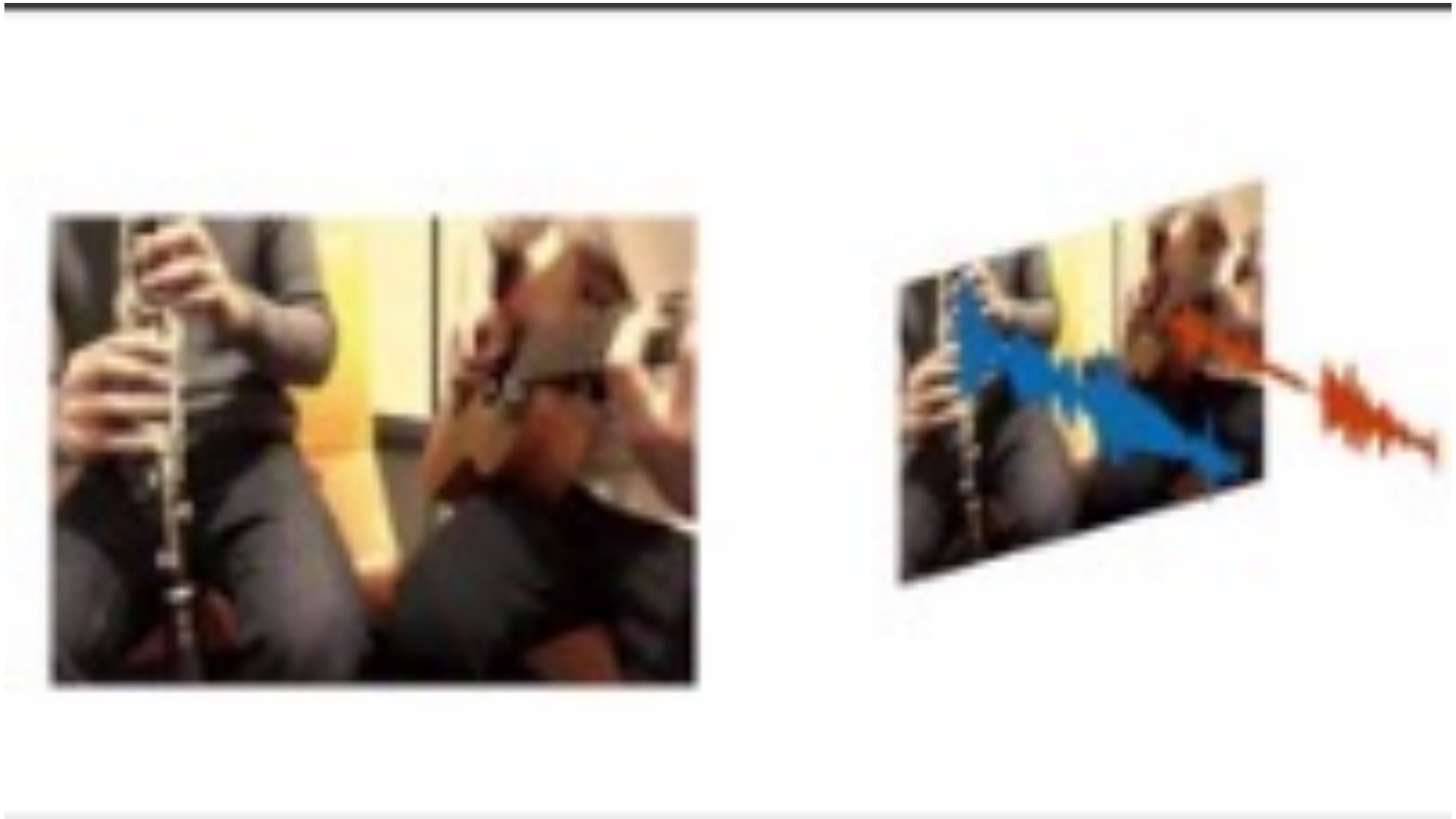
[Ephrat et al. 2018; Owens & Efros 2018 ; Zhao et al. 2018; Afouras et al. 2018; Gao & Grauman 2019; Gan et al. 2020]

Sound of Pixels



Sound of Pixels. Zhao et al., ECCV, 2018.

