max planck institut informatik

SIC Saarland Informatics Campus

High Level Computer Vision

Some Recent Trends: SAM & MAE & ImageBind

@ July 12, 2023

Bernt Schiele

cms.sic.saarland/hlcvss23/

Max Planck Institute for Informatics & Saarland University,
Saarland Informatics Campus Saarbrücken

Overview of Today's Lecture

- Segment Anything Model (SAM)
  - [arxiv'23] - https://arxiv.org/abs/2304.02643
  - https://segment-anything.com/
- Masked Autoencoders (MAEs) are Scalable Vision Learners
  - [cvpr'22] - https://arxiv.org/abs/2111.06377
- ImageBind: One Embedding Space to Bind them All
  - [cvpr'23] - https://arxiv.org/abs/2305.05665

Segment Anything

slides credit:

Alexander Kirillov @ Meta AI

slide credit: Alexander Kirillov

Research Team


Alexander Kirillov


Eric Mintun


Nikhila Ravi


Hanzi Mao


Chloe Rolland


Laura Gustafson


Tete Xiao


Spencer Whitehead


Alex Berg


Wan-Yen Lo


Piotr Dollar


Ross Girshick

Project Contributors (alphabetical):

Aaron Adcock, Vaibhav Aggarwal, Morteza Behrooz, Cheng-Yang Fu, Ashley Gabriel, Ahuva Goldstand, Allen Goodman, Sumanth Gurram, Jiabo Hu, Somya Jain, Devansh Kukreja, Robert Kuo, Joshua Lane, Yanghao Li, Lilian Luong, Jitendra Malik, Mallika Malhotra, William Ngan, Omkar Parkhi, Nikhil Raina, Dirk Rowe, Neil Seejoor, Vanessa Stark, Bala Varadarajan, Bram Wasti, Zachary Winstrom

slide credit: Alexander Kirillov

Deep learning paradigm shift

[2012 – 2022] ⟶ [2022 – present]

- **Apply DL to task X**
- End-to-end
- Step change on **specific tasks** and **specific data** distributions
- Adaptation via **fine-tuning only**

- **Combine foundation/base models for X, Y, Z to do tasks Q, R, S**
- Powerful **components** (LLMs, CLIP …)
- Step change on **broad tasks** and **broad data** distributions
- Adaptation without parameter updates (**prompting**)

slide credit: Alexander Kirillov

Broadly applicable solution for segmentation

- The Segment Anything project / Segment Anything Model (SAM)
  - A component (SAM) that broadly solves segmentation
  - Useful to both humans and machines (i.e., a component in a system)
  - Broad generalization to new domains and use cases / tasks
  - Adaptation primarily via prompting

slide credit: Alexander Kirillov

Segment Anything Project Goal

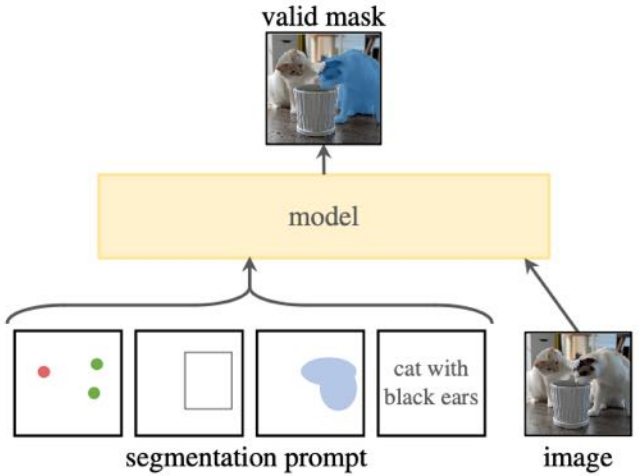
- Develop a **promptable model** and pre-train it on a broad **dataset** using a **task** that enables powerful generalization
  - Model: *Doesn't exist!*
  - Dataset: *Doesn't exist!*
  - Task: *Doesn't exist!*

- Model/data/task are highly coupled

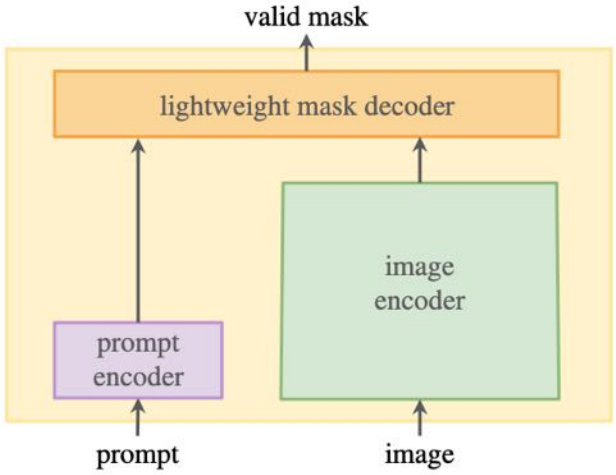
- We need a comprehensive solution

slide credit: Alexander Kirillov

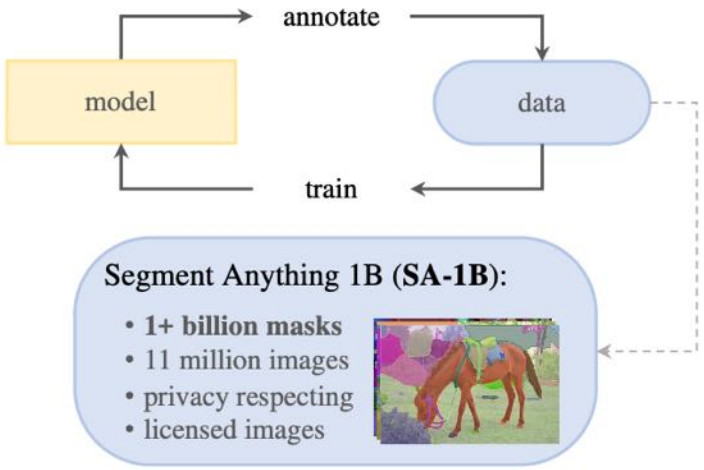
Segment Anything: Task, Model, Data



(a) **Task**: promptable segmentation



(b) **Model**: Segment Anything Model (**SAM**)



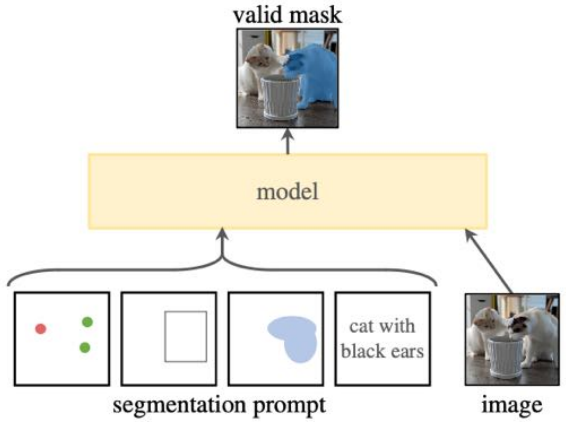
(c) **Data**: data engine (top) & dataset (bottom)

slide credit: Alexander Kirillov

Task: Promptable Segmentation

- Translate “prompt” from NLP to segmentation
  - Segmentation prompt: anything that specifies *what* to segment
  - E.g., foreground/background points, boxes, masks, text

- The task: Return a *valid* segmentation mask given *any prompt*



(a) **Task**: promptable segmentation

slide credit: Alexander Kirillov

Example: AR gaze-to-mask

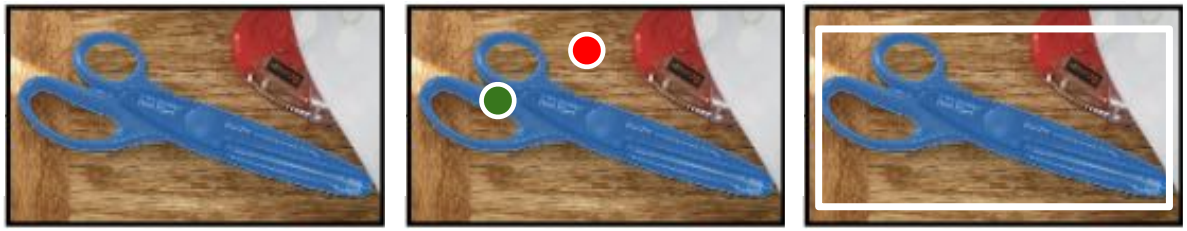
slide credit: Alexander Kirillov

slide credit: Alexander Kirillov

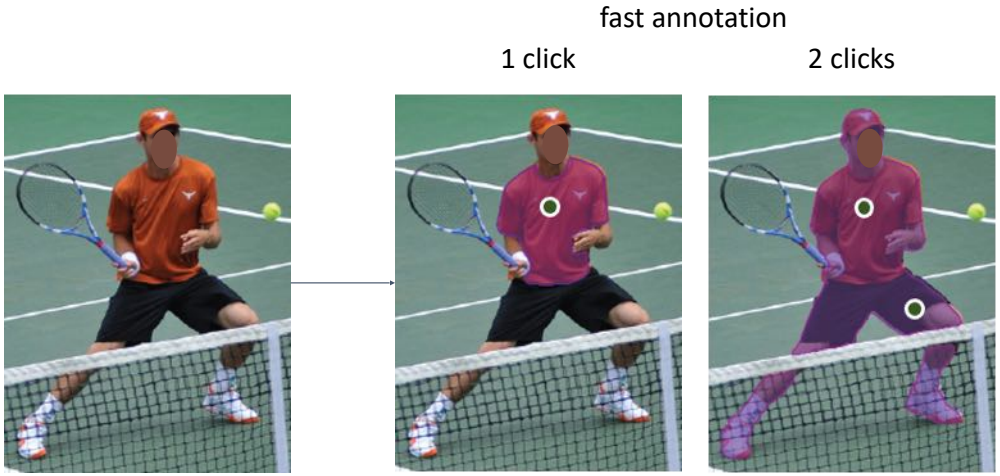
Why Promptable Segmentation?

- Composable with other models.

- Prompts easy to simulate at training.



- Enables human-in-loop annotation.

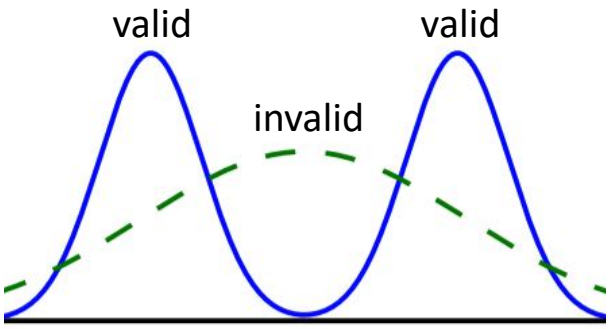


slide credit: Alexander Kirillov

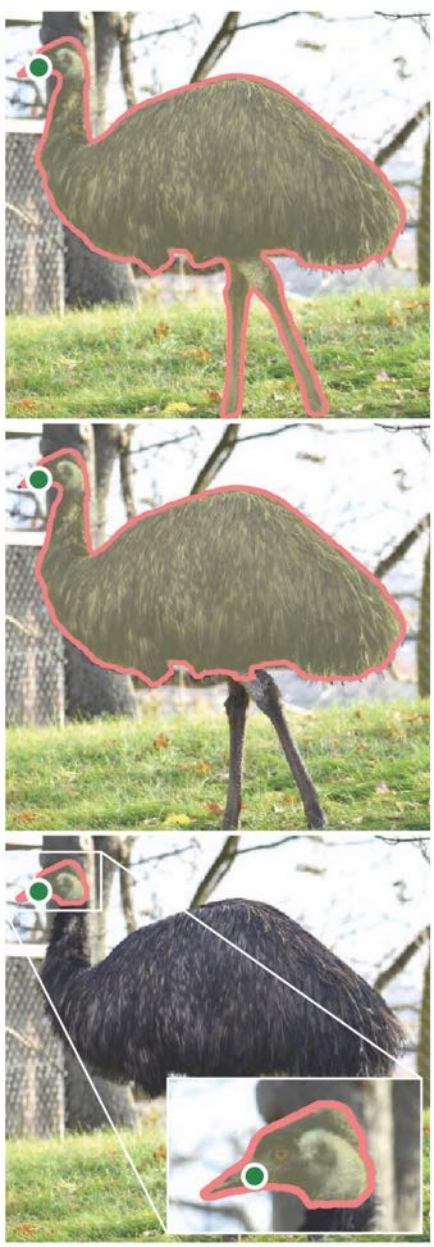
A “Valid” Mask?

- Prompts can be *ambiguous*
- Segmentation is *ambiguous*
- The task should train a model that handles ambiguity

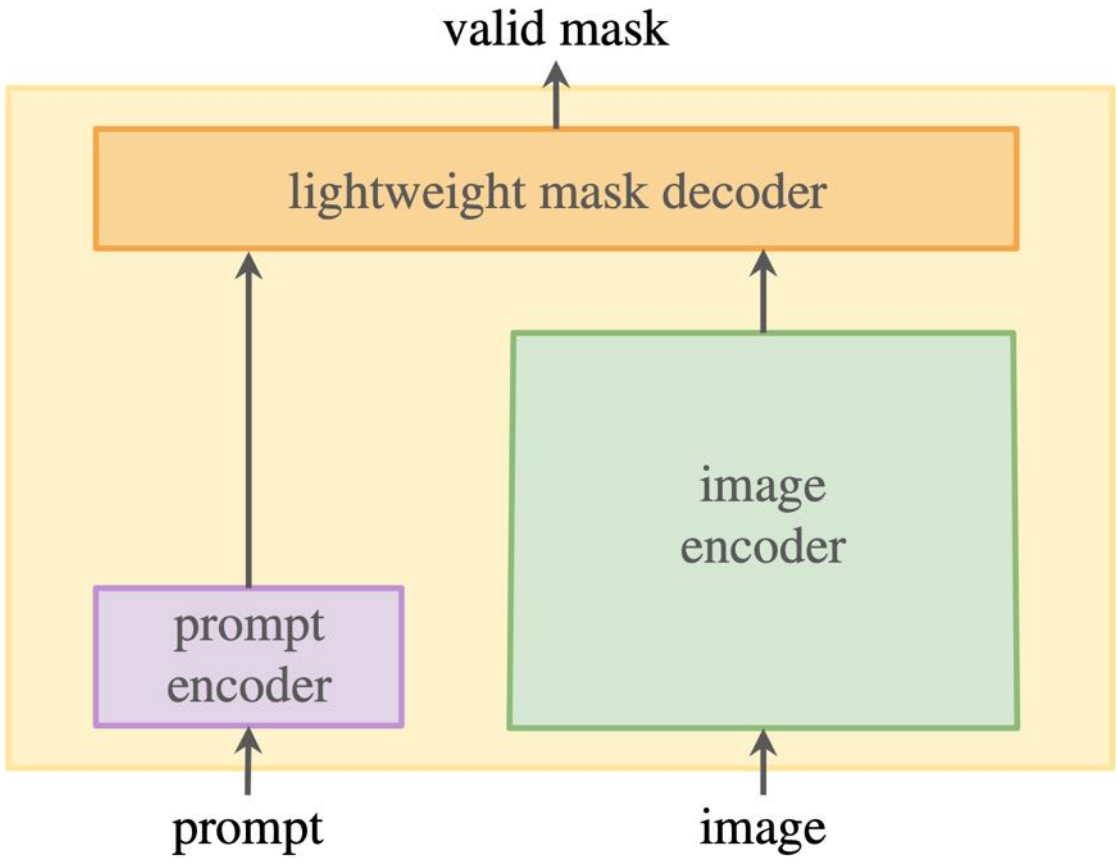
- Avoid “mode averaging”



Adapted from NIPS 2016 Tutorial: “Generative Adversarial Networks" by Ian Goodfellow

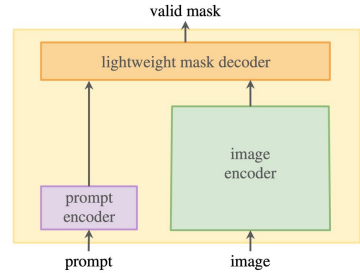


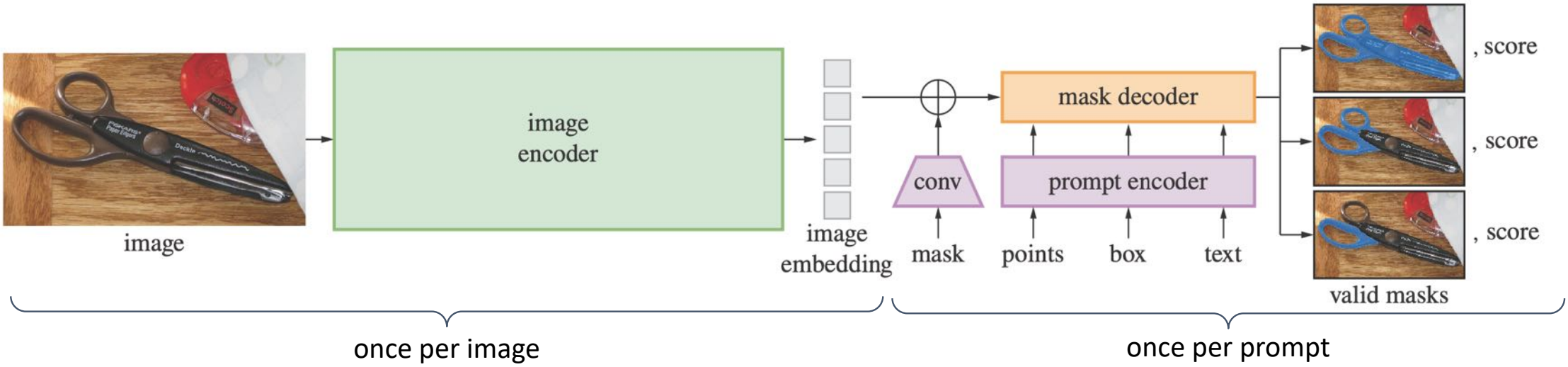
SAM: Model Overview



slide credit: Alexander Kirillov

SAM in More Detail





- **Image encoder:** ViT-H adapted for 1024x1024 input resolution
- **Prompt encoder:** Positional and label embeddings for points/boxes; CLIP for text
- **Mask decoder:** Lightweight transformer (~50ms on CPU)
- **Output:** 1 or 3 masks + model's prediction of mask quality

