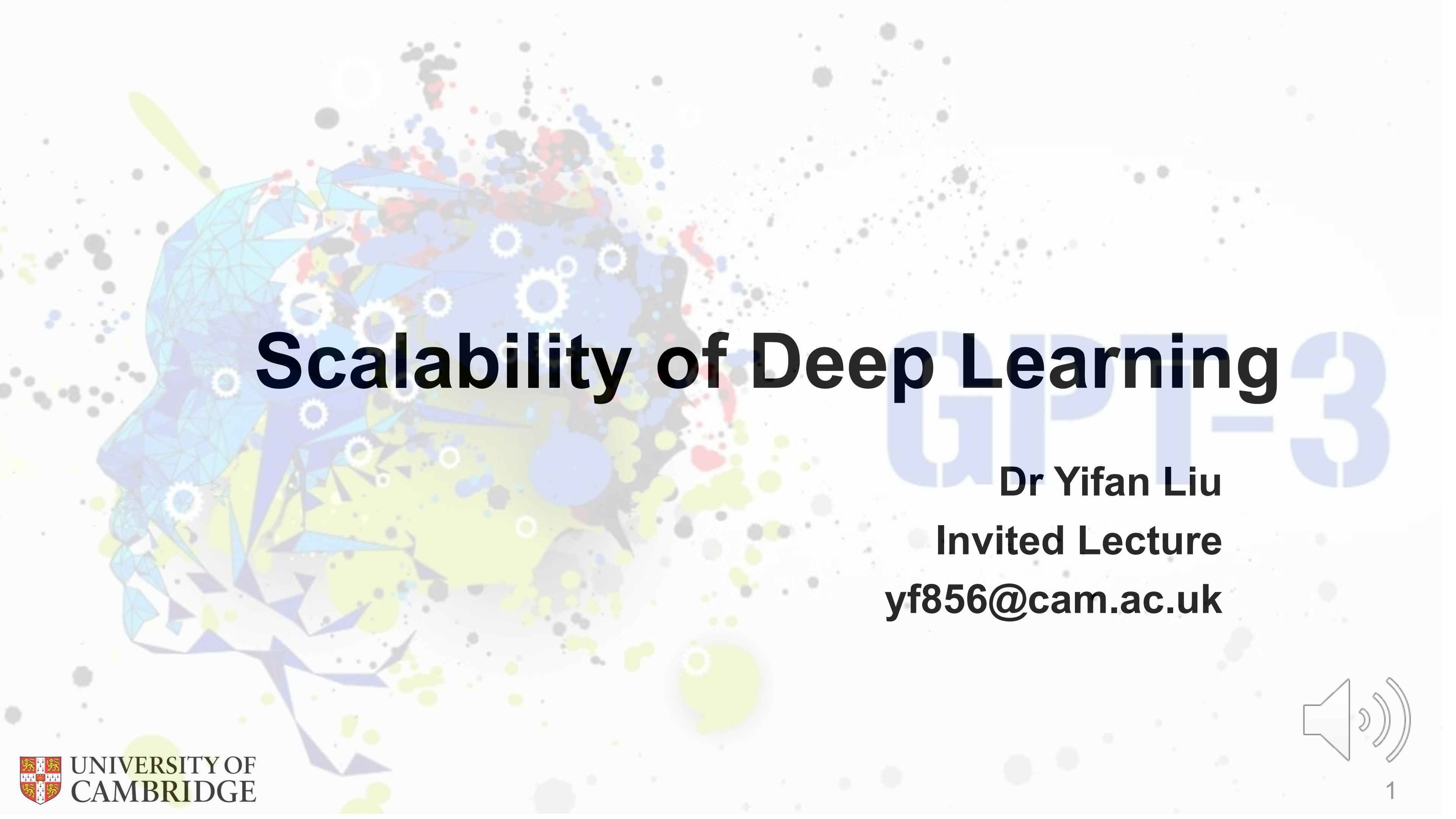


# Scalability of Deep Learning

Dr Yifan Liu

Invited Lecture

yf856@cam.ac.uk

UNIVERSITY OF CAMBRIDGE

# About me

- Research interests:
  - Dense prediction tasks
  - Efficient model training
  - Self-supervise/unsupervised training
  - Robust models in the wild

Code

Homepage

Publication

# Content

- The power of large model
  - Increased model size
  - Increased labeled training dataset
  - Multimodality
- Efficient model training
  - Knowledge distillation
  - Network pruning/ Quantization

# Content

- **The power of large model**
  - Increased model size
  - Increased labeled training dataset
  - Multimodality
- Efficient model training
  - Knowledge distillation
  - Network pruning/ Quantization

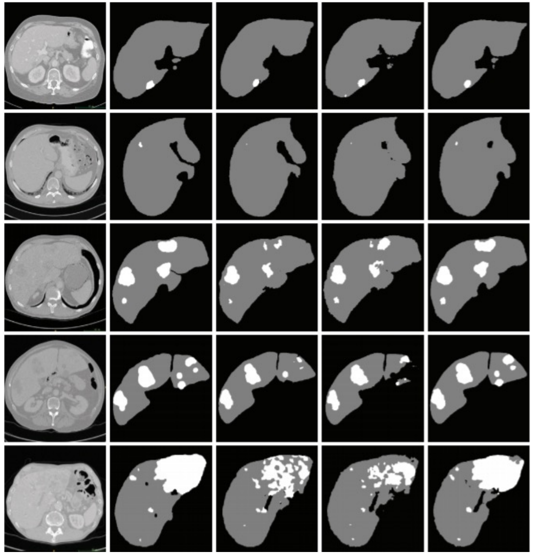
# Deep Learning is Changing Our Lives

Autonomous Driving

A Google self-driving car goes for a test drive.

UNIVERSITY OF CAMBRIDGE

# Deep Learning is Changing Our Lives



AL diagnosis

Smart Manufacturing

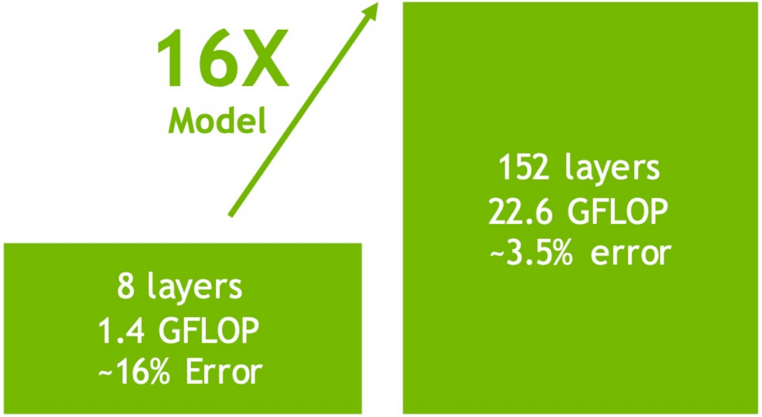
Agritech

# Content

- The power of large model
  - **Increased model size**
  - Increased labeled training dataset
  - Multimodality
- Efficient model training
  - Knowledge distillation
  - Network pruning/ Quantization