Sound of Pixels

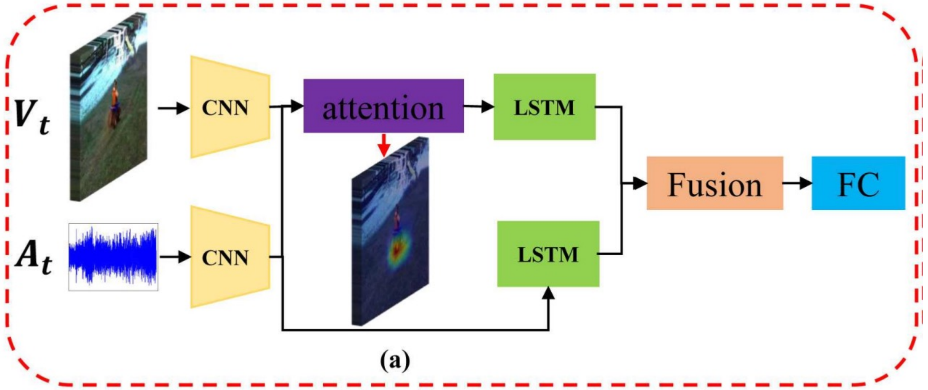


https://www.youtube.com/watch?v=2eVDLEQIKD0

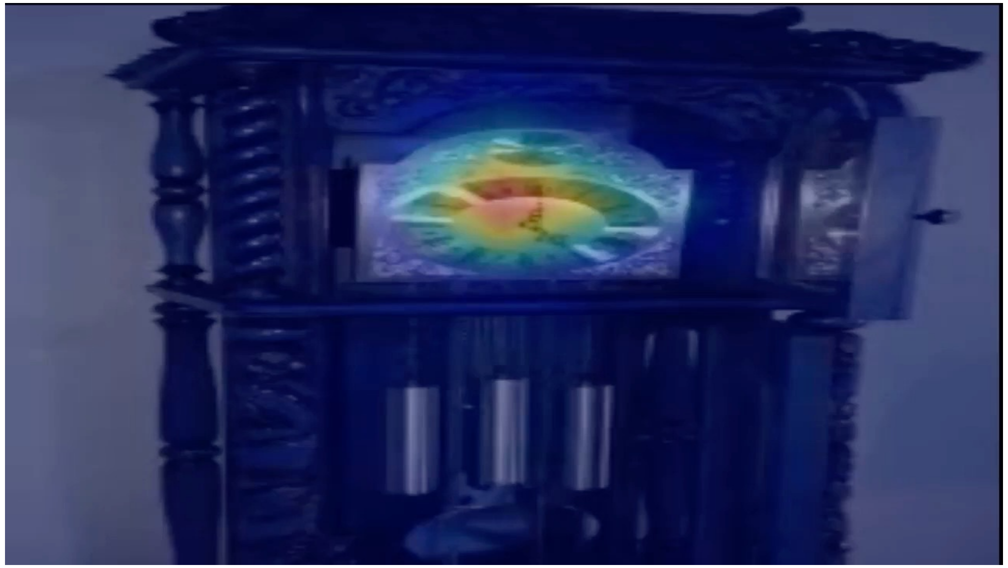
Sounding Object Localization

Spatially localize sound sources in video frames

Sounding Object Localization



Utilize audio-visual cross-modal attention to capture sounding objects in video frames



Localization results

Audio-Visual Event Localization in Unconstrained Videos. Tian et al., ECCV, 2018.

Universal Video Scenes

Videos contain various and diverse temporal video events, which are either audible (audio event), visible (visual event), or both (audio-visual event)

Audio Event: *Speech*
Visual Event: *Dog*

Visual Event: *Lawn mower*

Audio-Visual Event: *Basketball*

Questions for Understanding Video Scenes

These audio-visual examples are ubiquitous, which leads us to some basic questions

What events are in a video?