slide credit: Alexander Kirillov

SAM training

image encoder

x9

sample point

prompt encoder

mask decoder

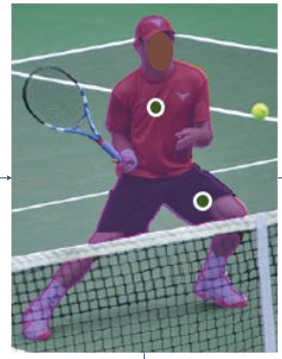
predict

sample point

prompt encoder

mask decoder

predict



total loss

dice + focal loss

dice + focal loss

slide credit: Alexander Kirillov

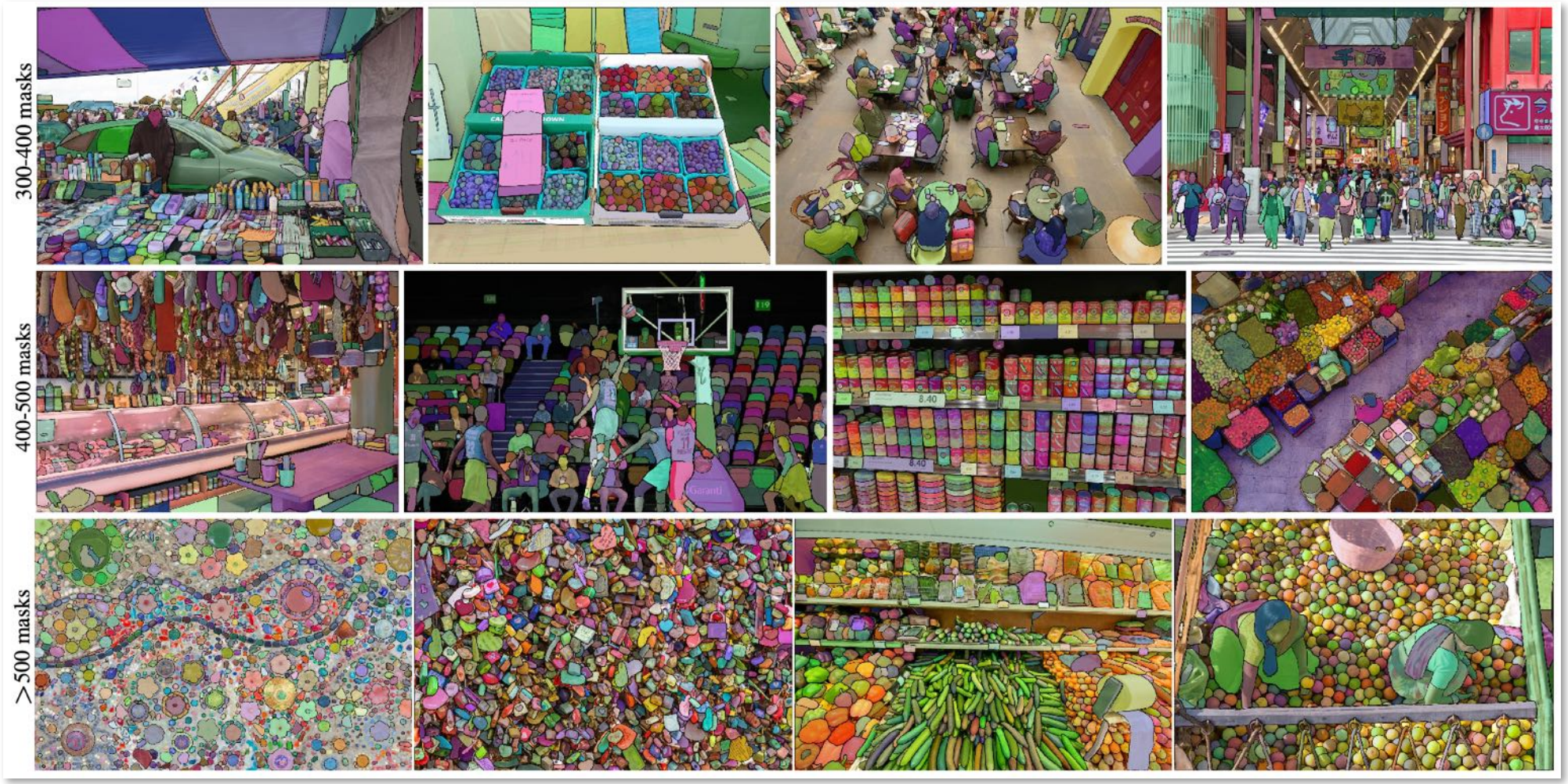
Transformer details

- Image encoder is ViT-H modified for large resolution (ViTDet):
  - initialized with MAE weights
  - Patch size 16
  - Size 14 windowed attention in all but three layers
  - Final 1x1 conv mapping ViT-H channel dim. (1280) to mask decoder dim. (256)
  - For 1024x1024 input image, output embeddings is 256x64x64
- Mask decoder
  - Large internal MLP dimension of 2048
  - Down-sampled q/k/v dimension of 128 in cross attention
  - Entire prompt token added as positional encoding to q/k in attention every layer (DETR style)
  - After transformer, embeddings upscaled to 256x256 by 2 conv transposes of dim 64 and 32

slide credit: Alexander Kirillov

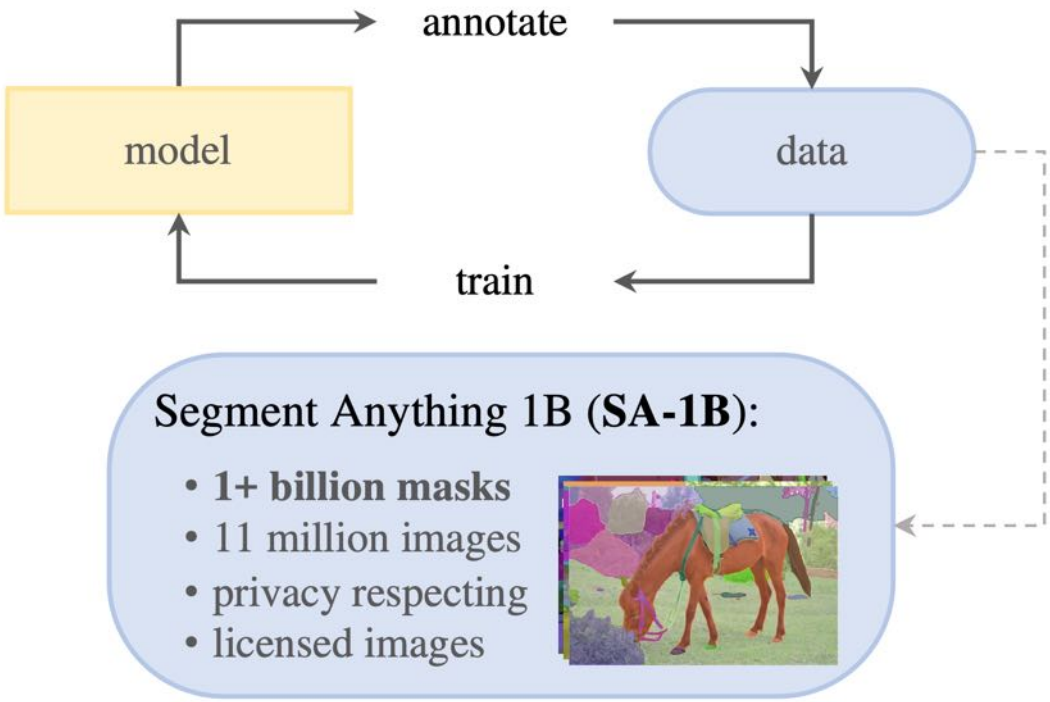
slide credit: Alexander Kirillov

Dataset Construction



Data Engine

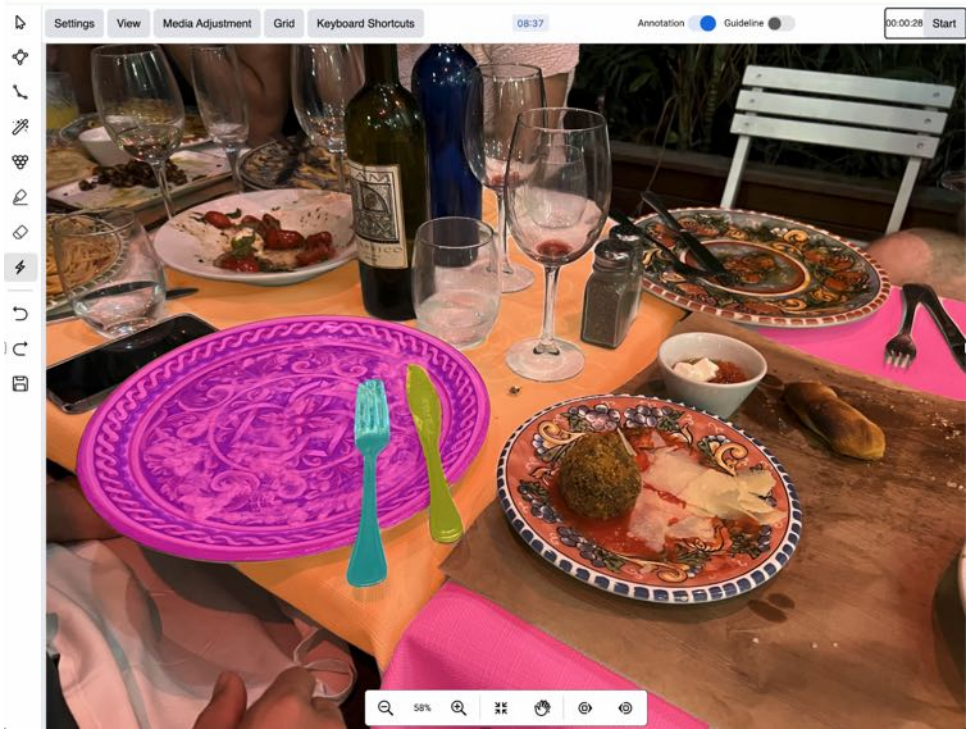
- Model-in-the-loop dataset construction
- Requires SAM to run in (amortized) real-time on CPU in a browser
- Three stages
  - *Assisted manual*,
  - *semi-manual*,
  - *fully automatic*



slide credit: Alexander Kirillov

Data Engine: Assisted Manual Stage

- SAM in a web-browser
  - Interactive segmentation tool
- 4.3 million masks (120k images)
- 6.5x faster than COCO masks
  - 2x slower than box annotation
- Retrained SAM 6 times
  - Scaled model size
  - Other improvements



slide credit: Alexander Kirillov

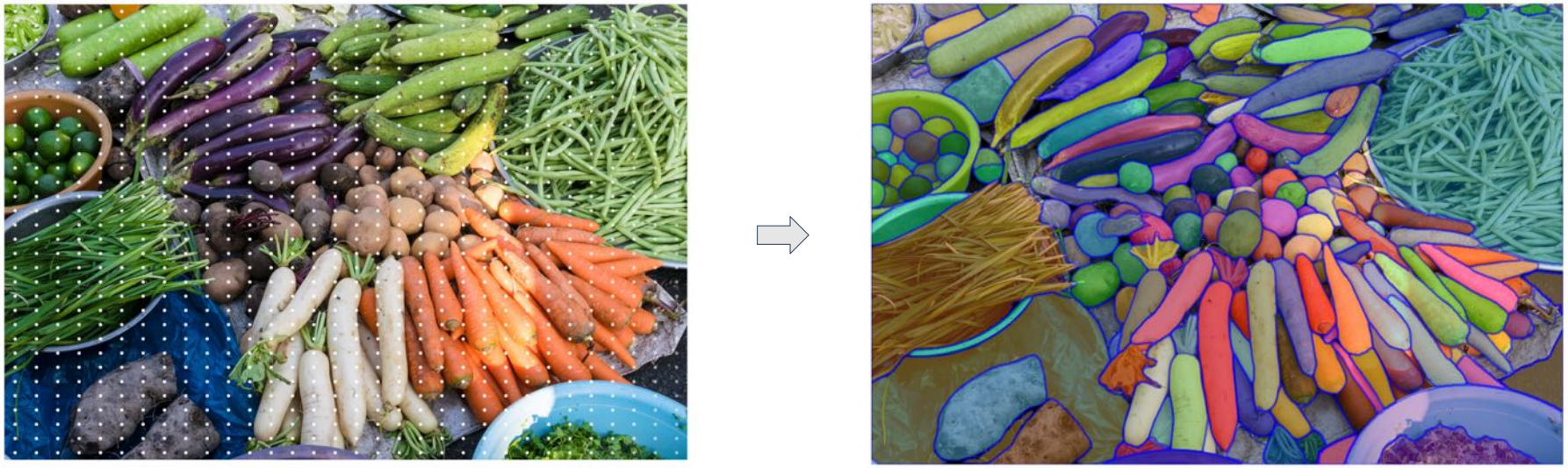
slide credit: Alexander Kirillov

Data Engine: Semi-Manual Stage

- Goal: increase diversity beyond most prominent objects
- Automatically segment many objects using detector + SAM
- Added 5.9 million more masks (180k more images)
- Retrained SAM 5 more times

Data Engine: Fully Automatic Stage

- We ran SAM in “segment everything” mode on 11 million images
- Generated 1.1 billion masks



slide credit: Alexander Kirillov

SA-1B Dataset

- 11 million images
- 1.1 **billion** masks
- All automatically generated by SAM
- 400x larger than previous largest segmentation dataset
- Quality evaluation
  - Estimated 94% have at least 90% IoU with professional masks
  - Prior work estimates human inter-annotator IoU at 85-91%

slide credit: Alexander Kirillov

SA-1B Examples (Automatically Labelled)

slide credit: Alexander Kirillov

SA-1B Examples (Automatically Labelled)

slide credit: Alexander Kirillov

SA-1B Examples (Automatically Labelled)

slide credit: Alexander Kirillov

> 500 masks

slide credit: Alexander Kirillov

Zero-Shot Transfer Tasks

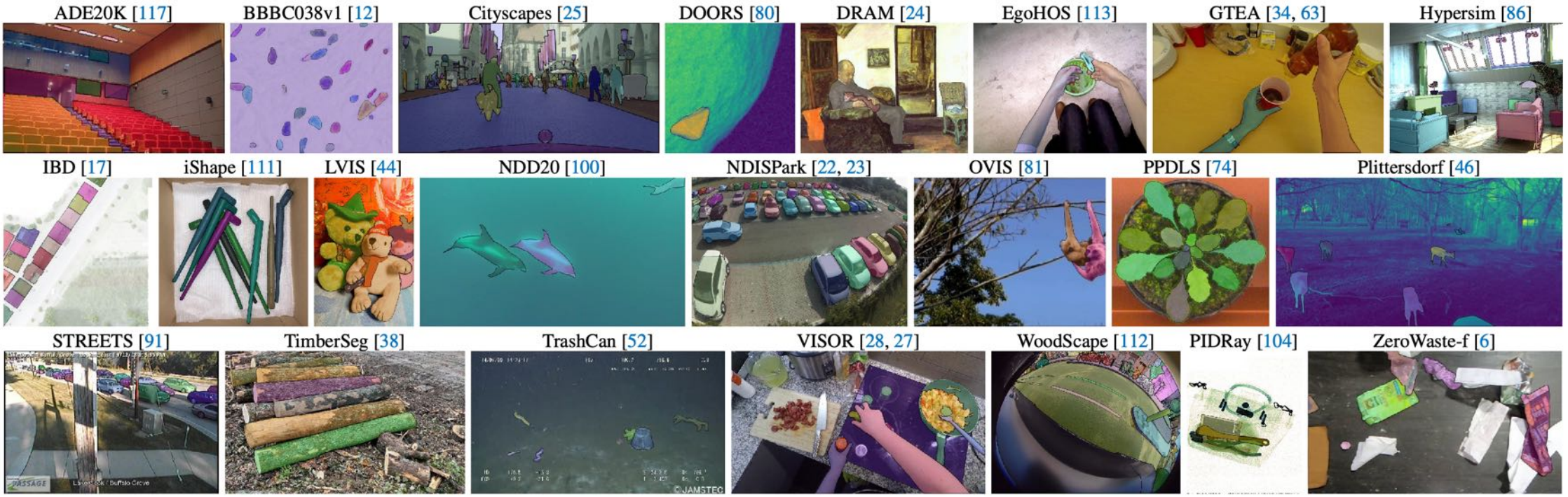
1. Single point valid mask evaluation
2. Edge detection
3. (Object proposals)
4. (Instance segmentation)
5. Text-to-mask

slide credit: Alexander Kirillov

slide credit: Alexander Kirillov

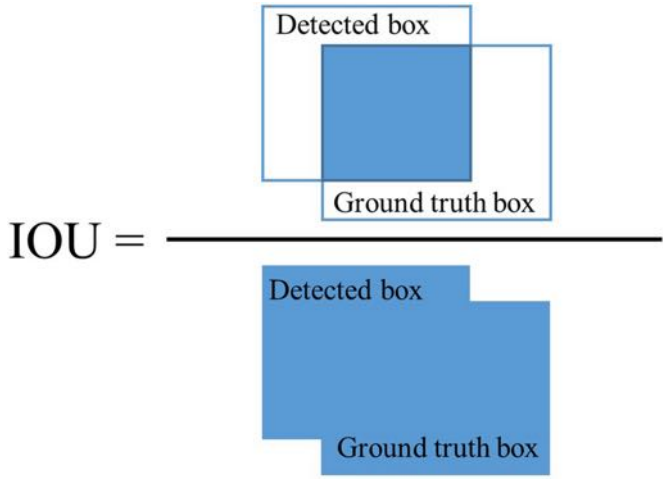
Single Point Valid Mask Evaluation

- SA-1B: mostly “scene” images (photographers walking in the world)
- 23 diverse dataset suite – mostly out-of-distribution vis-à-vis SA-1B