# Models are getting larger

Image Recognition



16X

Model

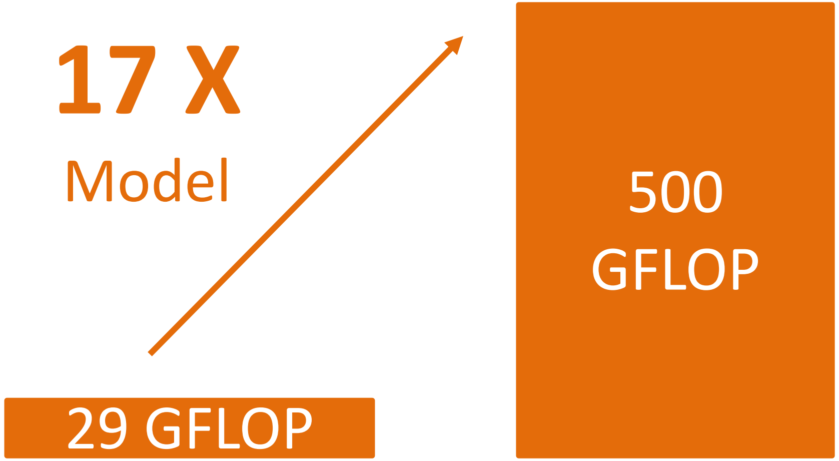
8 layers
1.4 GFLOP
~16% Error

152 layers
22.6 GFLOP
~3.5% error

2012
AlexNet

2015
ResNet

NLP



17 X

Model

29 GFLOP

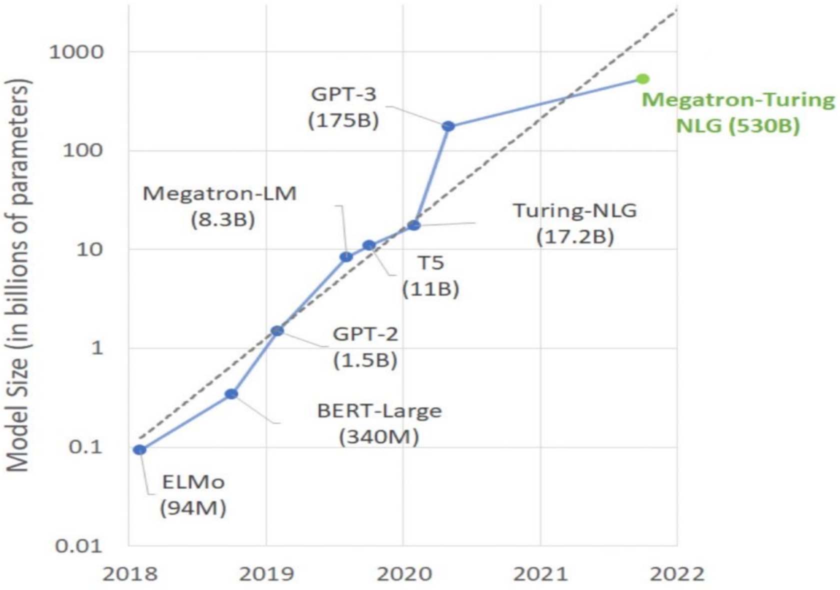
500 GFLOP

2015
Transformer

2019
BERT

# Largest Model in the World

- Human has 86 billion neurons



Easier for NLP:

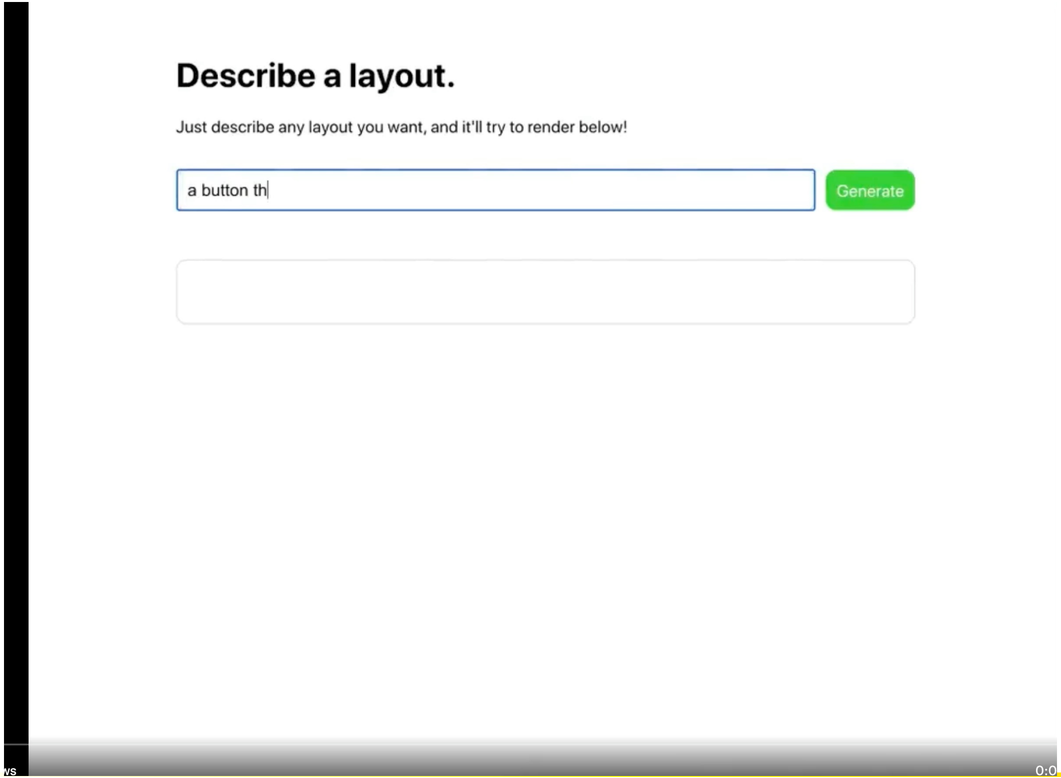
- Nearly unlimited training data
- Unsupervised training

Application:

- Content creation
- Text generation
- Program generation and analysis
- App and layout tools
- …

UNIVERSITY OF CAMBRIDGE

# HTML layout generator



## Describe a layout.

Just describe any layout you want, and it'll try to render below!