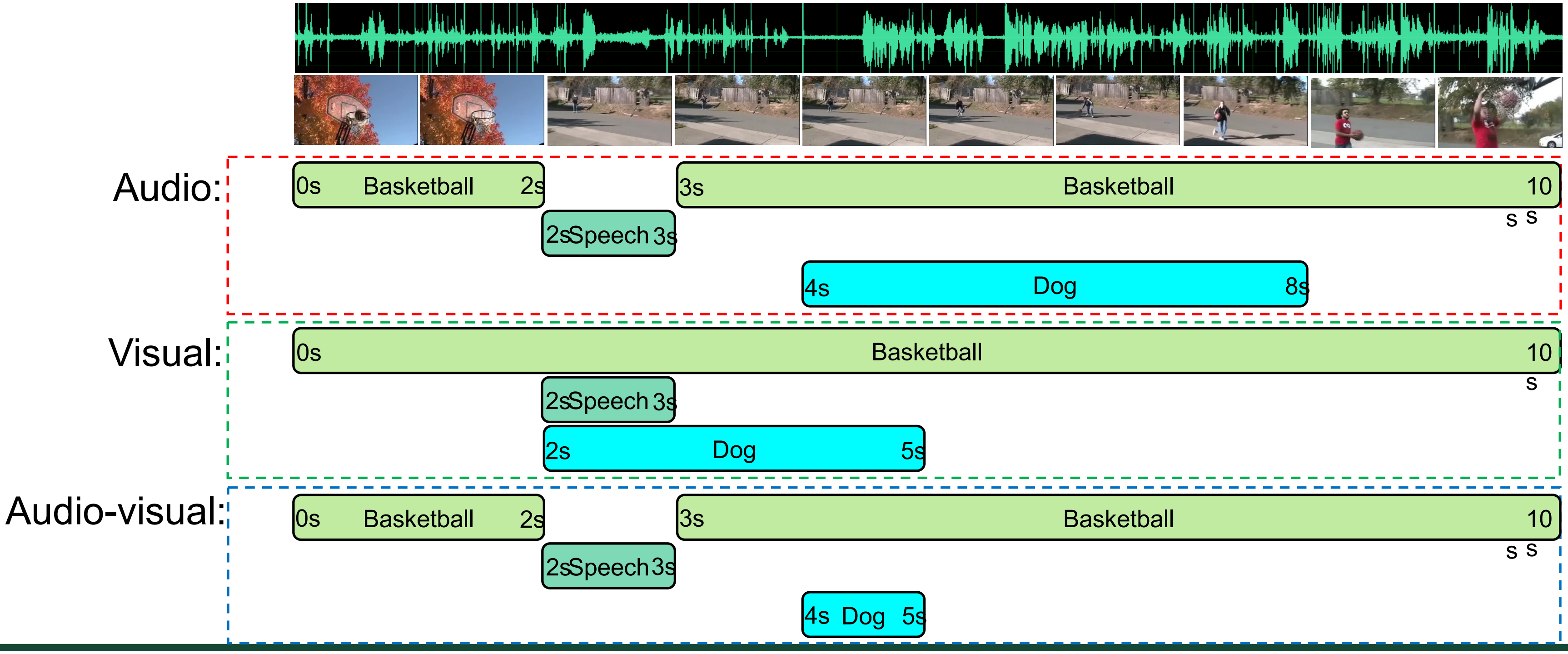
Which modalities perceive the events?

Where are these events?

How can we effectively detect them?

Modality-Aware Scene Understanding

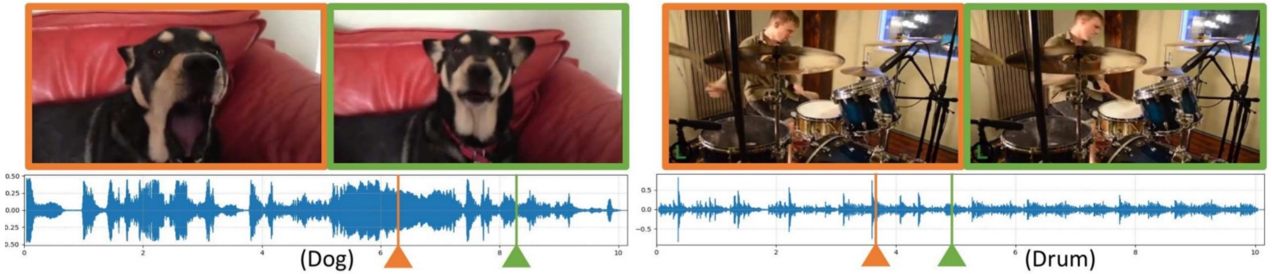
Audio-visual video parsing - recognizes *event categories* bind to *sensory modalities*, and meanwhile, finds *temporal boundaries* of when such an event starts and ends.

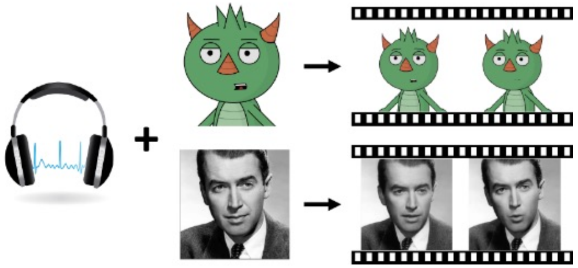


Unified Multisensory Perception: Weakly-Supervised Audio-Visual Video Parsing. Tian et al., ECCV 2020

Cross-Modal Generation

- Visual to sound generation
- Audio-driven visual generation (e.g., talking face)