Single Point Valid Mask Evaluation Protocol

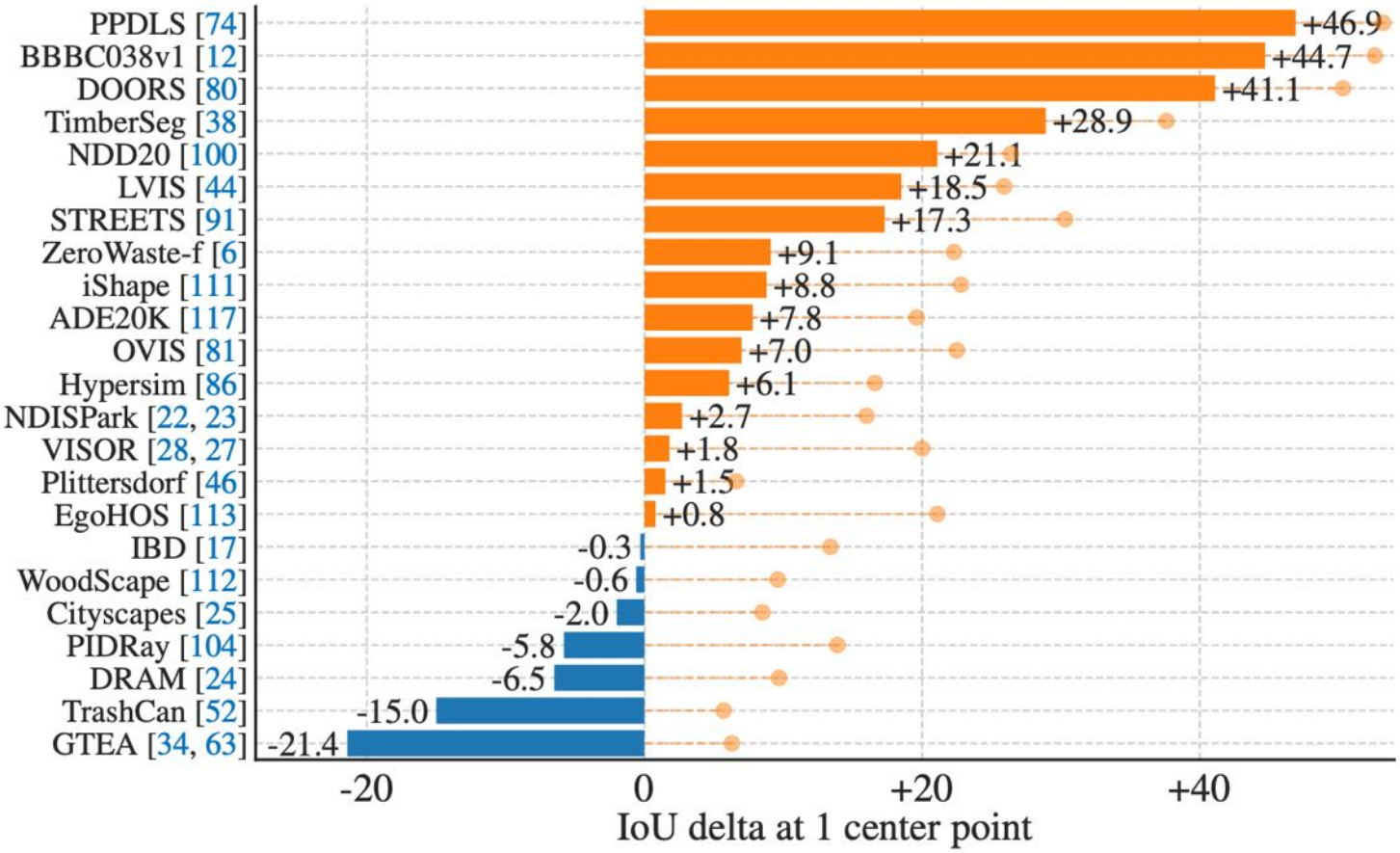
- Select g.t. mask from dataset
  - Sample point from mask
  - Prompt model with point, model predicts mask
- Compute IoU (prediction, g.t.)
  - Metric: mean IoU (mIoU) over mask
- Oracle evaluation
  - Recall SAM can predict 3 masks to handle ambiguity
  - Oracle evaluation: mask with best g.t. IoU is selected
  - Regular evaluation: model puts forth most confident mask



slide credit: Alexander Kirillov

Single Point Valid Mask Evaluation

- SAM vs. RITM
- Circle = oracle



slide credit: Alexander Kirillov

RITM: Konstantin Sofiiuk, Ilya A Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. ICIP, 2022

Edge Detection

⇨

- BSDS500
- Segment everything (regular grid of foreground points)
  → filter edges from masks → edge NMS

method	year	ODS	OIS	AP	R50
HED [108]	2015	.788	.808	.840	.923
EDETR [79]	2022	.840	.858	.896	.930
zero-shot transfer methods:					
Sobel filter	1968	.539	-	-	-
Canny [13]	1986	.600	.640	.580	-
Felz-Hutt [35]	2004	.610	.640	.560	-
SAM	2023	.768	.786	.794	.928

slide credit: Alexander Kirillov

Text-to-Mask

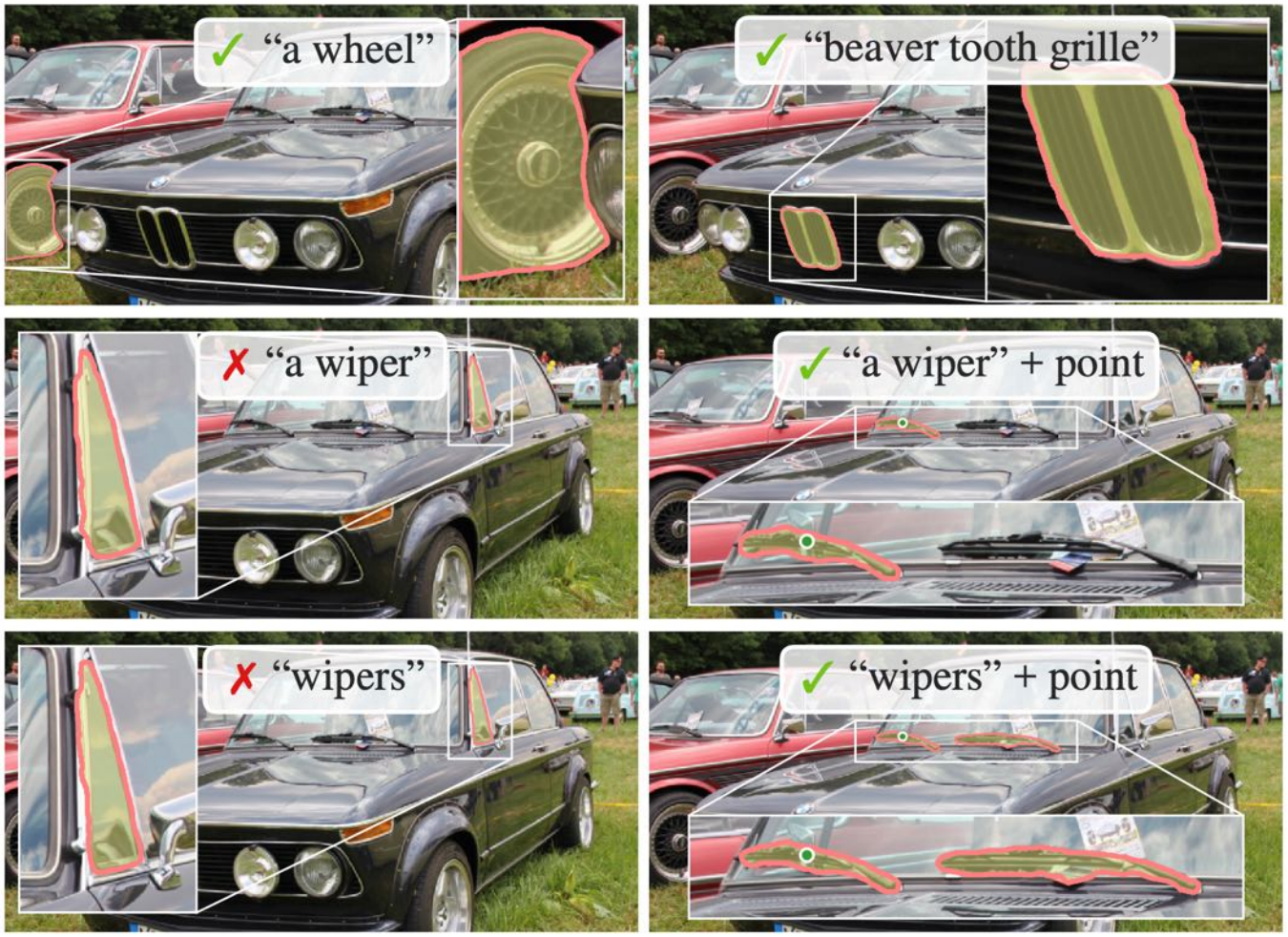
- Training:
  - For each mask (larger than 100x100), compute the CLIP image encoder embedding
  - Prompt with embedding during simulated interactions

- Recall in CLIP image and text embeddings are *aligned*

- During inference: prompt with text embedding

slide credit: Alexander Kirillov

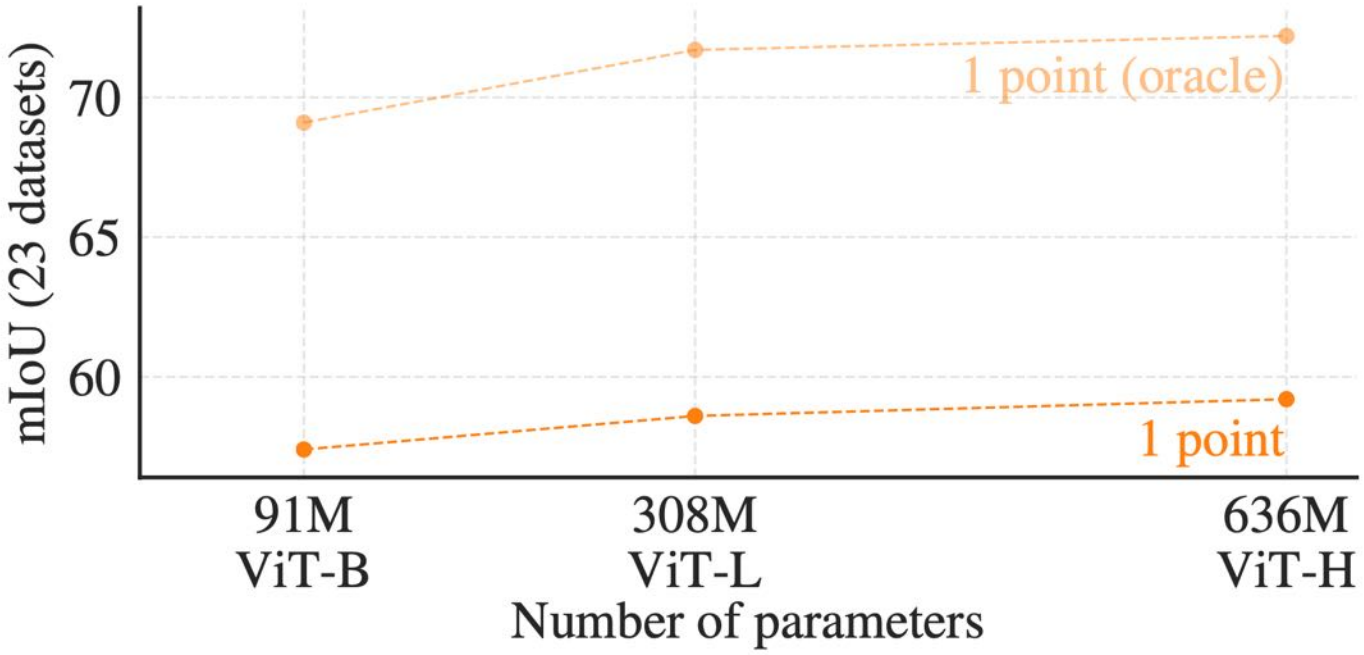
Text-to-Mask Qualitative Examples



slide credit: Alexander Kirillov

Ablations

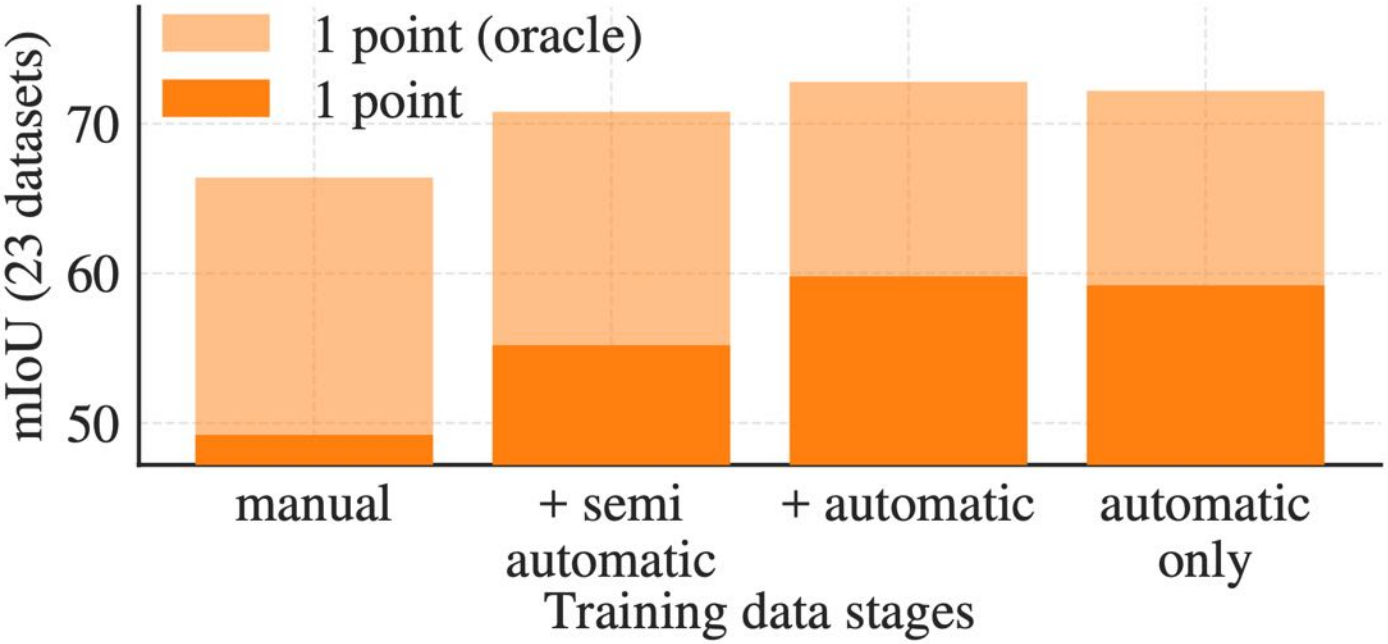
- Scaling saturation



slide credit: Alexander Kirillov

Ablations

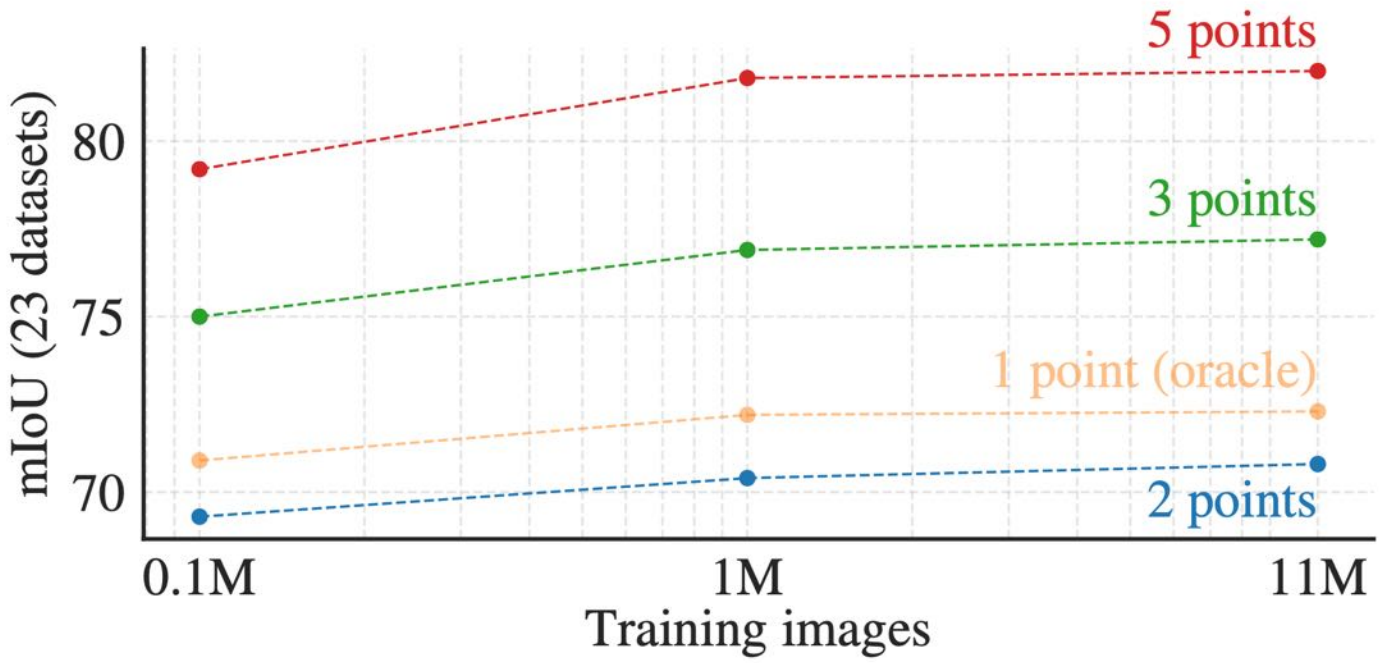
- Automatic data is good enough



slide credit: Alexander Kirillov

Ablations

- Training with 10% of the data is good



slide credit: Alexander Kirillov

Limitations

- Can miss fine structures
- Small “sprinkles” and holes (fixed with post-processing)
- Boundaries not as crisp as “zoom-in” approaches
- Interactive-only models may work better with many points
- Amortized real-time, but not real-time (with ViT-H anyway)
- Text-to-mask exploration is preliminary
- Domain-specialized tools may work better

slide credit: Alexander Kirillov

slide credit: Alexander Kirillov

Conclusion

- Main contributions
  - New task (yields pre-training and inference)
  - New model (SAM)
  - New dataset (SA-1B – 1 billion masks!)
- The SA Project: attempt at a base/foundation model for segmentation
  - Up to community use and evaluation to see if that is the case

- NLP:LLMs :: CV:???
  - Is there something as simple, general, and powerful for CV?