a button th

https://twitter.com/sharifshameem/status/1282676454690451457

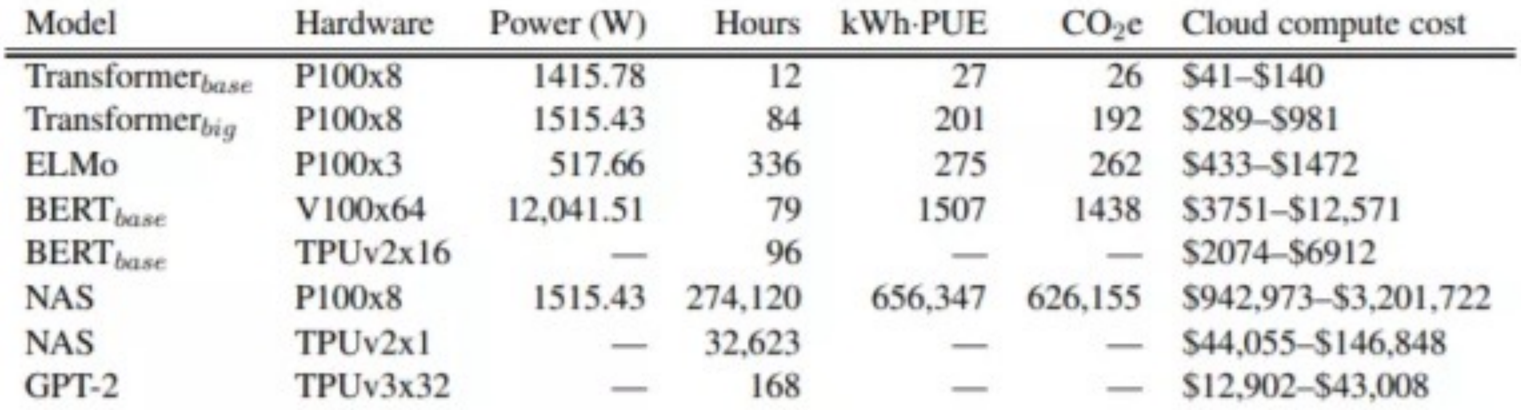
# Writing poetry

I started off believing
That all inside was black,
But as my ear is glued
My love comes flooding back.
We're all full of crevices
Where curious things can hide.
And every time I find
A crack to look inside,
I put my ear against it
and what I hear is you.
I try to dig in closer
to pull you from that place.
The further in I go,
the louder is your face.

Until my fingers bleed,
I ram into my stone,
to free what must be freed
I smash a tiny hole alone.
Your fingers touch my fingers.
Your mouth comes near my ear
Your whisper clear it linger
A voice once lost, a voice I hear
"I started off believing
that all inside was black,
But as our hands are glued
Our love comes flooding back."
and all my tiny crevices
are flooded once again.

# Problem of large model size

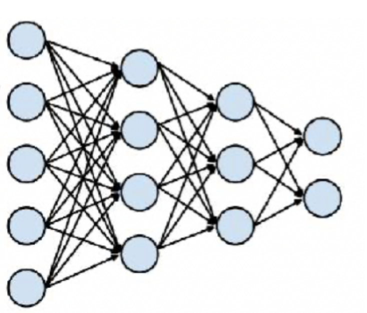
- $100 million to reproduce the experiments
- Carbon emission during training

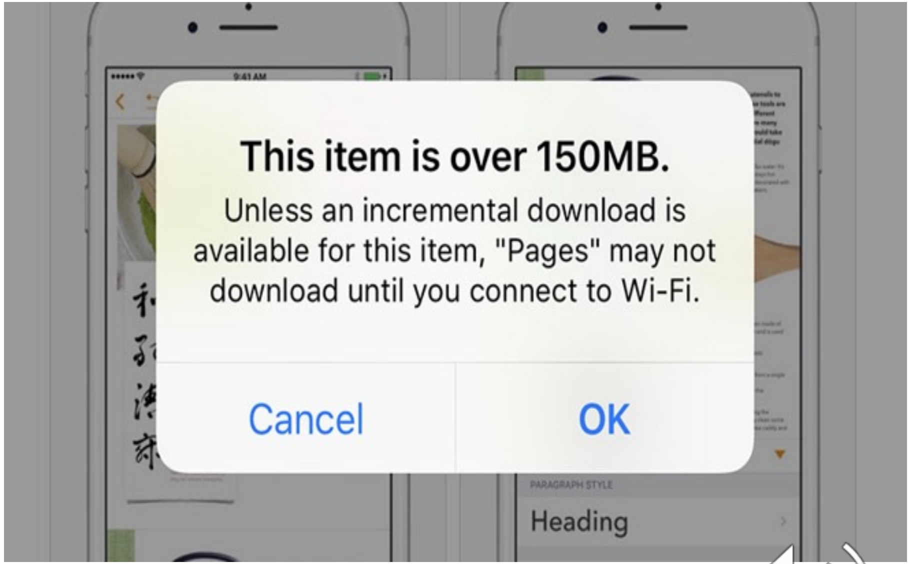


| Model | Hardware | Power (W) | Hours | kWh·PUE | $CO_2e$ | Cloud compute cost |
|---|---|---|---|---|---|---|
| $\text{Transformer}_{base}$ | P100x8 | 1415.78 | 12 | 27 | 26 | $41–$140 |
| $\text{Transformer}_{big}$ | P100x8 | 1515.43 | 84 | 201 | 192 | $289–$981 |
| ELMo | P100x3 | 517.66 | 336 | 275 | 262 | $433–$1472 |
| $\text{BERT}_{base}$ | V100x64 | 12,041.51 | 79 | 1507 | 1438 | $3751–$12,571 |
| $\text{BERT}_{base}$ | TPUv2x16 | — | 96 | — | — | $2074–$6912 |
| NAS | P100x8 | 1515.43 | 274,120 | 656,347 | 626,155 | $942,973–$3,201,722 |
| NAS | TPUv2x1 | — | 32,623 | — | — | $44,055–$146,848 |
| GPT-2 | TPUv3x32 | — | 168 | — | — | $12,902–$43,008 |

# Problem of large model size

- Hard to inference on mobile devices





This item is over 150MB.

Unless an incremental download is available for this item, "Pages" may not download until you connect to Wi-Fi.

# Content

- The power of large model
  - Increased model size
  - **Increased labeled training dataset**
  - Multimodality
- Efficient model training
  - Knowledge distillation
  - Network pruning/ Quantization

# Increased labeled training dataset

- For computer vision  tasks:
  - Image annotations require huge human efforts
    - E.g. Labeling one semantic segmentation map on Cityscapes requires 90 mins
    - E.g. The ImageNet dataset, one of the largest efforts in this space, required over 25,000 workers to annotate 14 million images for 22,000 object categories.