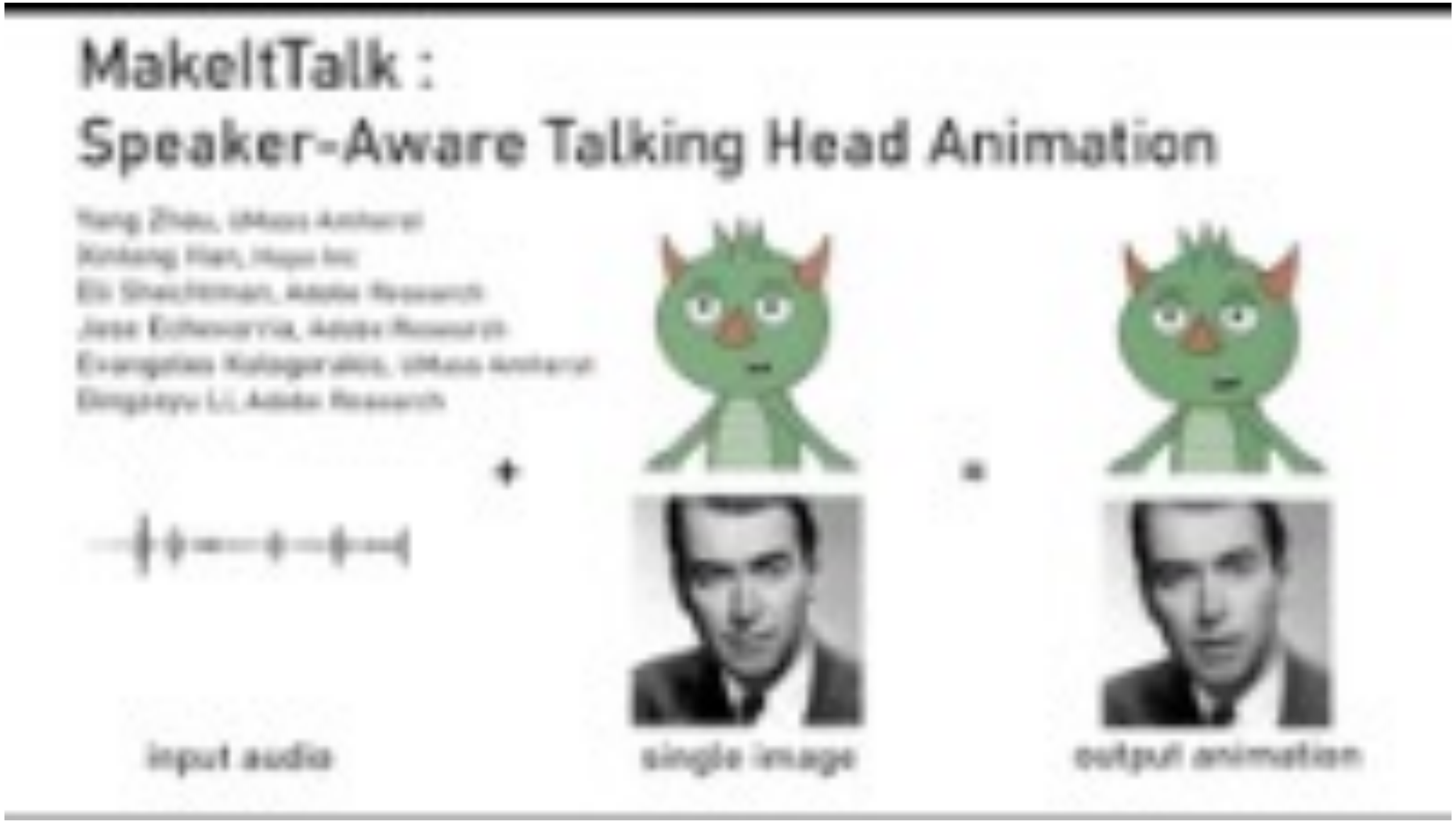


Visual to Sound: Generating Natural Sound for Videos in the Wild. Zhou et al., CVPR, 2018.
MakeItTalk: Speaker-Aware Talking-Head Animation. Zhou et al., SIGGRAPH Asia, 2020.

Visual to Sound

https://www.youtube.com/watch?v=Kgy919U295c

Audio to Visual: Talking Head Generation



https://www.youtube.com/watch?v=vUMGKASgbf8

Further Reading

Deep Visual-Semantic Alignments for Generating Image Descriptions, 2015 https://arxiv.org/abs/1412.2306

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, 2015 https://arxiv.org/abs/1502.03044

MDETR - Modulated Detection for End-to-End Multi-Modal Understanding, 2021 https://arxiv.org/abs/2104.12763

VQA: Visual Question Answering, 2015 https://arxiv.org/abs/1505.00468

Learning Transferable Visual Models From Natural Language Supervision, 2021 https://arxiv.org/abs/2103.00020

Sound of Pixels, 2018 http://sound-of-pixels.csail.mit.edu/

Audio-Visual Event Localization in Unconstrained Videos, 2018 https://openaccess.thecvf.com/content_ECCV_2018/papers/Yapeng_Tian_Audio-Visual_Event_Localization_ECCV_2018_paper.pdf

Unified Multisensory Perception: Weakly-Supervised Audio-Visual Video Parsing, 2020 https://arxiv.org/pdf/2007.10558.pdf

Visual to Sound: Generating Natural Sound for Videos in the Wild, 2018 https://arxiv.org/abs/1712.01393

MakeItTalk: Speaker-Aware Talking-Head Animation, 2020. https://arxiv.org/abs/2004.12992