Links

- Project website: https://segment-anything.com
- Interactive demo: https://segment-anything.com/demo
- Dataset explorer: https://segment-anything.com/dataset/index.html
- Code (inference only for now): https://github.com/facebookresearch/segment-anything

slide credit: Alexander Kirillov

Overview of Today's Lecture

- Segment Anything Model (SAM)
  - [arxiv'23] - https://arxiv.org/abs/2304.02643
  - https://segment-anything.com/

- Masked Autoencoders (MAEs) are Scalable Vision Learners
  - [cvpr'22] - https://arxiv.org/abs/2111.06377

- ImageBind: One Embedding Space to Bind them All
  - [cvpr'23] - https://arxiv.org/abs/2305.05665

Masked Auto-Encoders as Scalable Vision Learners

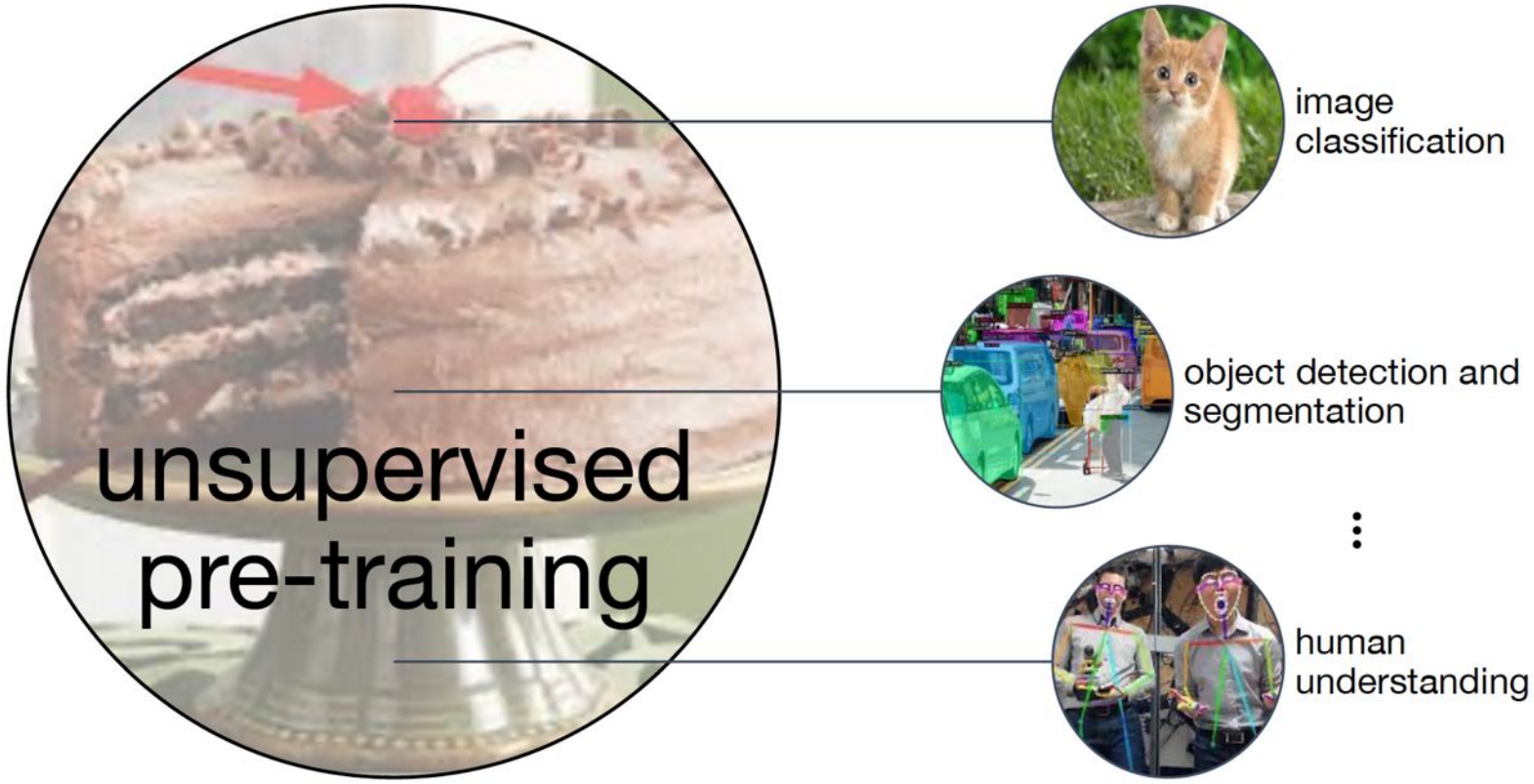
Xinlei Chen

ECCV 2022 tutorial on self-supervised representation learning in computer vision

facebook
Artificial Intelligence Research

Self-Supervised Learning

- Pre-train representations without labels for downstream tasks

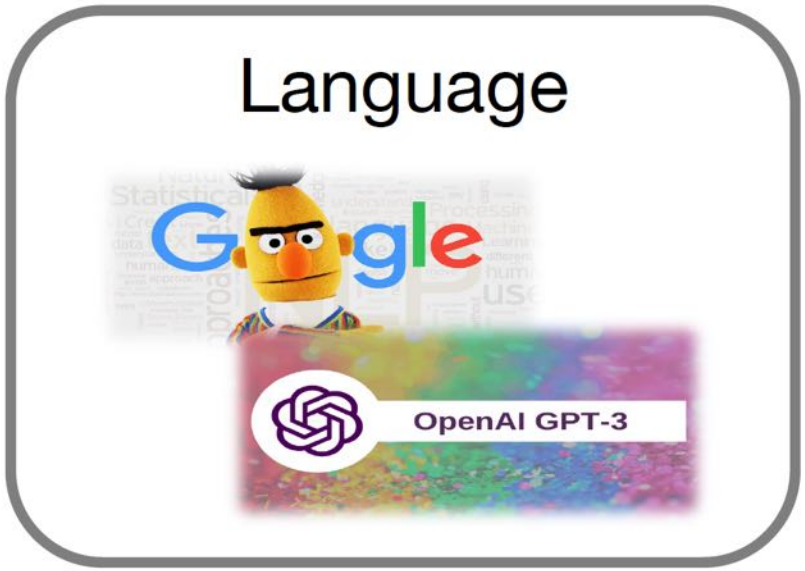


slide credit: Xinlei Chen

[Devlin et al, NAACL 2019] [Brown et al, NeurIPS 2020]

Self-Supervised Representation Learning

- Scalable: use unlimited data to train unlimited-sized models
- Tremendously successful in NLP

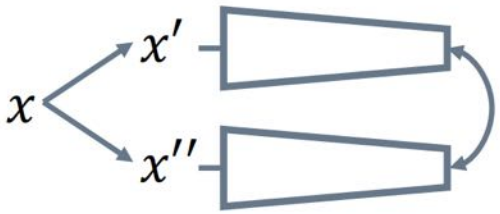


slide credit: Xinlei Chen

[Chen et al, ICML 2020] [He et al, CVPR 2022]

Self-Supervised Paradigms Covered

- Contrastive / Siamese



- Reconstructive / Auto-Encoding

Masked **A**uto-**E**ncoders Are Scalable Vision Learners:

Kaiming, Xinlei, Saining, Yanghao, Piotr, Ross

CVPR 2022

slide credit: Xinlei Chen

[Devlin et al, NAACL 2019] [He et al, CVPR 2022]

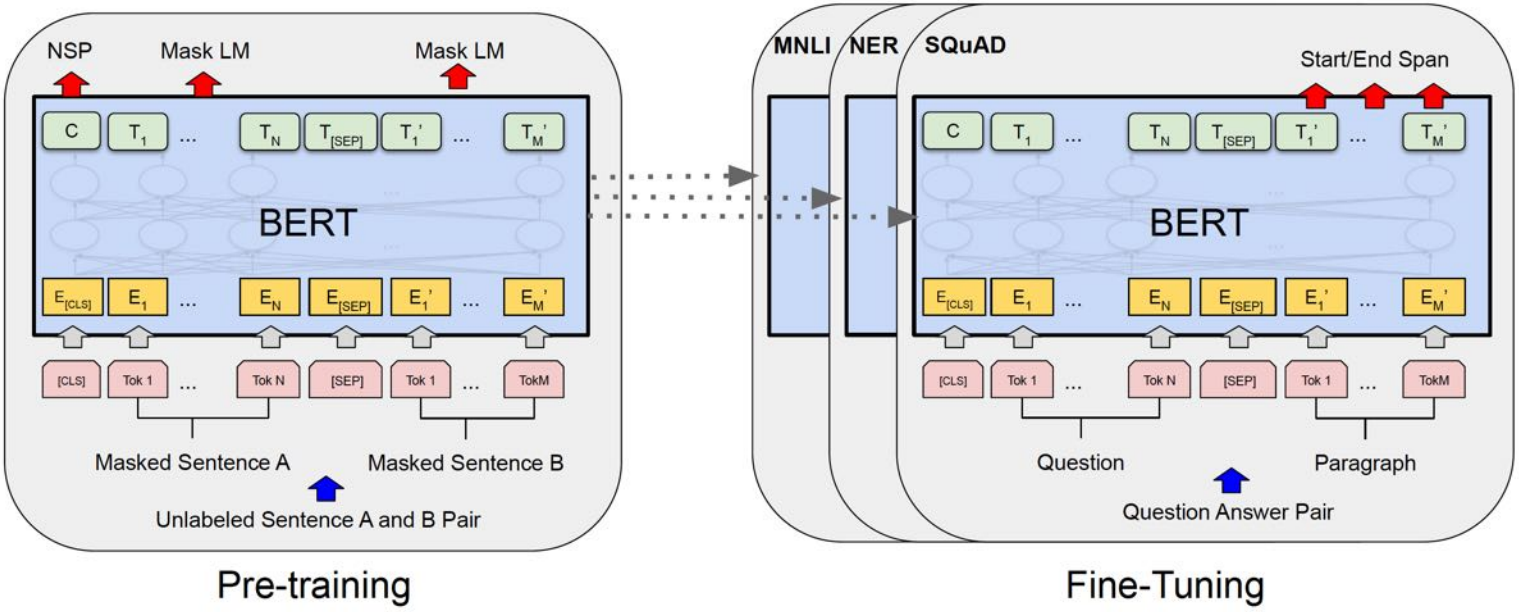
What is MAE?

- Very simple method, but highly effective
- BERT-like algorithm, but with crucial design changes for vision
- Intriguing properties – better scalability and more from analysis

slide credit: Xinlei Chen

BERT: Pretraining of Bidirectioanal Transformers for Language Understanding @NAACL 2019 — https://arxiv.org/abs/1810.04805

- Pretraining & Finetuning
- Pretraining Losses:
  - Mask LM: prediction of words masked out (without a direction !)
  - NSP: next sentence prediction (does sentence B follow sentence A)

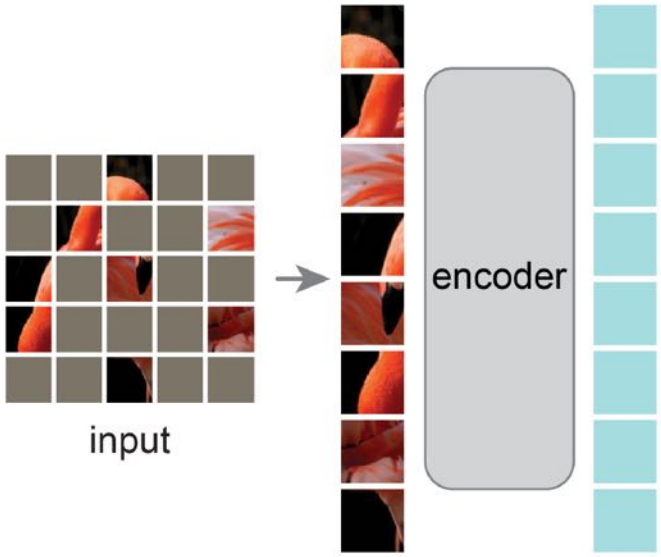


How MAE Works?

Random masking

slide credit: Xinlei Chen

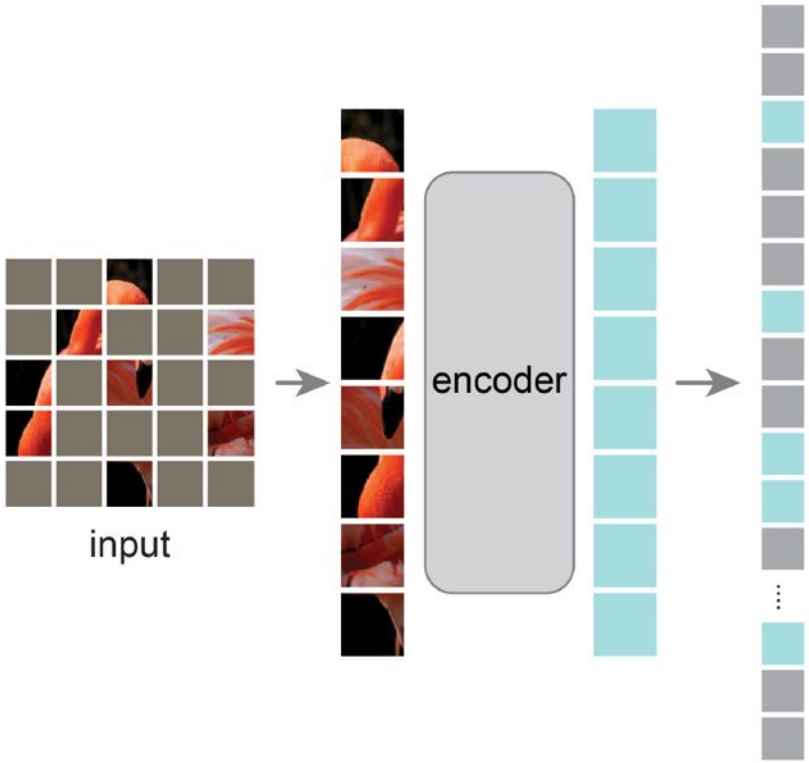
How MAE Works?



Encode visible patches

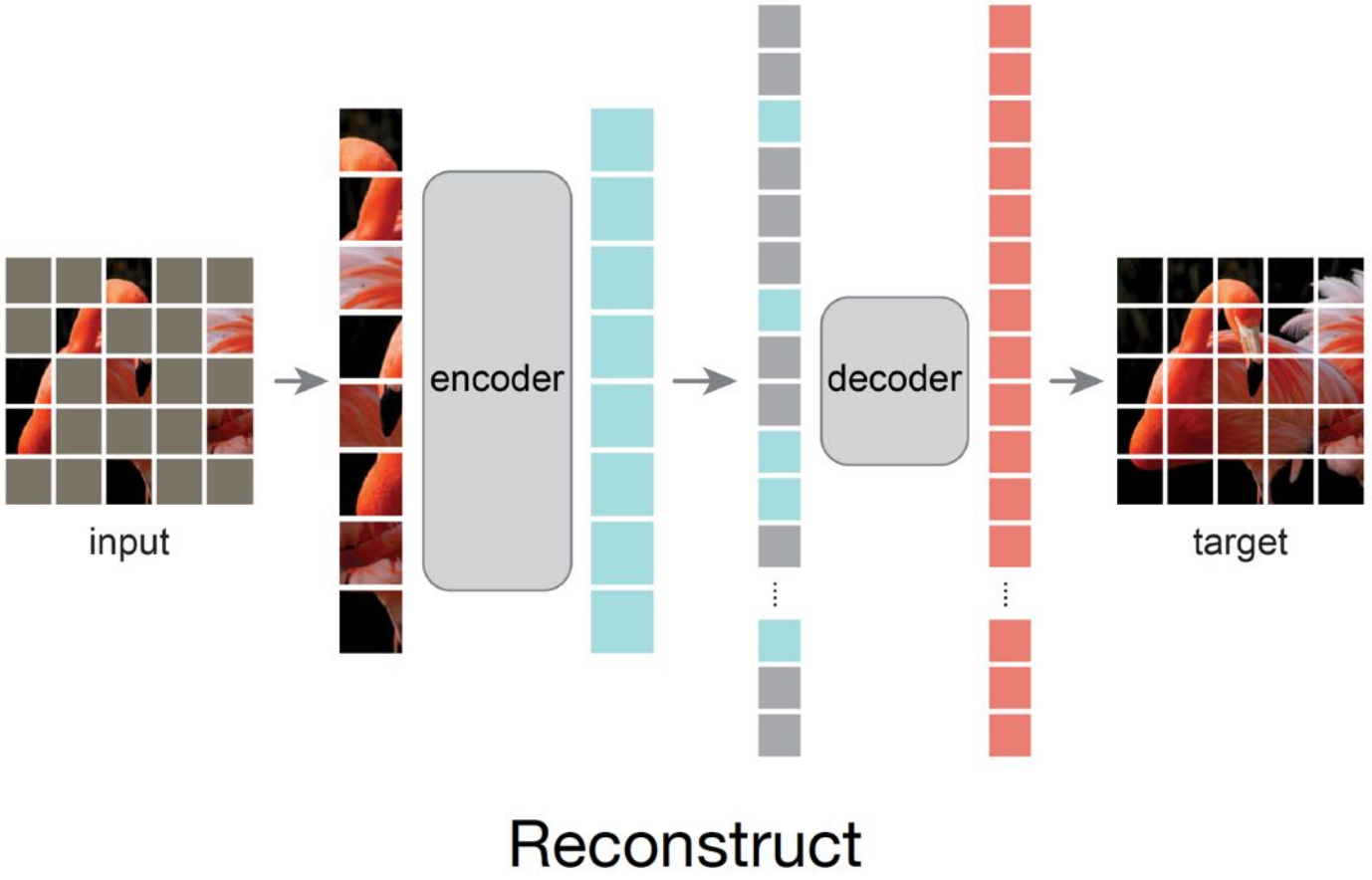
slide credit: Xinlei Chen

How MAE Works?



slide credit: Xinlei Chen

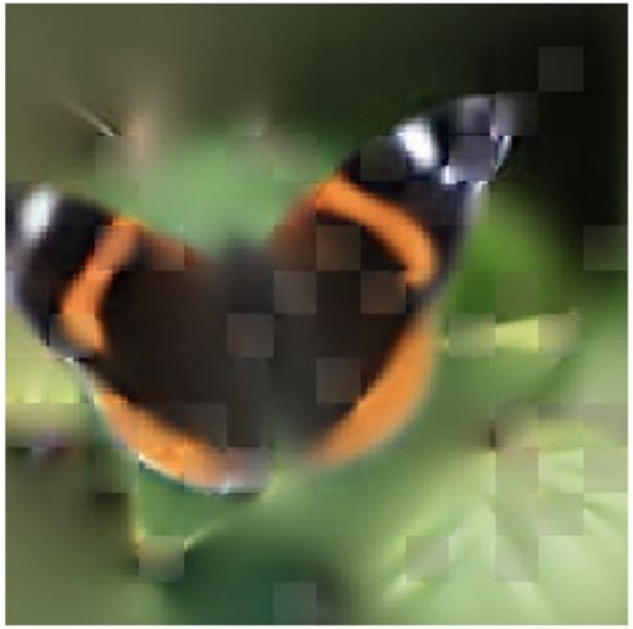
How MAE Works?