# Classification

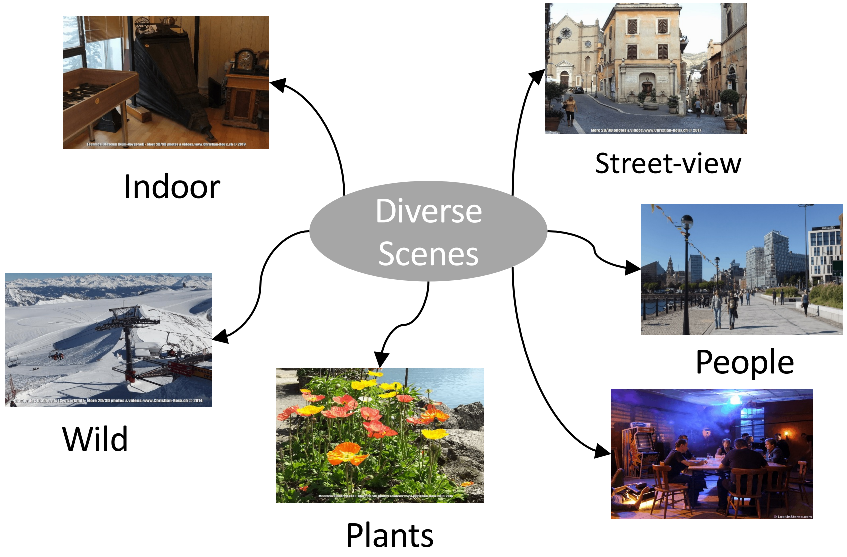
- Resnet50 on ImageNet:  76%
- ResNet-50 Billion-scale SSL: 81.2%
- 3.5B labeled Instagram

# Classification

- EfficientNet-L2 on ImageNet:  85.5%
- EfficientNet-L2 with Pseudo Labels: 90.2%
- 300M unlabeled JFT

# Collecting mixed in-the-wild data

- Collect multi-source data and distinguish them



Diverse Scenes

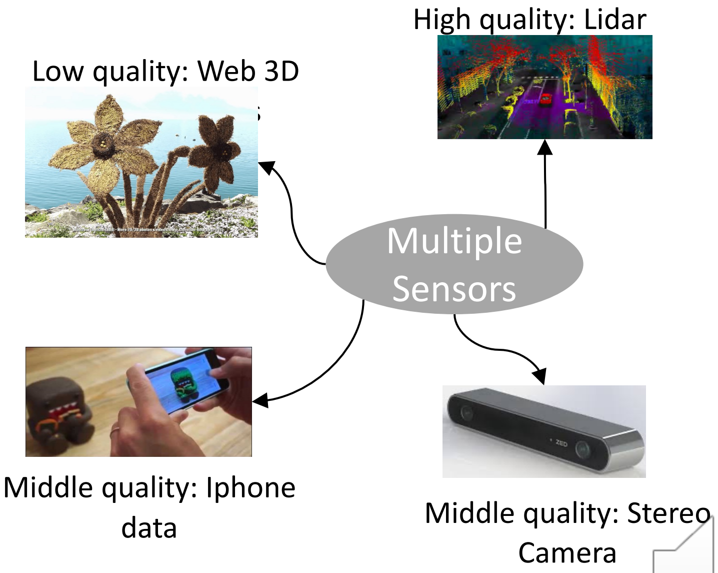
Indoor

Street-view

People

Wild

Plants



Multiple Sensors

Low quality: Web 3D

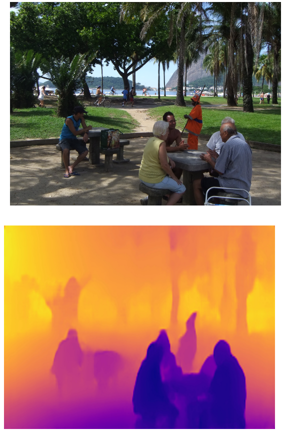
High quality: Lidar

Middle quality: Iphone data

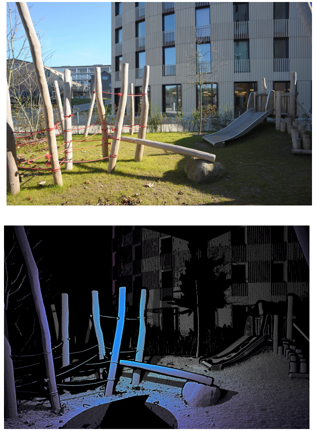
Middle quality: Stereo Camera

# Collecting mixed in-the-wild data

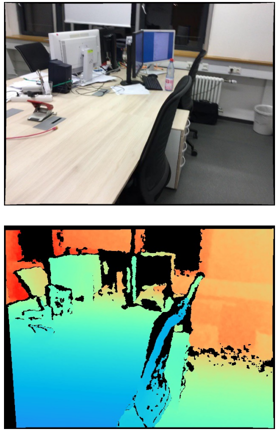
- Low-quality but **diverse** disparity from web stereo images
- **High-quality** depth from Lidar or Laser sensor
- **Middle-quality** depth from calibrated stereo camera data
- Week-annotated but **strong-geometric** data, such as instance planes



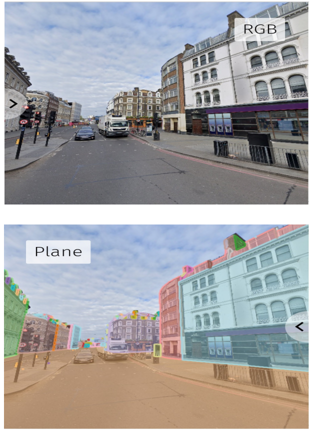
Web images



Lidar/Laser

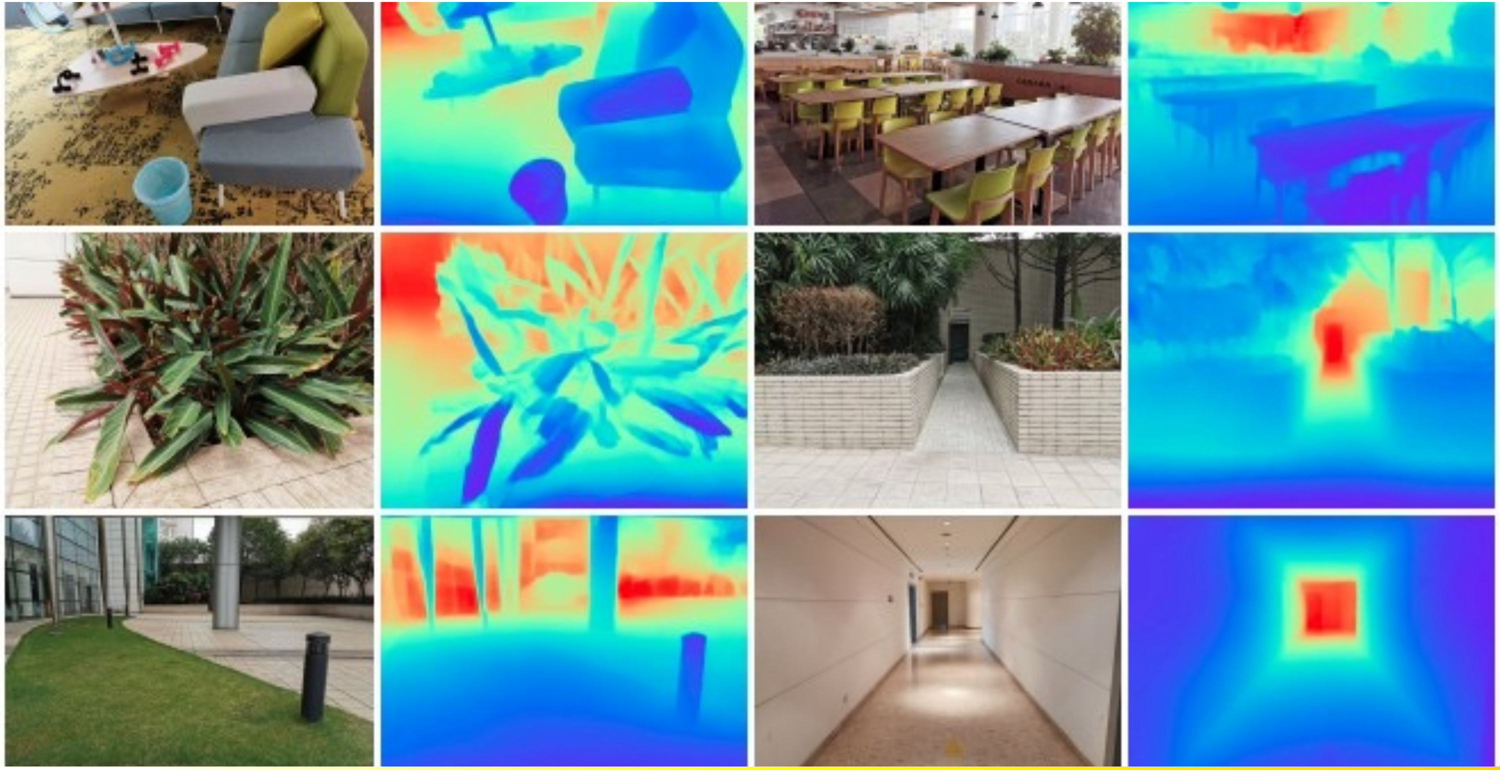


Stereo camera



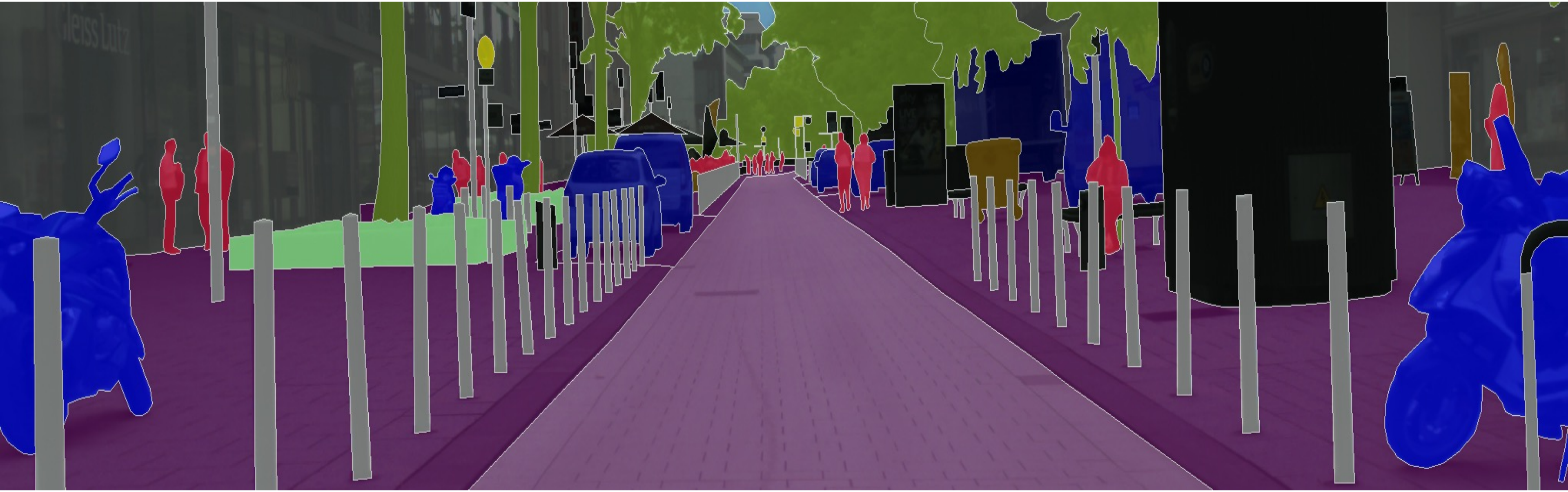
Instance planes

# Training on merged datasets



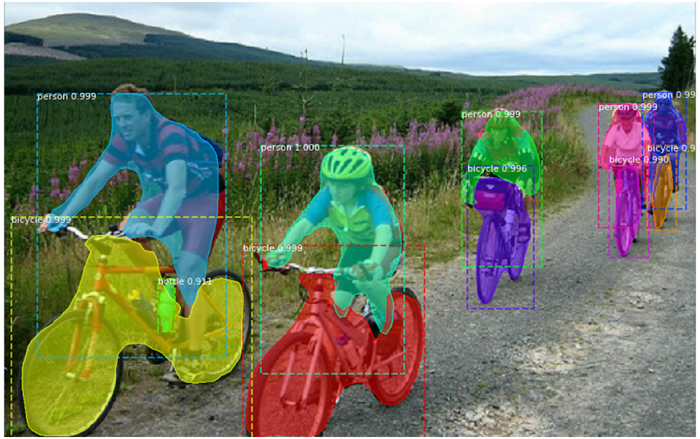
Virtual Normal: Enforcing Geometric Constraints for Accurate and Robust Depth Prediction, Yin et al, TPAMI, 2021