slide credit: Xinlei Chen

MAE Reconstruction Example

Masked input: 80%



MAE's guess

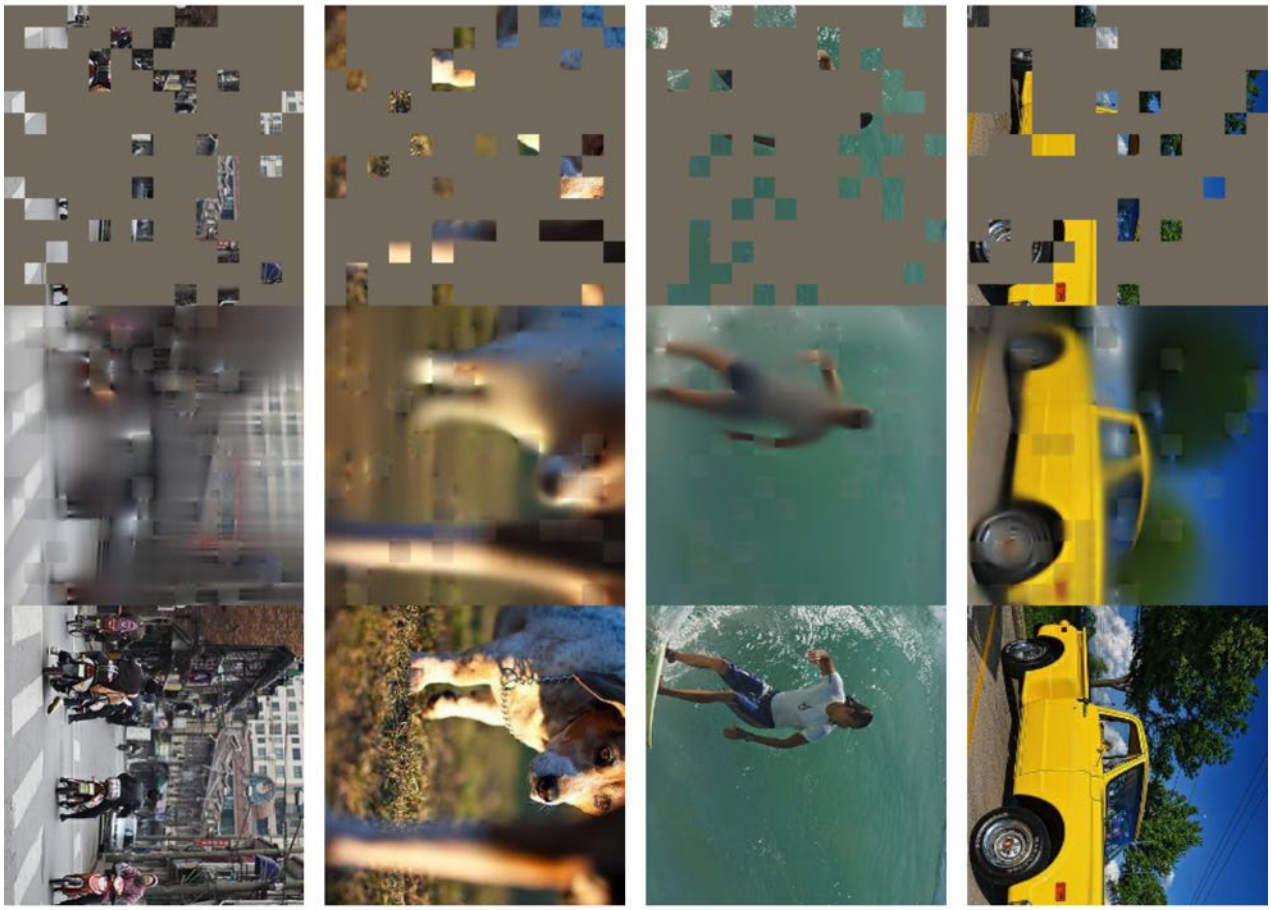
Ground truth

slide credit: Xinlei Chen

ImageNet val set (unseen)

slide credit: Xinlei Chen

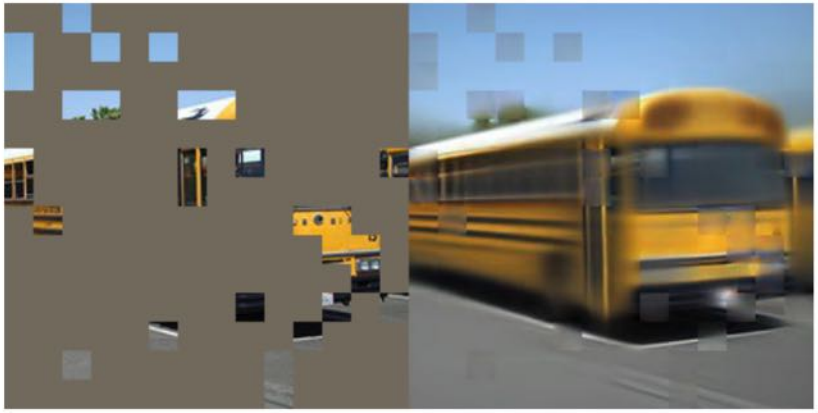
COCO val set (unseen)



slide credit: Xinlei Chen

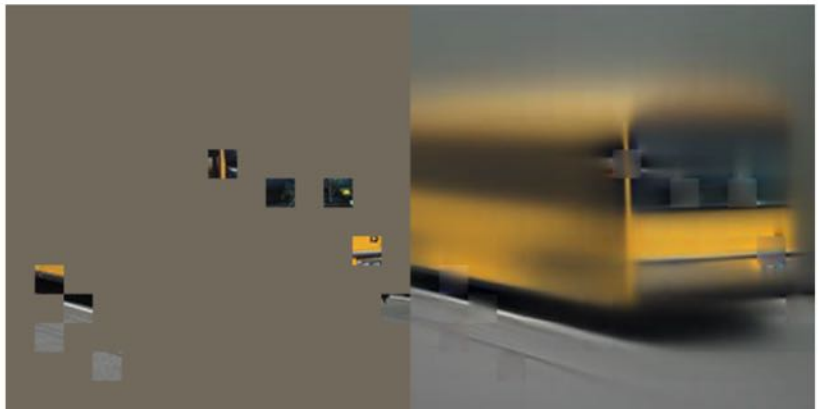
original

75% mask



85% mask

95% mask



MAE Can Generalize

slide credit: Xinlei Chen

original

75% mask

85% mask

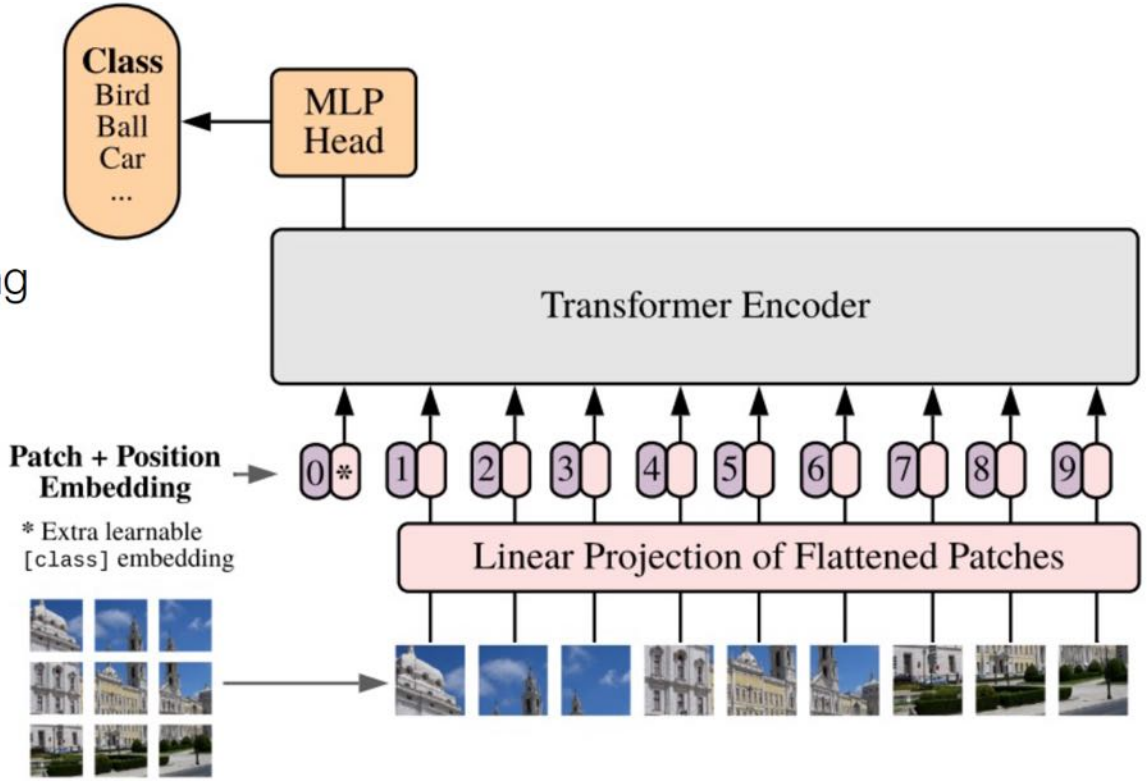
95% mask

MAE Can Generalize

[Dosovitskiy et al, ICLR 2021]

BERT-like: Transformers

- Vision Transformer (ViT)
  - Less inductive bias
  - Non-overlapping tokenization
    - Easier for masked auto-encoding
- *Scalable*
  - with larger models
  - on larger datasets

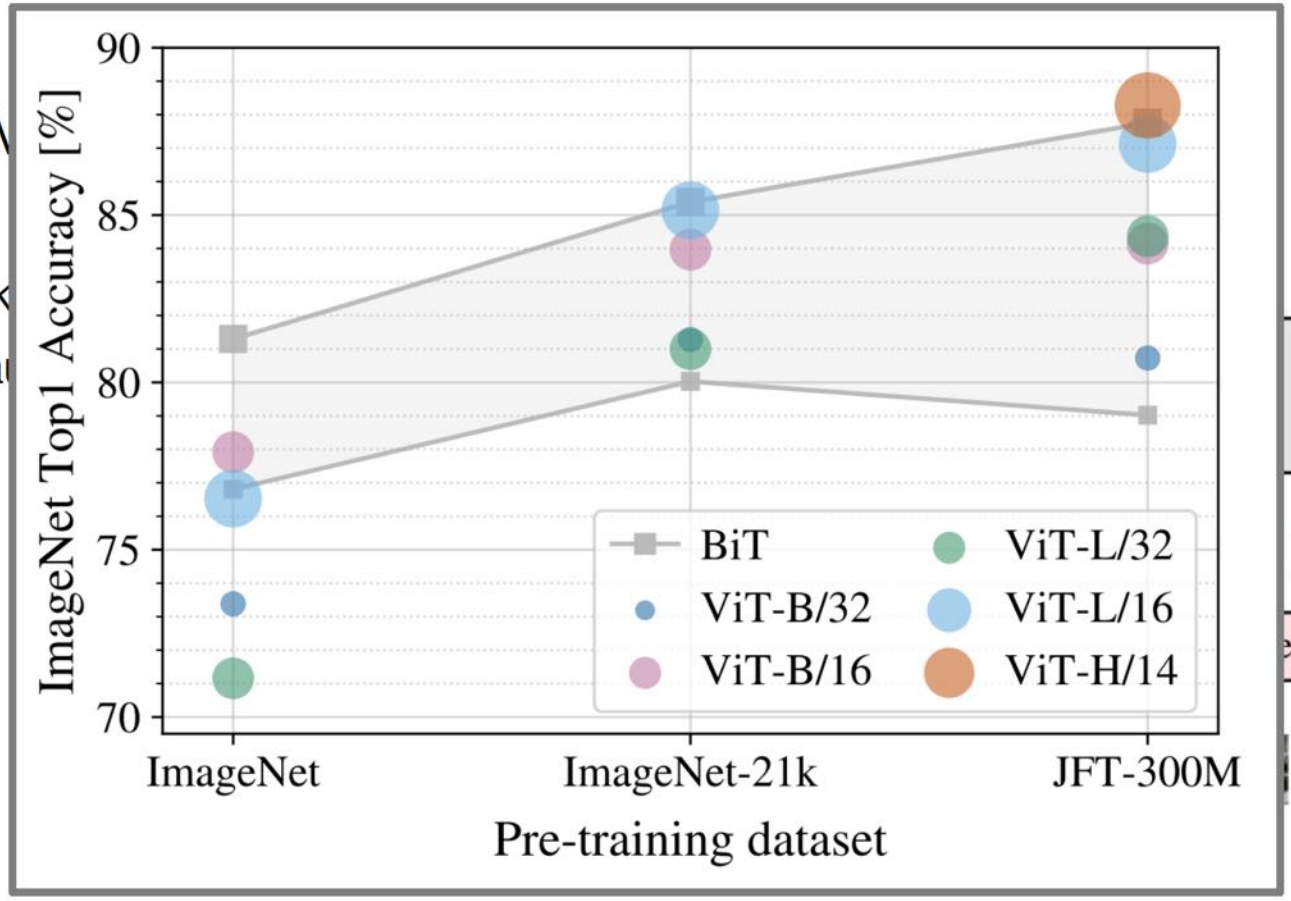


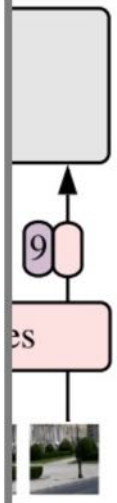
slide credit: Xinlei Chen

[Dosovitskiy et al, ICLR 2021]

BERT-like: Transformers

- Vision Transformer (V
  - Less inductive bias
  - Non-overlapping tok
    - Easier for masked a
- *Scalable*
  - with larger models
  - on larger datasets



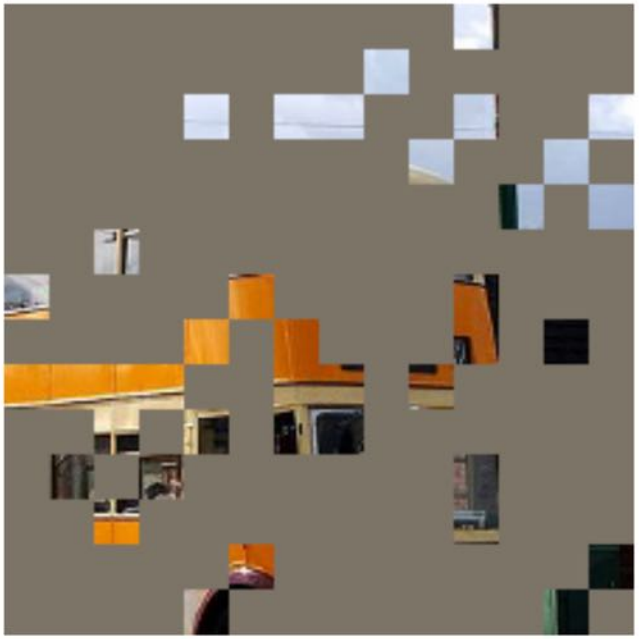


slide credit: Xinlei Chen

[Devlin et al, NAACL 2019]

BERT-*unlike*: Mask Ratio

- BERT: 15% is enough to create a challenging task
- MAE: a high ratio of 75% - 80% is about optimal

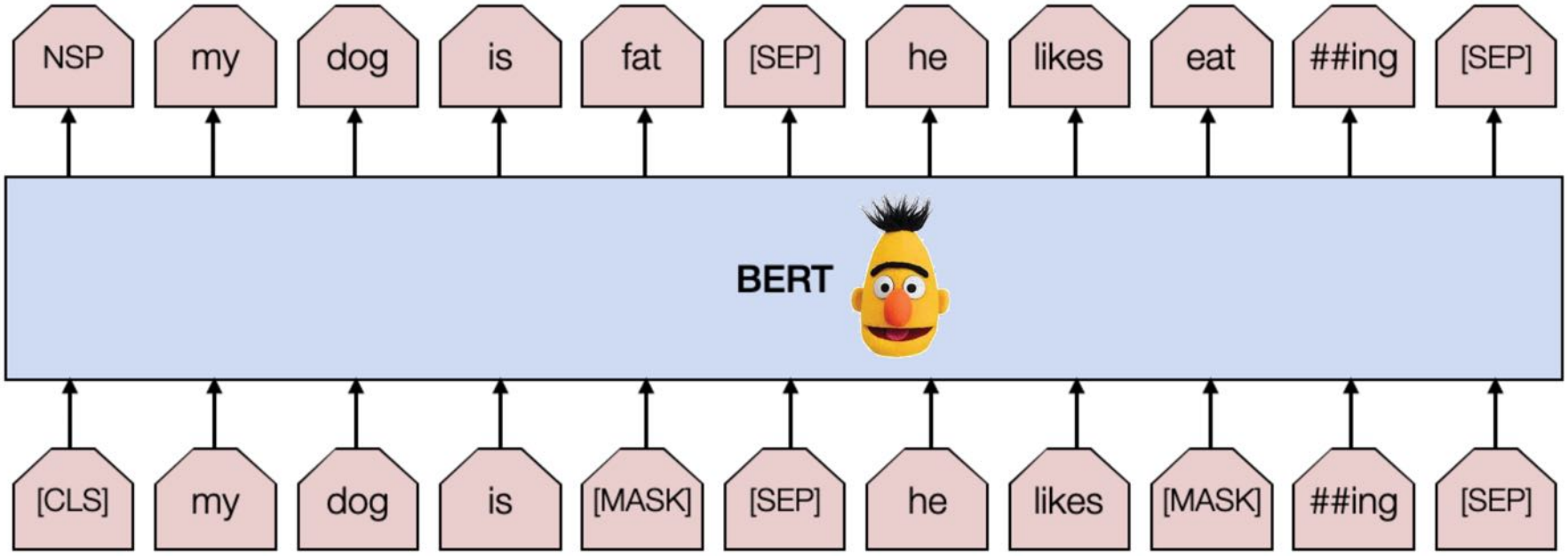


slide credit: Xinlei Chen

[Devlin et al, NAACL 2019]