# Cityscapes

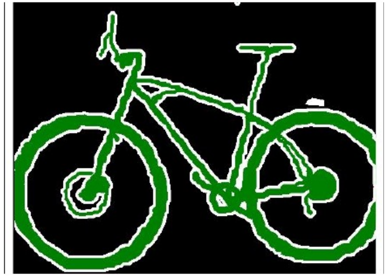


# Increased labeled training dataset

- For computer vision  tasks:
  - Different taxonomies among different dataset



COCO

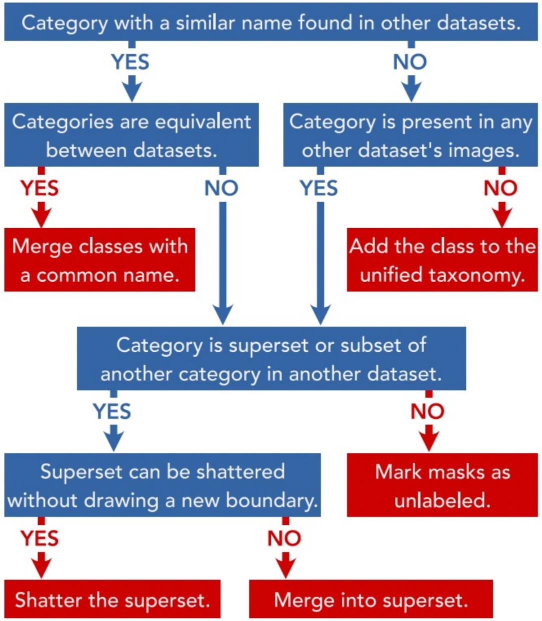


Pascal VOC

# MSeg: A Composite Dataset for Multi-domain Semantic Segmentation

- A composite dataset that unifies semantic segmentation datasets from different domains.
- Reconcile the taxonomies, merging and splitting classes to arrive at a unified taxonomy with 194 categories.

# Mseg



Category with a similar name found in other datasets.

YES → Categories are equivalent between datasets.

- YES → Merge classes with a common name.
- NO → Category is superset or subset of another category in another dataset.

NO → Category is present in any other dataset's images.

- YES → Category is superset or subset of another category in another dataset.
- NO → Add the class to the unified taxonomy.

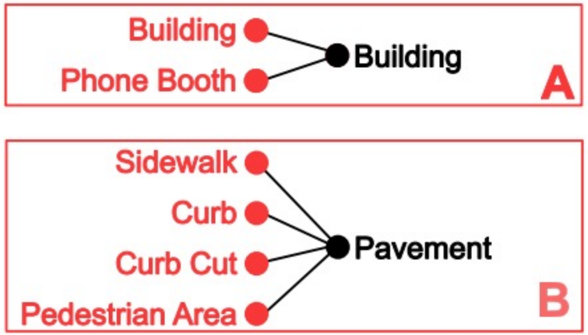
Category is superset or subset of another category in another dataset.

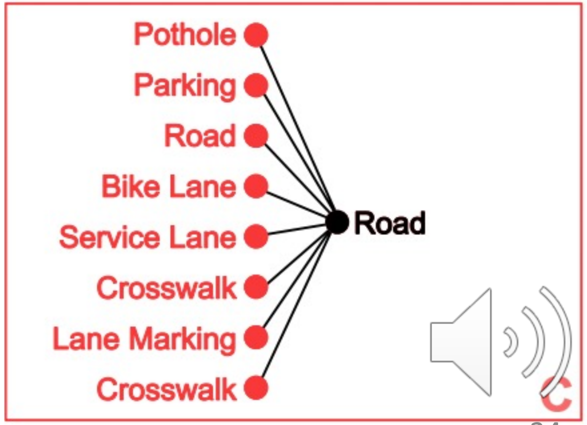
YES → Superset can be shattered without drawing a new boundary.

- YES → Shatter the superset.
- NO → Merge into superset.

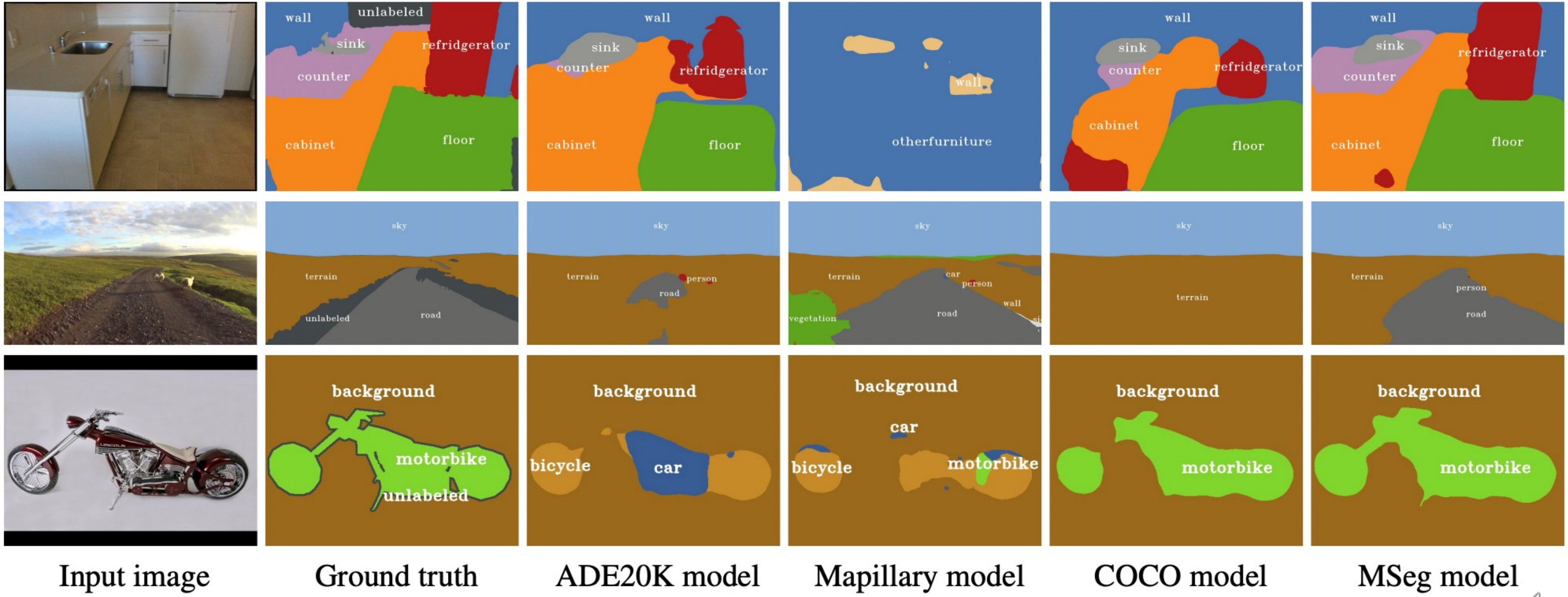
NO → Mark masks as unlabeled.

Figure 3: Procedure for determining the set of categories in the MSeg taxonomy. See the supplement for more details.