BERT-*unlike*: Encoder-Decoder

- BERT: encoder-*only* pre-training



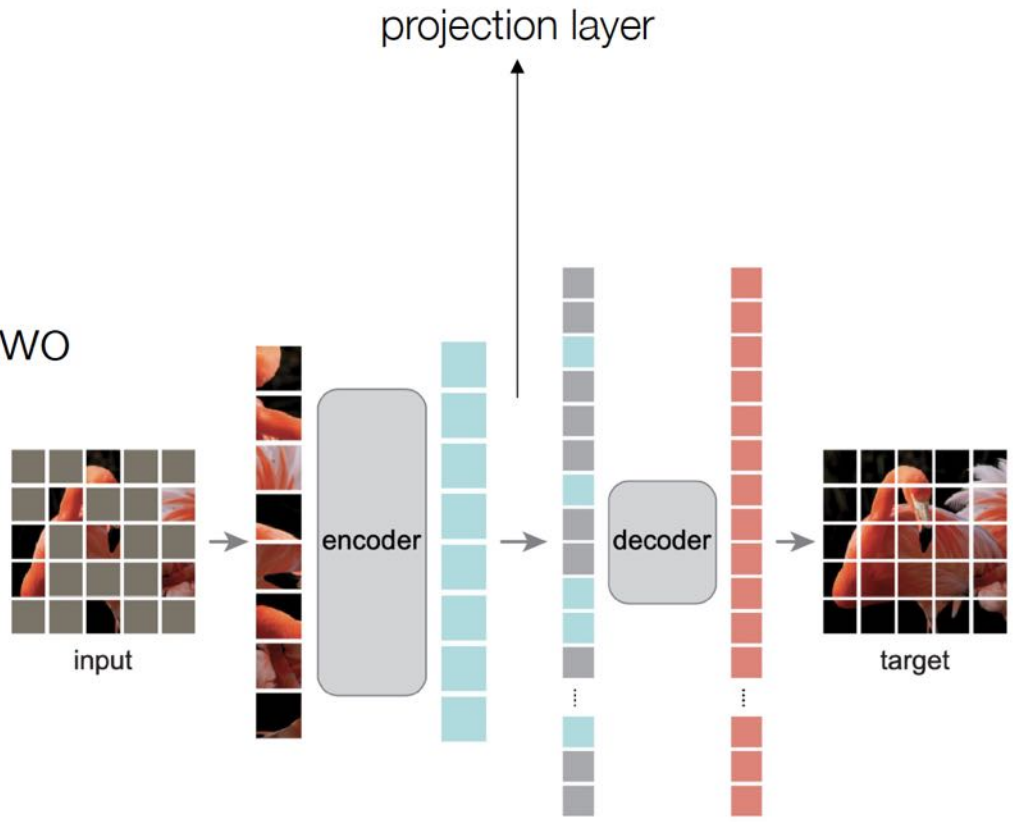
slide credit: Xinlei Chen

[Devlin et al, NAACL 2019]

BERT-*unlike*: Encoder-Decoder

- MAE:
  - *Large* encoder on *visible* tokens
  - Small decoder on *all* tokens
  - *Projection* layer to connect the two

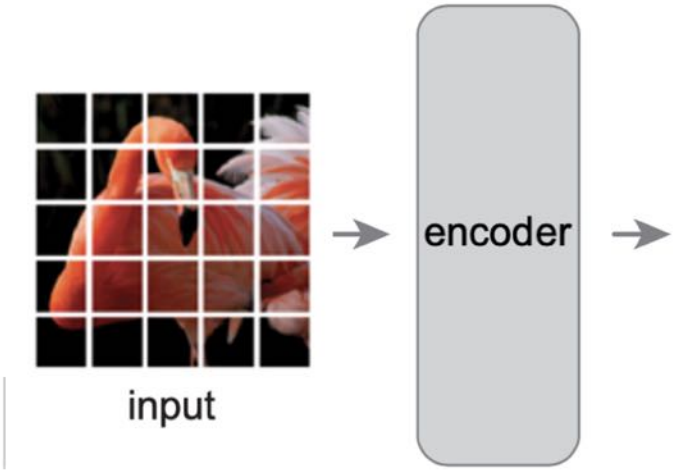
- Very efficient when coupled with high mask ratio (75%)



slide credit: Xinlei Chen

MAE for Downstream Tasks: *Encoder Only*

- After MAE pre-training, just *throw away* the decoder
- Encoder is used for representations with *full-sequence* input



slide credit: Xinlei Chen

Experimental Protocols

- Pre-training dataset: ImageNet-1K
- Architecture: ViT-*Large* encoder, 512-dim decoder
- Transfer task: ImageNet-1K classification
  - "*ft*": end-to-end tuning with MAE as an initialization
  - "*lin*": linear probing, a single classifier on top of frozen encoder features

slide credit: Xinlei Chen

Analysis: Decoder Size

- Encoder has 24-blocks, 1024-dimensional

blocks	ft	lin
1	84.8	65.5
2	**84.9**	70.0
4	**84.9**	71.9
8	**84.9**	**73.5**
12	84.4	73.3

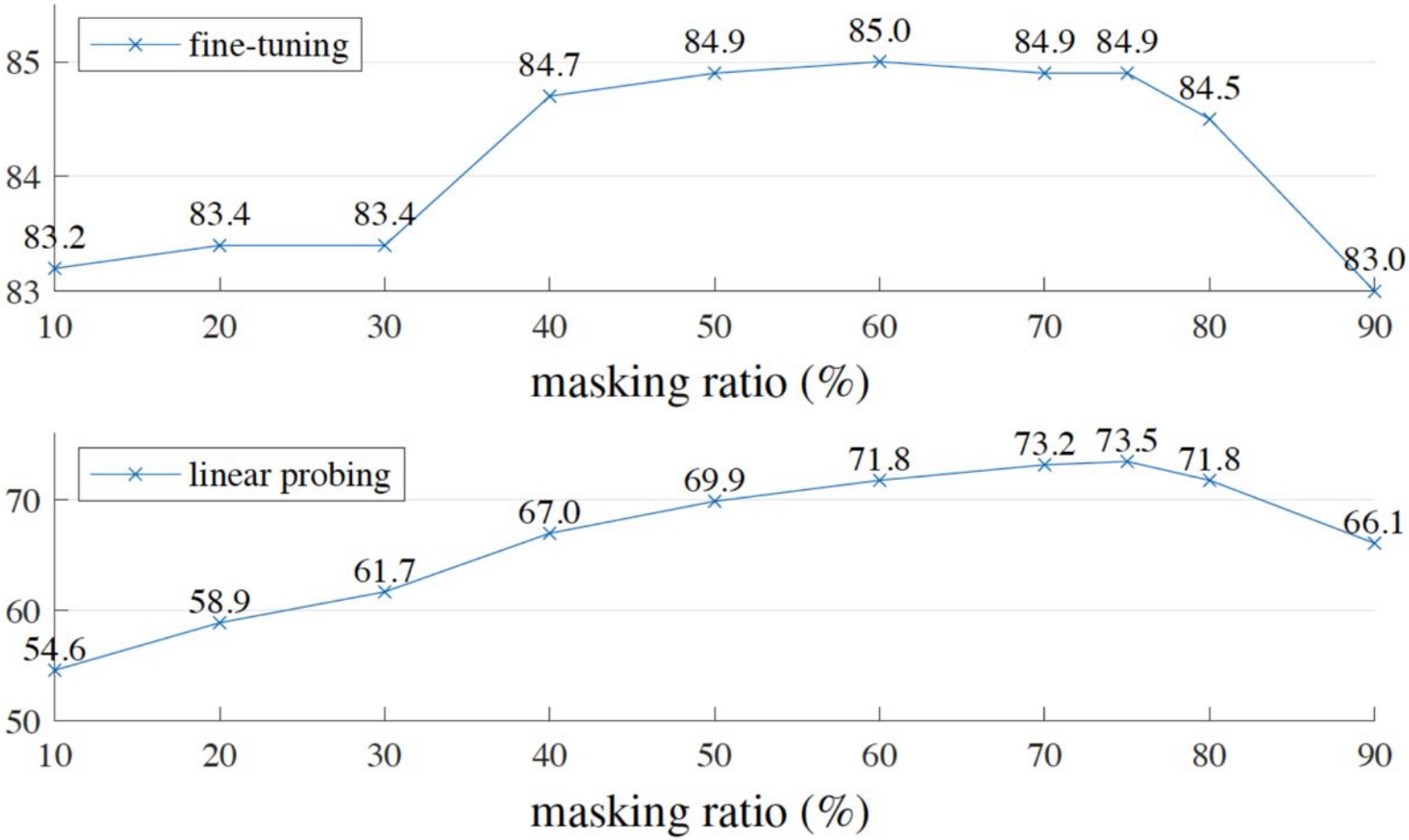
Decoder depth

dim	ft	lin
128	**84.9**	69.1
256	84.8	71.3
512	**84.9**	**73.5**
768	84.4	73.1
1024	84.3	73.1

Decoder width

slide credit: Xinlei Chen

Analysis: Mask Ratio



slide credit: Xinlei Chen

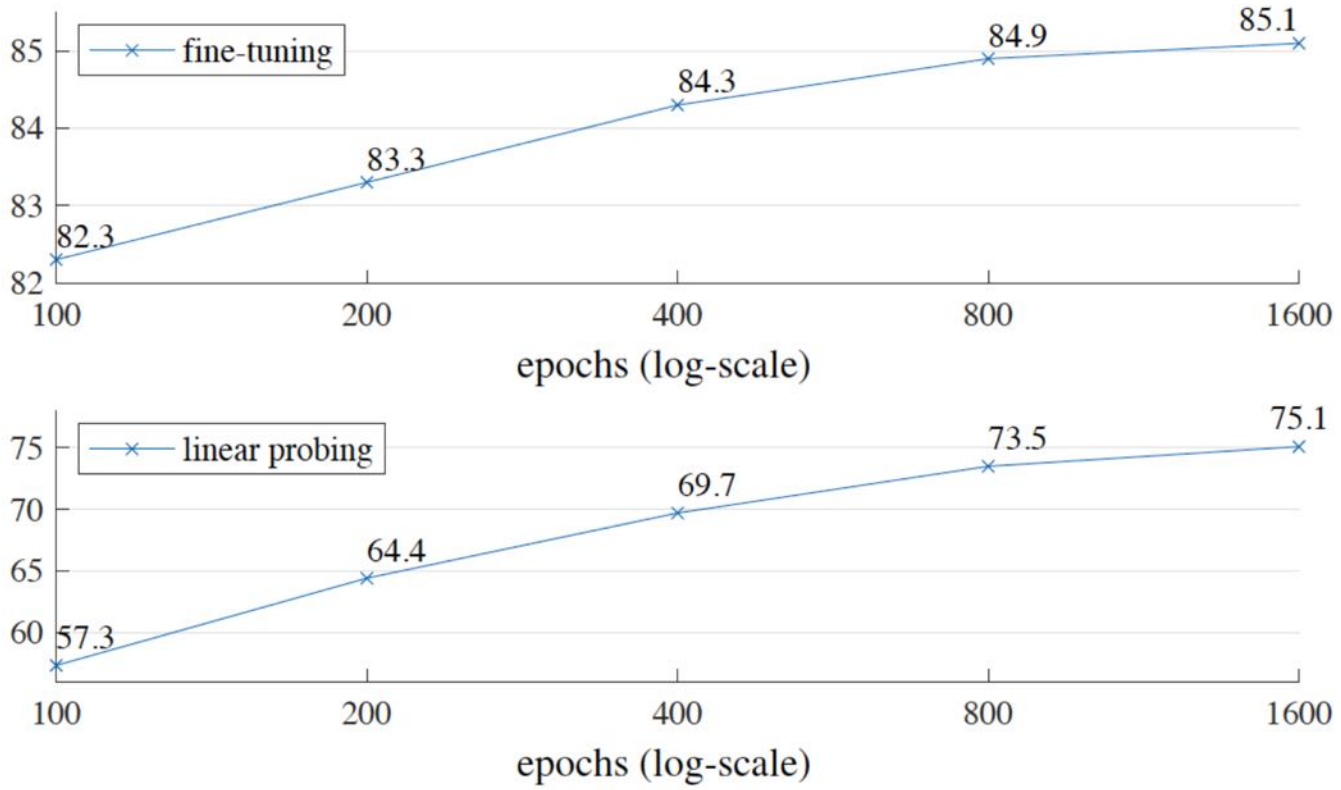
Analysis: Augmentations

case	ft	lin
none	84.0	65.7
crop, fixed size	84.7	73.1
crop, rand size	**84.9**	**73.5**
crop + color jit	84.3	71.9

- MAE can work with minimal data augmentation

slide credit: Xinlei Chen

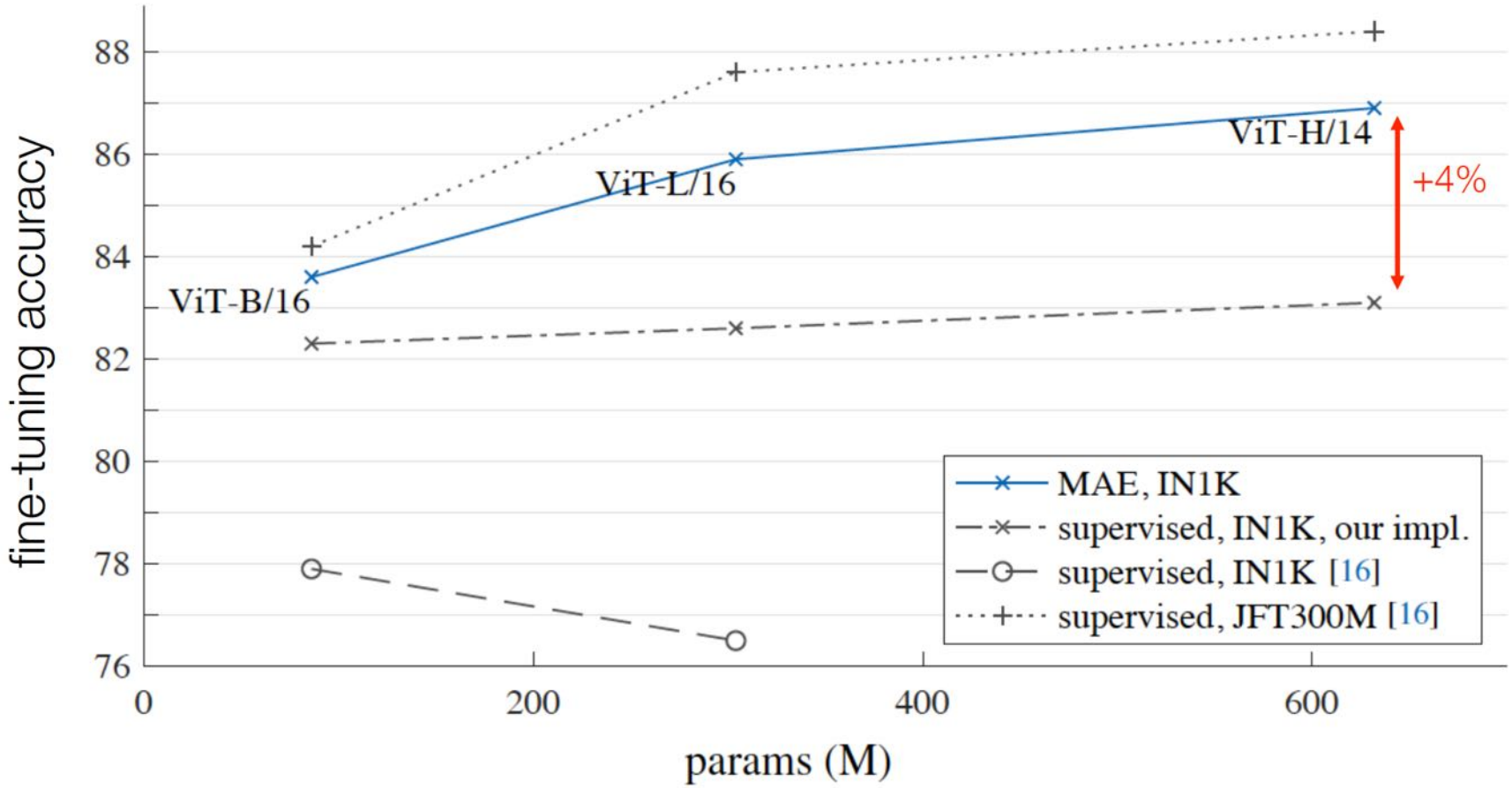
Scalability: Longer Training



Wall-clock speed still efficient thanks to MAE design

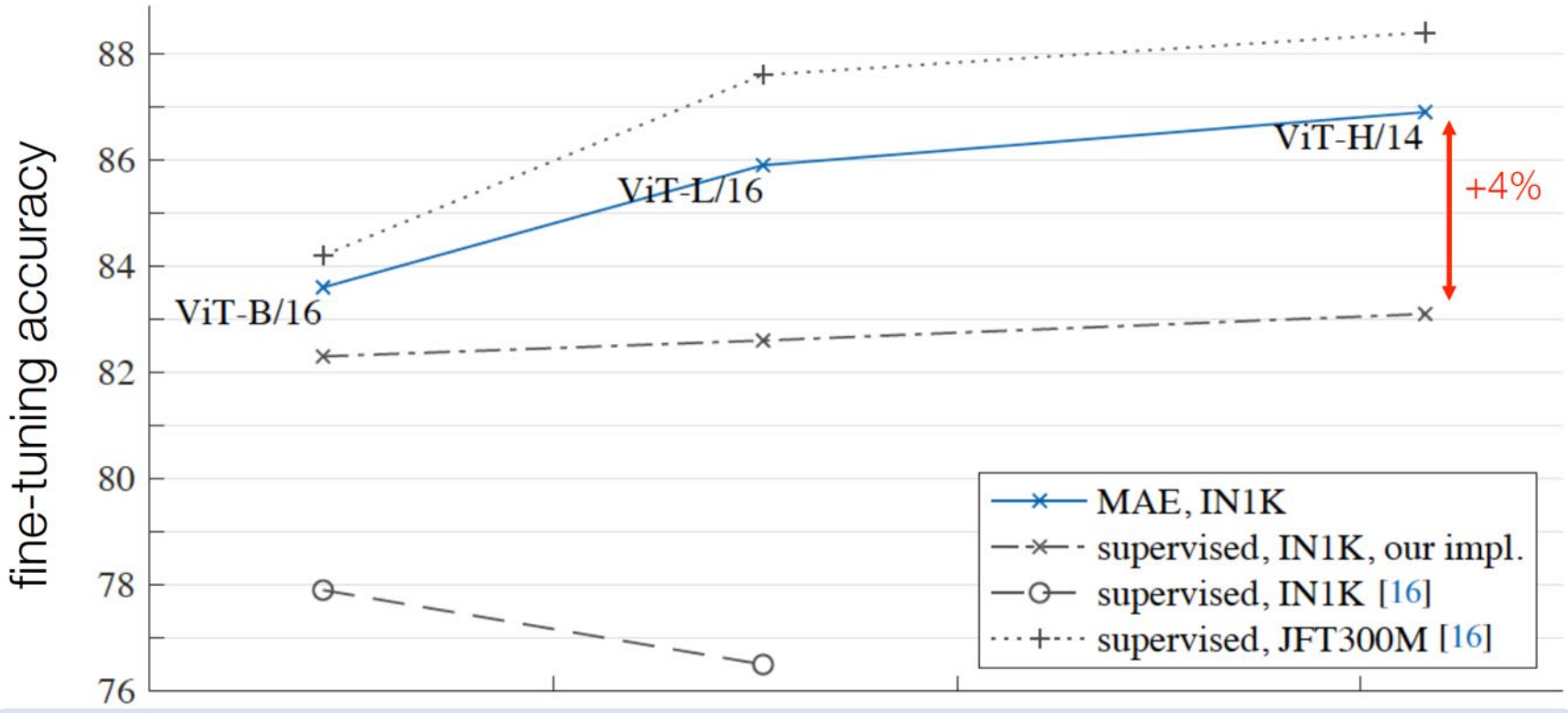
slide credit: Xinlei Chen

Scalability: Larger Models



slide credit: Xinlei Chen

Scalability: Larger Models



new SOTA on ImageNet-1K (no extra data): **87.8%**

slide credit: Xinlei Chen

slide credit: Xinlei Chen

Scalability: Larger Models

dataset	ViT-B	ViT-L	ViT-H	ViT-H$_{448}$	prev best
iNat 2017	70.5	75.7	79.3	**83.4**	75.4 [50]
iNat 2018	75.4	80.1	83.0	**86.8**	81.2 [49]
iNat 2019	80.5	83.4	85.7	**88.3**	84.1 [49]
Places205	63.9	65.8	65.9	**66.8**	66.0 [19]†
Places365	57.9	59.4	59.8	**60.3**	58.0 [36]‡