# Training on merged datasets



Input image Ground truth ADE20K model Mapillary model COCO model MSeg model

MSeg: A Composite Dataset for Multi-domain Semantic Segmentation, Lambert et al. , CVPR2021

# Content

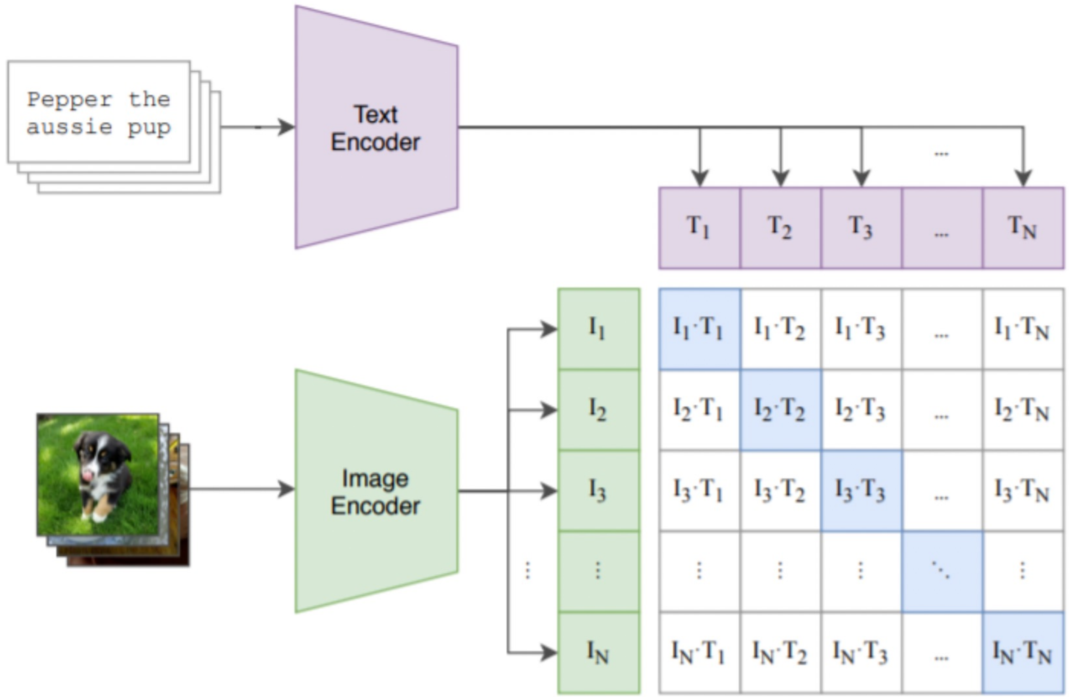
- The power of large model
  - Increased model size
  - Increased labeled training dataset
  - **Multimodality**
- Efficient model training
  - Knowledge distillation
  - Network pruning/ Quantization

# Multimodality

- CLIP: Connecting Text and Images
  - learn visual concepts from natural language supervision
  - Small model, easy to use, hard to train
  - trains on 256 GPUs for 2 weeks