new SOTA on 5 large-scale classification datasets

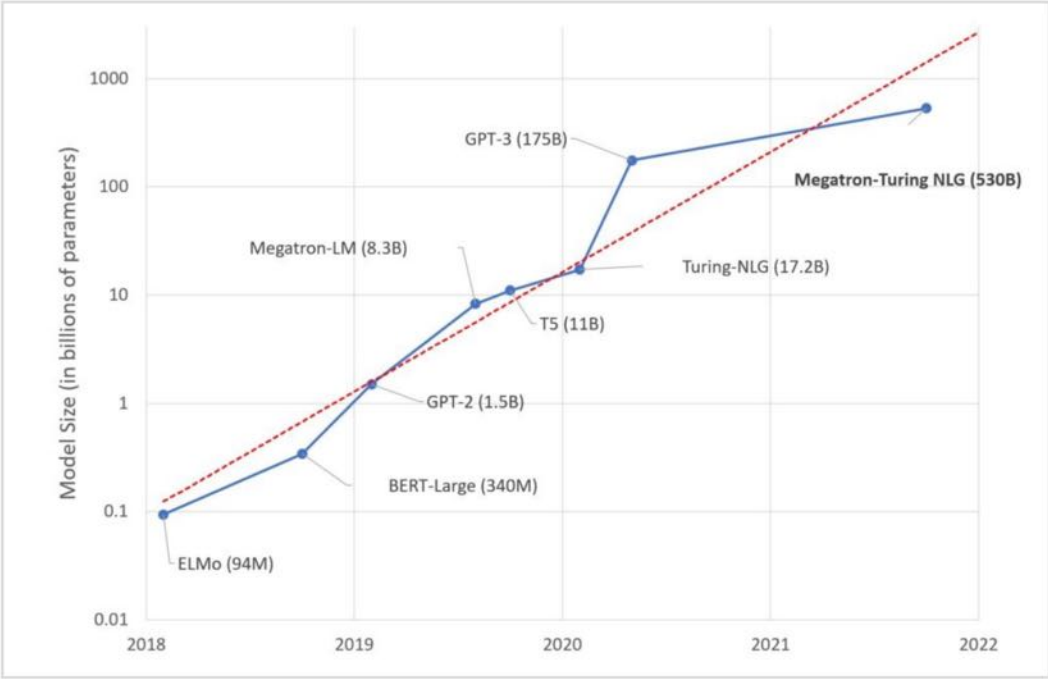
dataset	ViT-B	ViT-L	ViT-H	ViT-H$_{448}$	prev best
IN-Corruption ↓ [27]	51.7	41.8	**33.8**	36.8	42.5 [32]
IN-Adversarial [28]	35.9	57.1	68.2	**76.7**	35.8 [41]
IN-Rendition [26]	48.3	59.9	64.4	**66.5**	48.7 [41]
IN-Sketch [60]	34.5	45.3	49.6	**50.9**	36.0 [41]

new SOTA on 4 ImageNet robust evaluations

Is the Journey 99% Done?

- NLP has witnessed amazing progress in scaling since BERT

- It's just *starting* in vision:
  - Temporal data
  - Architectures – ConvNets?
  - Other modalities? 3D?
  - Other downstream tasks?
  - Other axes to scale?
  - [`Your exploration`] here!



slide credit: Xinlei Chen

code (GPU): https://github.com/facebookresearch/mae
code (TPU): https://github.com/facebookresearch/long_seq_mae

Take-aways

- Self-supervised learning aims at *scalable* representation learning

- Masked auto-encoders can serve as scalable vision learners

- Exciting years ahead in this direction!

slide credit: Xinlei Chen

Overview of Today's Lecture

- Segment Anything Model (SAM)
  - [arxiv'23] - https://arxiv.org/abs/2304.02643
  - https://segment-anything.com/

- Masked Autoencoders (MAEs) are Scalable Vision Learners
  - [cvpr'22] - https://arxiv.org/abs/2111.06377

- ImageBind: One Embedding Space to Bind them All
  - [cvpr'23] - https://arxiv.org/abs/2305.05665

Transformers are Modality Guzzlers

Can guzzle all modalities

- Structured input - text, images, speech, audio
- Unordered input - points, graphs

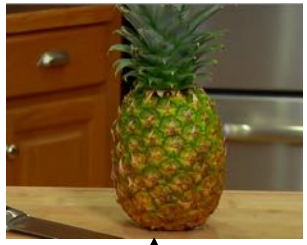
Seem to scale well with data/model size

Great candidate for **multi-modal** learning!

slide credit: Ishan Misra

Recipe for multimodal learning

- Get billions of (image, text) pairs
- Learn representations that “align” images with text



A pineapple sitting on the counter

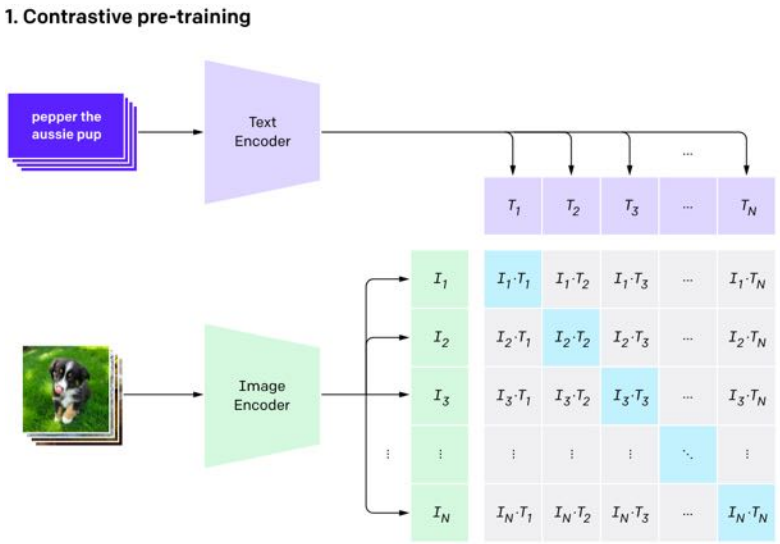


Image source: CLIP - Radford et al., 2021

slide credit: Ishan Misra

Aligned image-text features

- Aligned representations are *really* useful

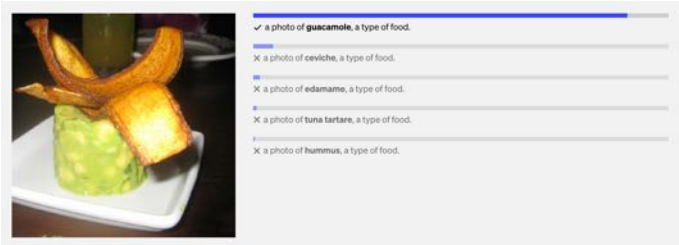
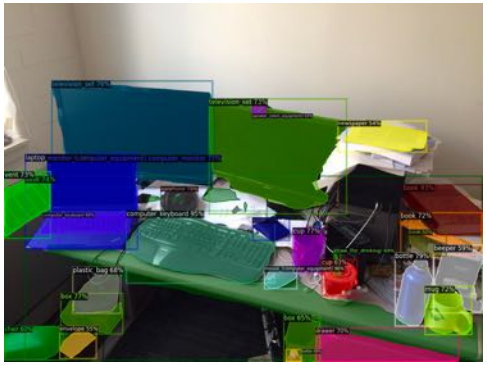


Image-text retrieval
Open-vocabulary classification[1]



Open-vocabulary detection and segmentation[2]

Text to image generation[3]

[1] CLIP - Radford et al., 2021
[2] Detic - Zhou et al., 2022
[3] GLIDE - Nichol et al., 2022, LAFITE - Zhou et al., 2022

slide credit: Ishan Misra

Aligned image-text features

- Aligned representations are *really* useful

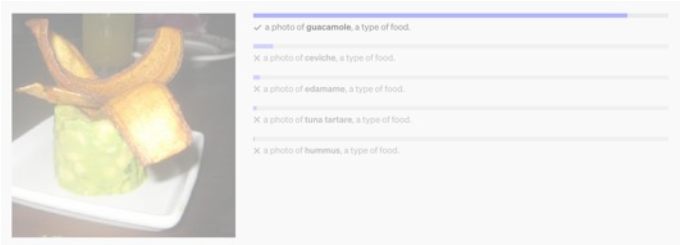
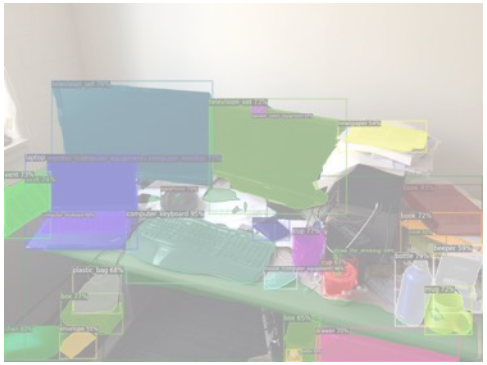


Image-text retrieval
Open-vocabulary classification[1]



Open-vocabulary detection and segmentation[2]

Text to image generation[3]

So have we "solved" multi-modal learning?

slide credit: Ishan Misra

[1] CLIP - Radford et al., 2021
[2] Detic - Zhou et al., 2022
[2] GLIDE - Nichol et al., 2022, LAFITE - Zhou et al., 2022

Problem 1: **Multi**-modal != **Bi**-modal

There are other modalities …

Image source: Rawpixel, The Rijksmuseum

slide credit: Ishan Misra

Problem 2: **Aligned** data is hard to get

Depth

Thermal

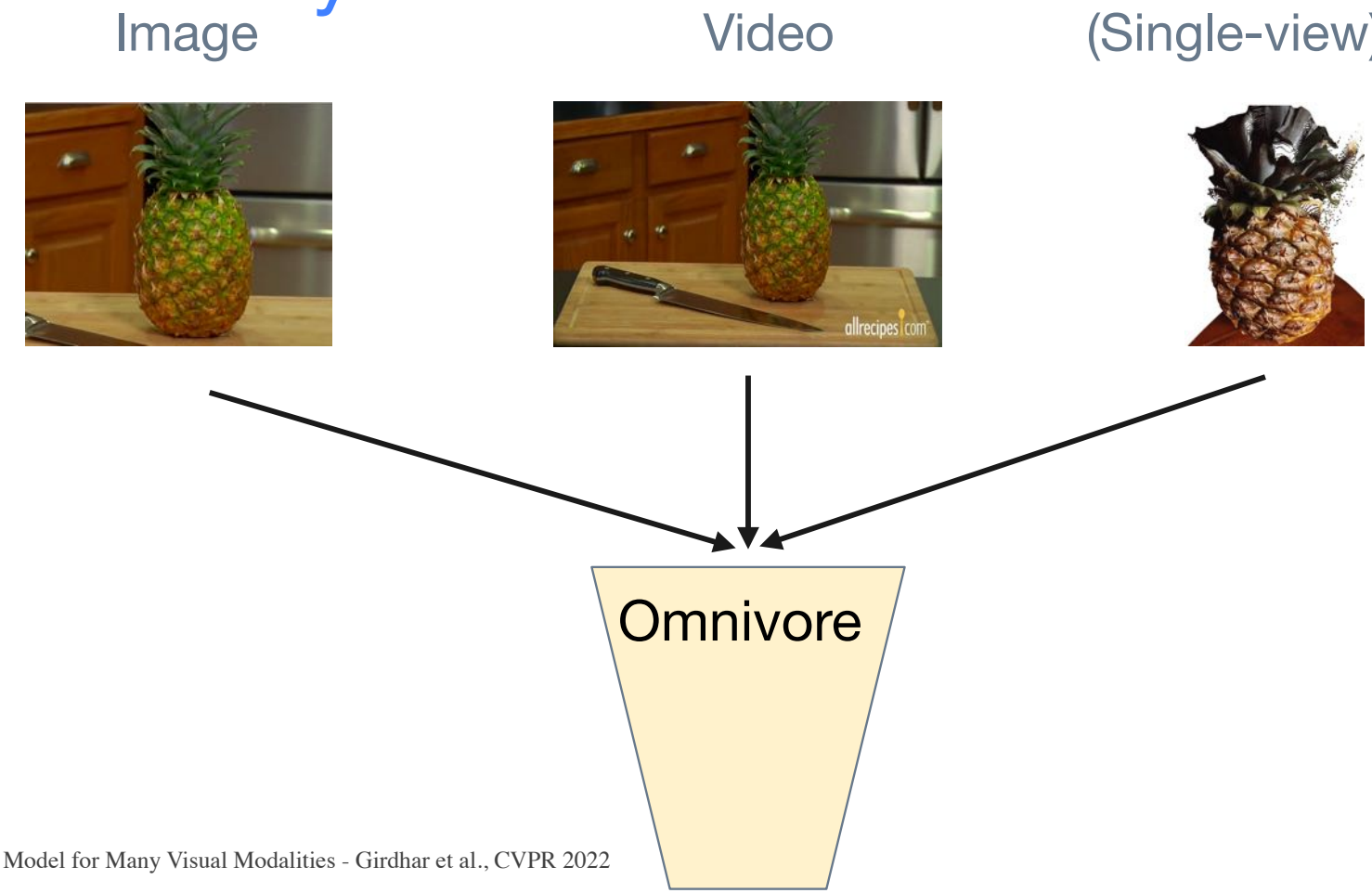
Motion (IMU)

Audio

Image source: Rawpixel, The Rijksmuseum

slide credit: Ishan Misra

Solution 1: Single model: Omnivore: A Single Model for Many Visual Modalities



Omnivore: A Single Model for Many Visual Modalities - Girdhar et al., CVPR 2022

slide credit: Ishan Misra

Emergent Property 1

Train on three modalities with completely separate/unaligned datasets

Cross-modal alignment emerges because of

- Shared parameters
- Shared information in modalities (visual structure)

slide credit: Ishan Misra

Omnivore: Cross-modal alignment emerges!

slide credit: Ishan Misra

How do we push this further?

More visual modalities

- Thermal

Non-visual modalities

- Text, Motion signals (IMU)

slide credit: Ishan Misra

Images are a universal language

Depth

Thermal

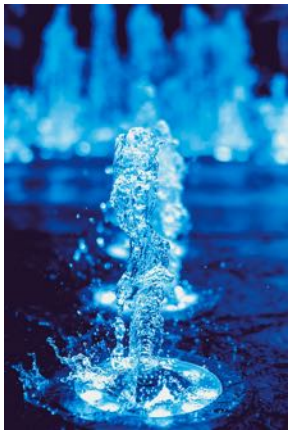
Motion (IMU)

Audio

RGB

RGB

RGB



RGB

slide credit: Ishan Misra

Image source (L to R): SUN RGB-D, LLVIP, Isaque Pereira, Ego4D, Wikimedia,

Images are a universal language

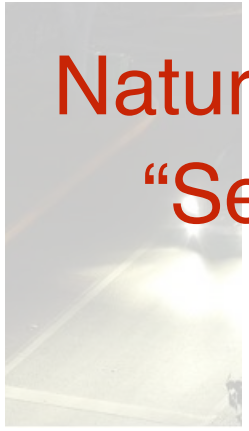
Depth

Thermal

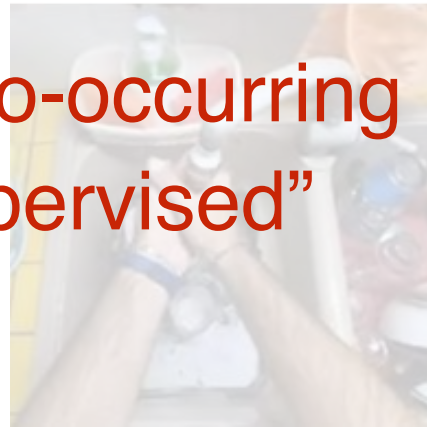
Motion (IMU)

Audio

RGB



RGB



RGB

RGB

Naturally co-occurring
"Self-supervised"

slide credit: Ishan Misra

Image source (L to R): SUN RGB-D, LLVIP, Isaque Pereira, Ego4D, Wikimedia,

ImageBind: One Embedding to Rule them All

Rohit Girdhar*, Alaaeldin El-Nouby*, Zhuang Liu, Mannat Singh,
Kalyan Vasudev Alwala, Armand Joulin, Ishan Misra*

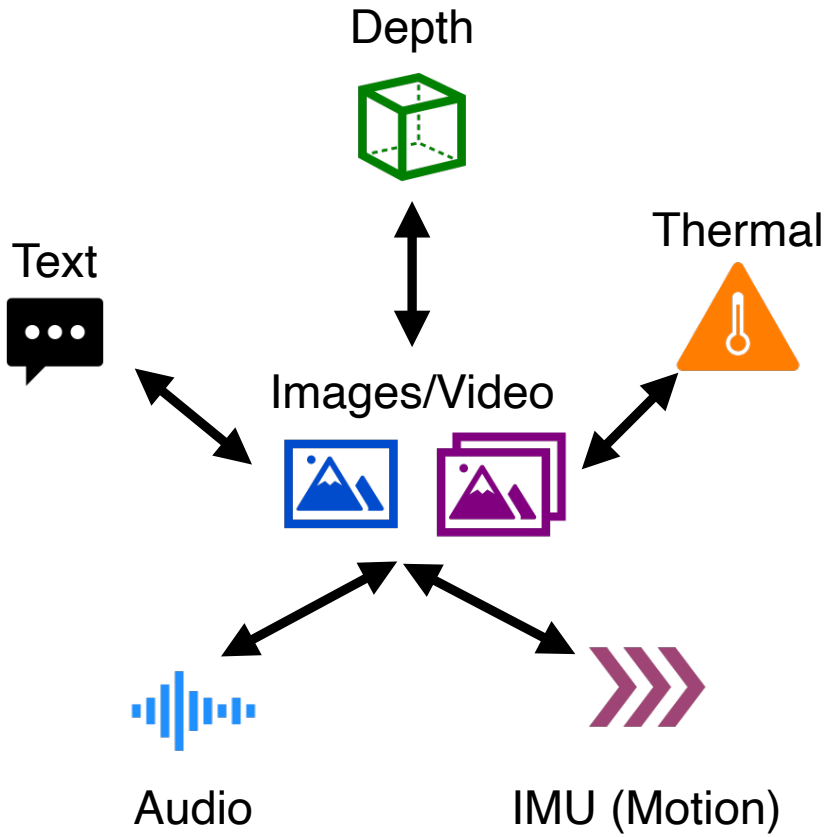
https://github.com/facebookresearch/ImageBind

CVPR 2023

slide credit: Ishan Misra

Key Idea

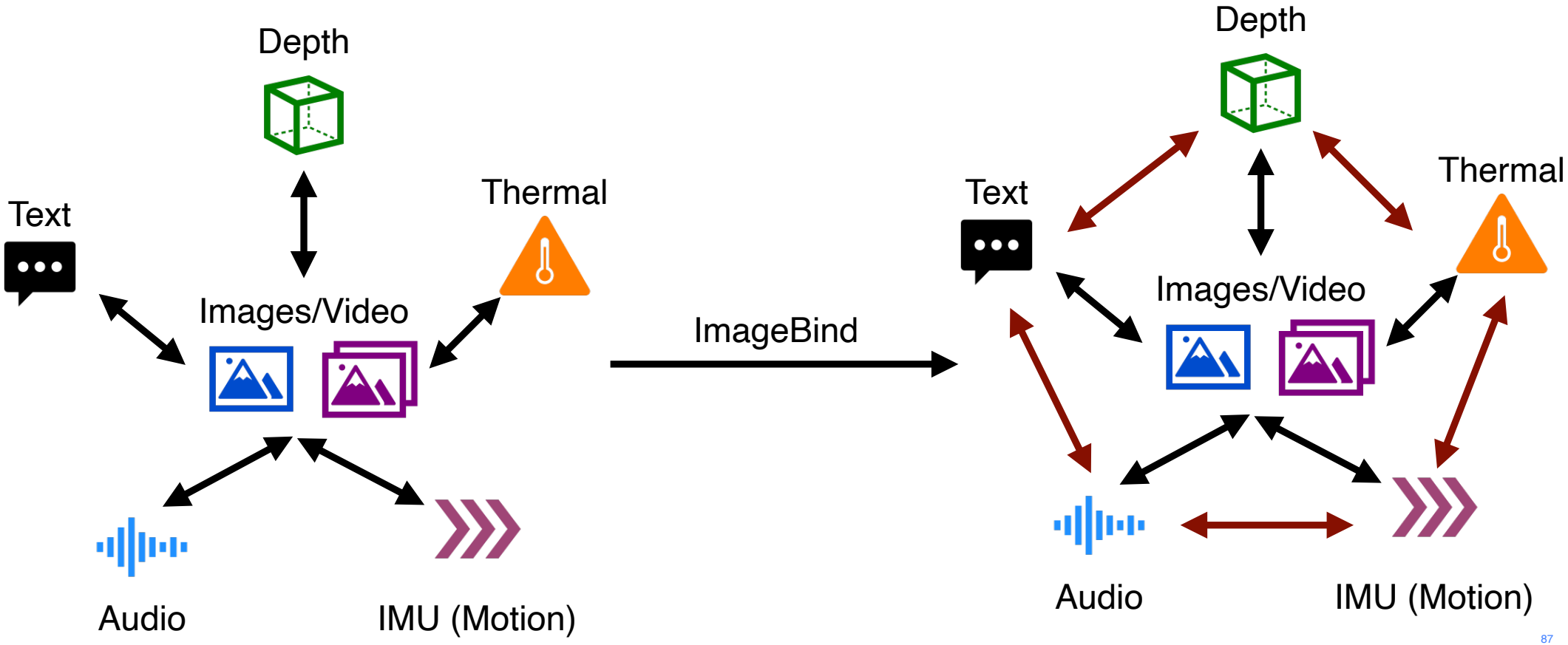
- Images naturally co-occur with different modalities
- Align every modality's representation with images
- Heavily leverage self-supervised learning



slide credit: Ishan Misra

Emergent behavior (Transitive alignment!)

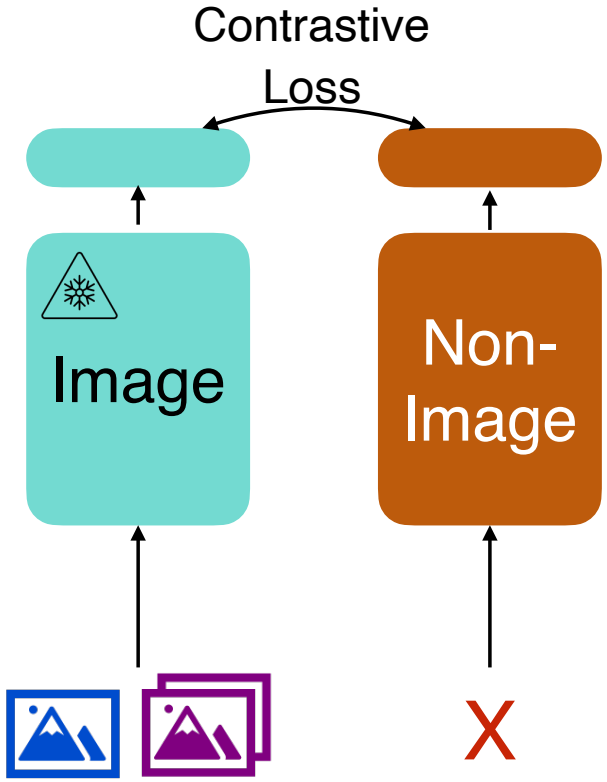
- After training **all** modalities are aligned



slide credit: Ishan Misra

Training setup

- 6 modalities — Image/Video, Text, Audio, Depth, IMU, Thermal
- Train only with image-paired data
- Separate encoder per modality
- Initialize image & text encoder from CLIP/OpenCLIP and keep frozen



slide credit: Ishan Misra

Measuring emergent alignment to text

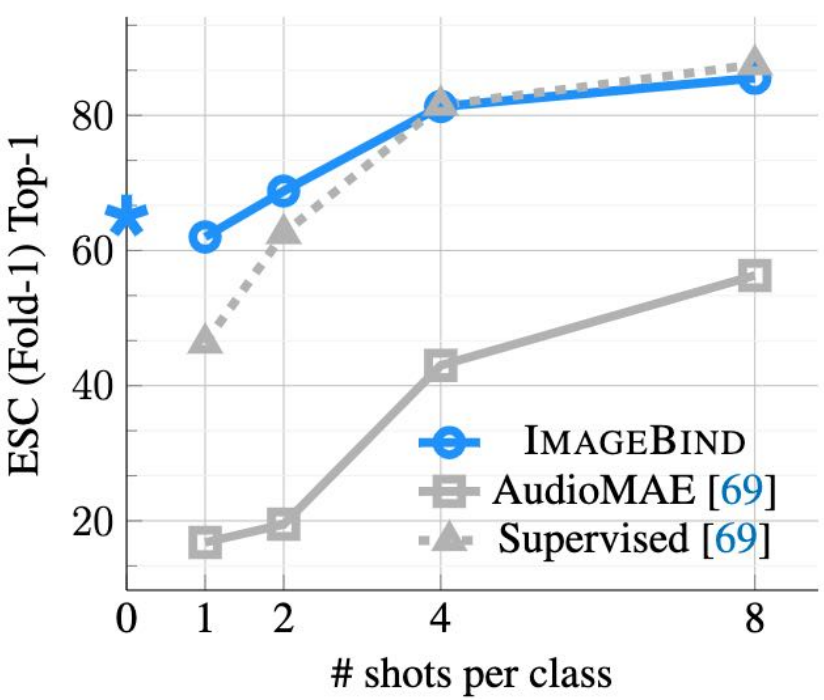
- Train on (Image, X) (Image, Text)
- Test on (X, Text) —> **“Emergent”** zero-shot classification

	Image		Video		Depth		Audio			Thermal	IMU
	IN1k	P365	K400	MSVTT	NYU	SUN	AudioSet	VGGS	ESC	LLVIP	Ego4D
Random	0.1	0.27	0.25	0.1	10.0	5.26	0.62	0.32	2.75	50.0	0.9
ImageBind	77.7	45.4	50.0	36.1	**54.0**	**35.1**	17.6	**27.8**	66.9	**63.4**	**25.0**
Text paired	-	-	-	-	41.9	25.4	**28.4**	-	**68.6**	-	-
Absolute SOTA	91.0	60.7	89.9	57.7	76.7	64.9	49.6	52.5	97.0	-	-

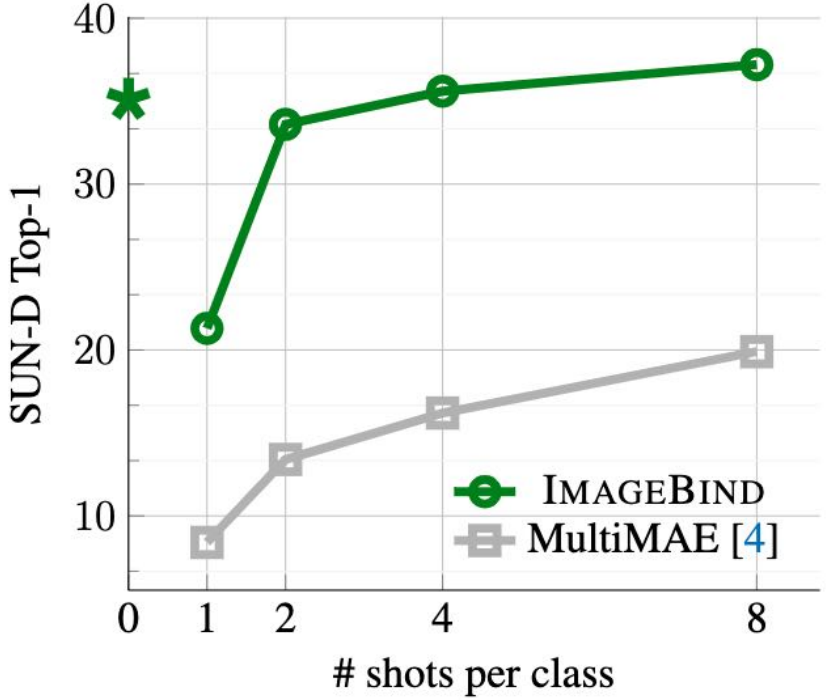
slide credit: Ishan Misra

Measuring performance on few-shot classification

Few-shot audio

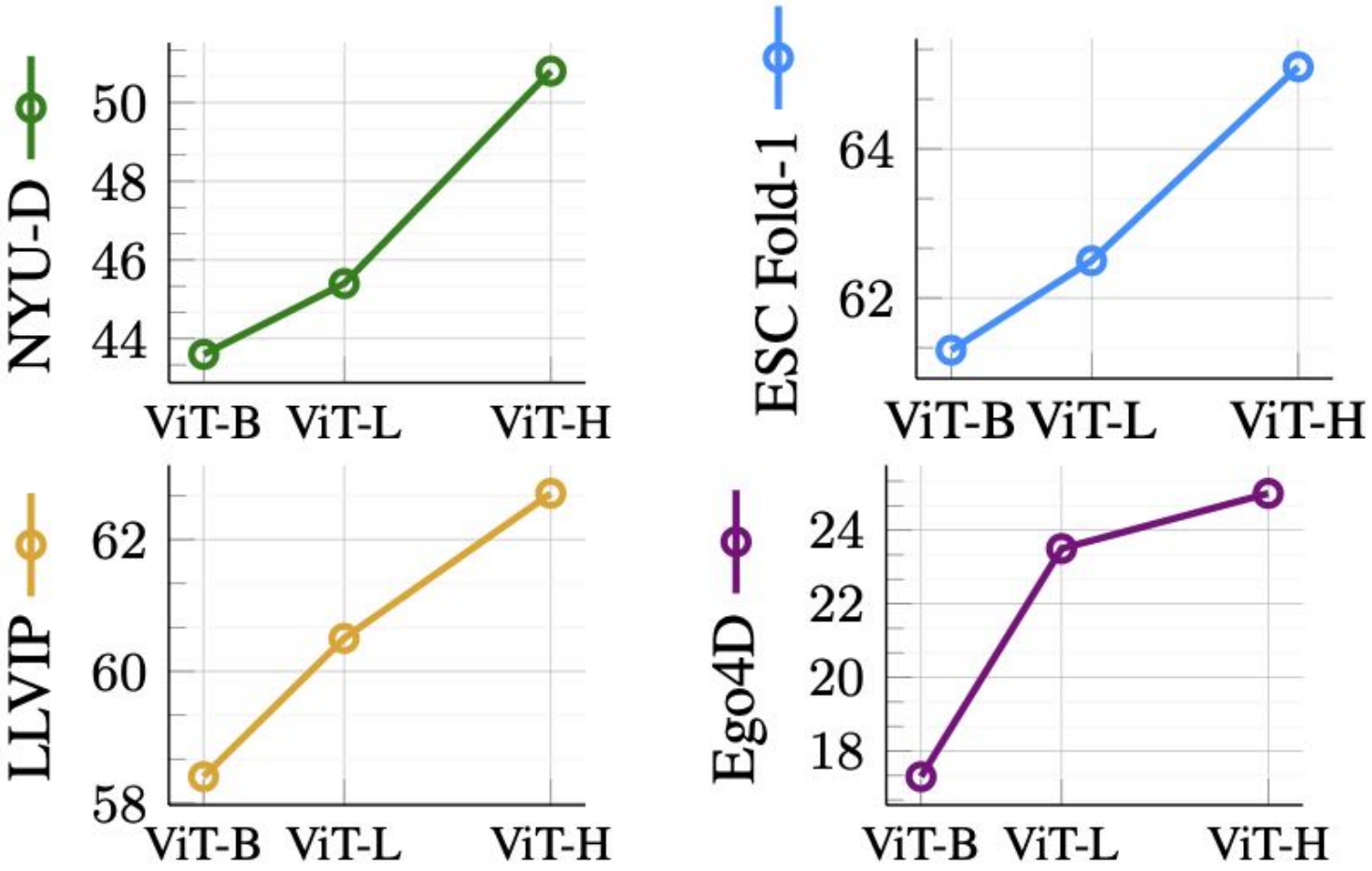


Few-shot depth



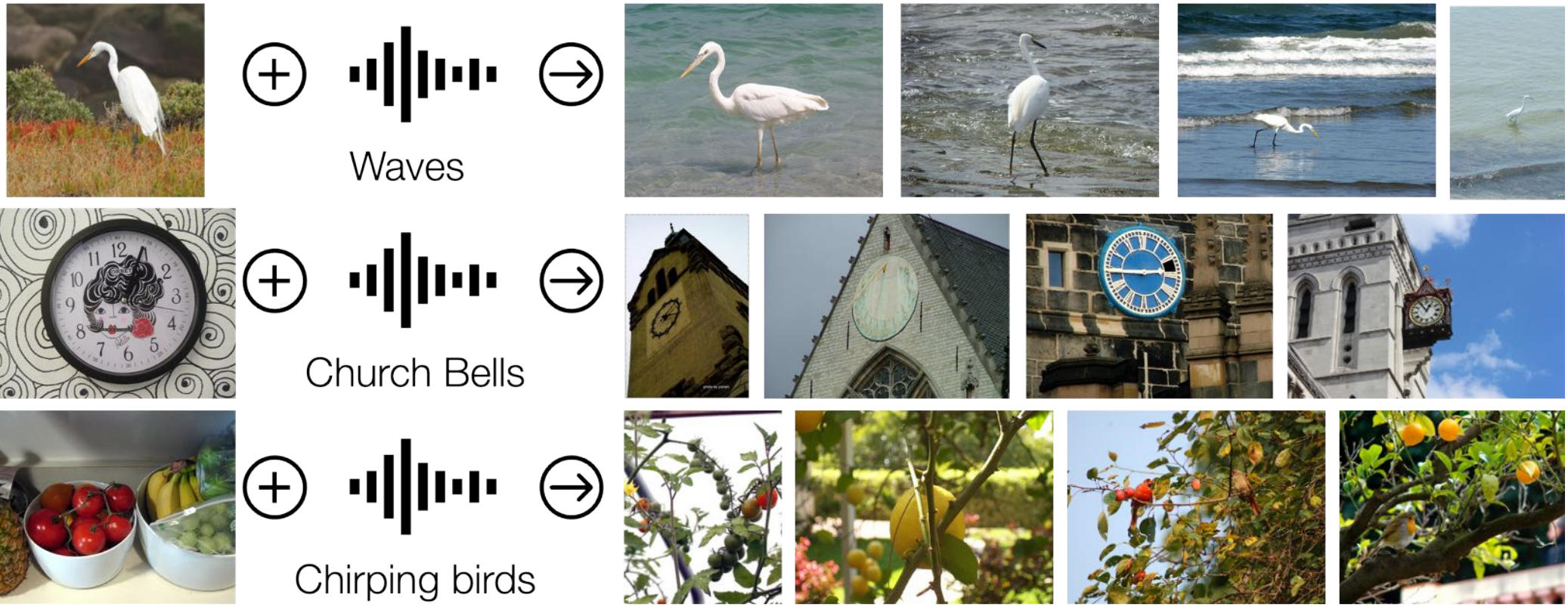
slide credit: Ishan Misra

Binding gets stronger with image-model size



slide credit: Ishan Misra

Aligned embeddings can be “added”



slide credit: Ishan Misra