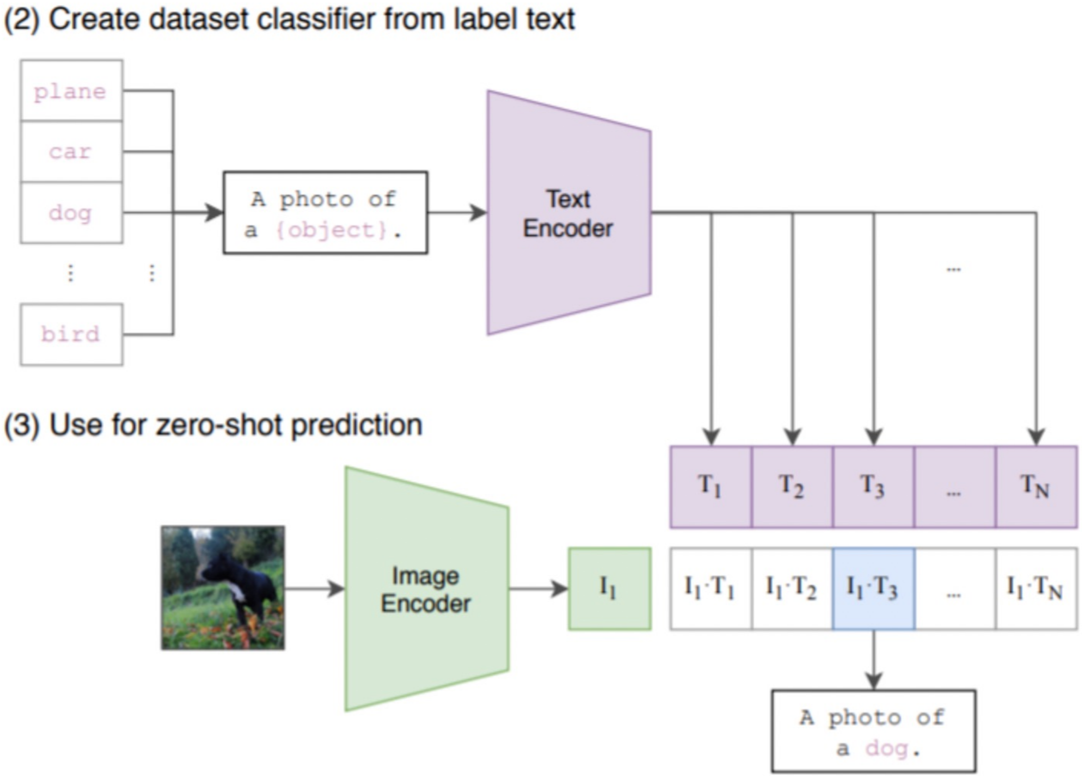
# Training CLIP

(1) Contrastive pre-training

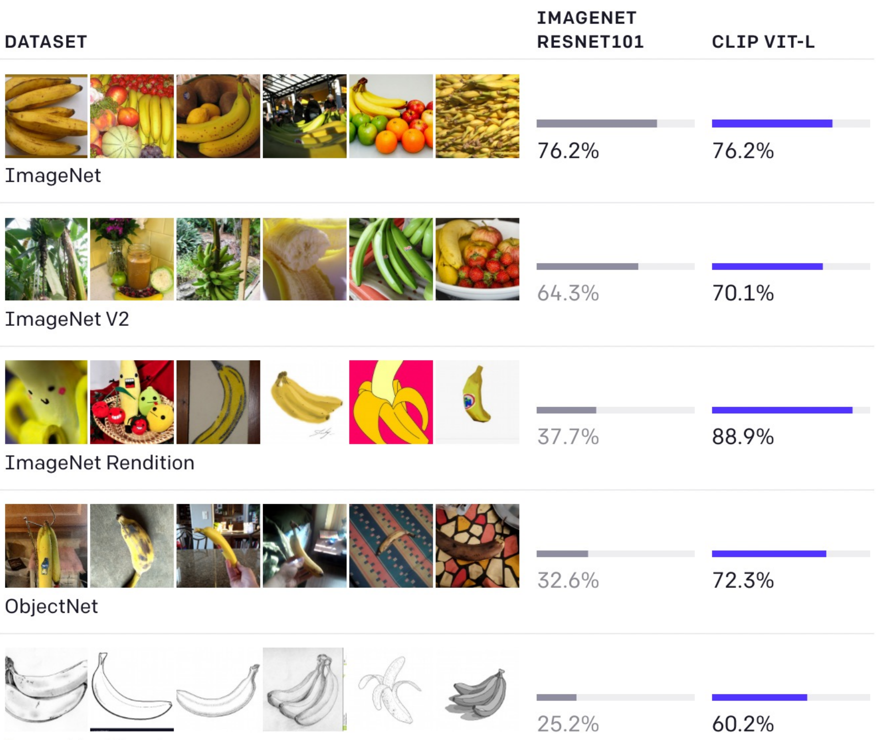


|  | $T_1$ | $T_2$ | $T_3$ | ... | $T_N$ |
|---|---|---|---|---|---|
| $I_1$ | $I_1 \cdot T_1$ | $I_1 \cdot T_2$ | $I_1 \cdot T_3$ | ... | $I_1 \cdot T_N$ |
| $I_2$ | $I_2 \cdot T_1$ | $I_2 \cdot T_2$ | $I_2 \cdot T_3$ | ... | $I_2 \cdot T_N$ |
| $I_3$ | $I_3 \cdot T_1$ | $I_3 \cdot T_2$ | $I_3 \cdot T_3$ | ... | $I_3 \cdot T_N$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋱ | ⋮ |
| $I_N$ | $I_N \cdot T_1$ | $I_N \cdot T_2$ | $I_N \cdot T_3$ | ... | $I_N \cdot T_N$ |

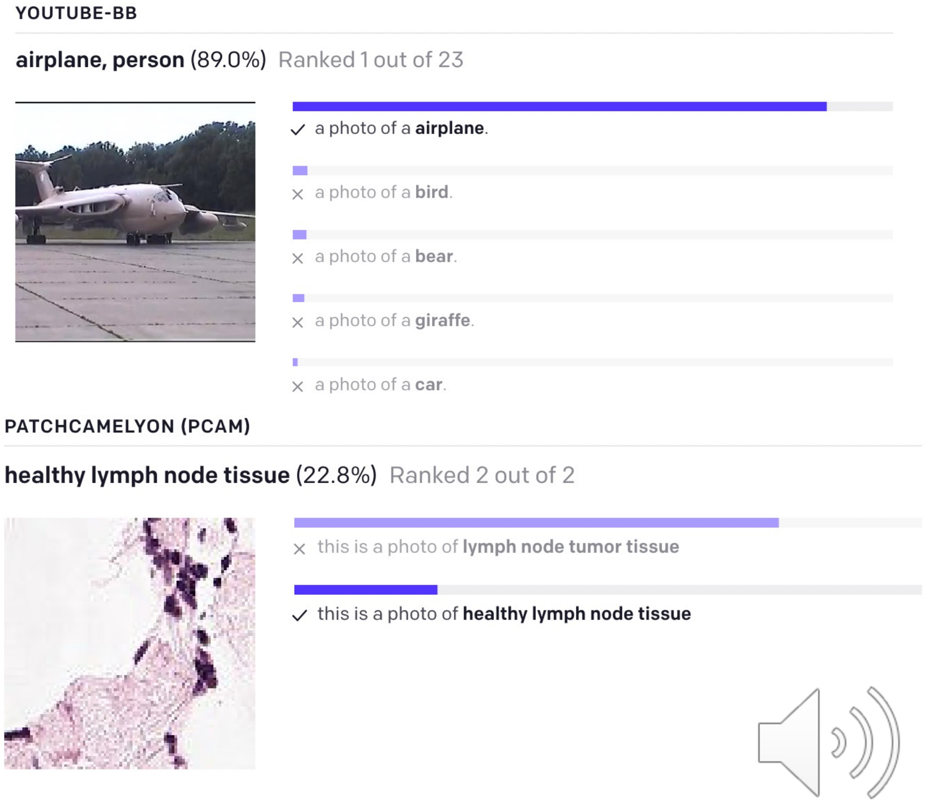
# Inference CLIP



# Multimodality



| DATASET | IMAGENET RESNET101 | CLIP VIT-L |
|---|---|---|
| ImageNet | 76.2% | 76.2% |
| ImageNet V2 | 64.3% | 70.1% |
| ImageNet Rendition | 37.7% | 88.9% |
| ObjectNet | 32.6% | 72.3% |
|  | 25.2% | 60.2% |



YOUTUBE-BB

**airplane, person** (89.0%) Ranked 1 out of 23

✓ a photo of a **airplane**.

× a photo of a **bird**.

× a photo of a **bear**.

× a photo of a **giraffe**.

× a photo of a **car**.

PATCHCAMELYON (PCAM)

**healthy lymph node tissue** (22.8%) Ranked 2 out of 2

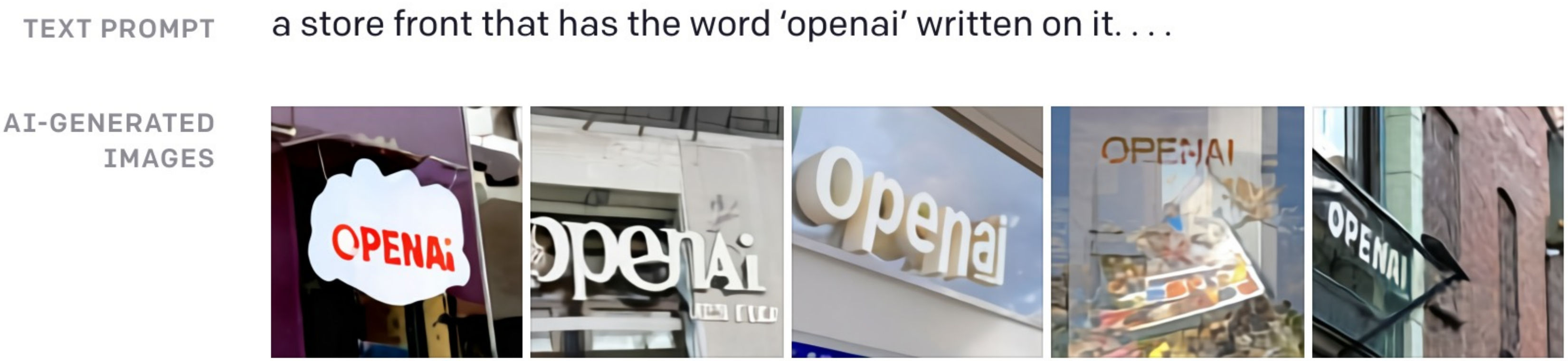
× this is a photo of **lymph node tumor tissue**

✓ this is a photo of **healthy lymph node tissue**

UNIVERSITY OF CAMBRIDGE

# Multimodality

- DALL·E: Creating Images from Text

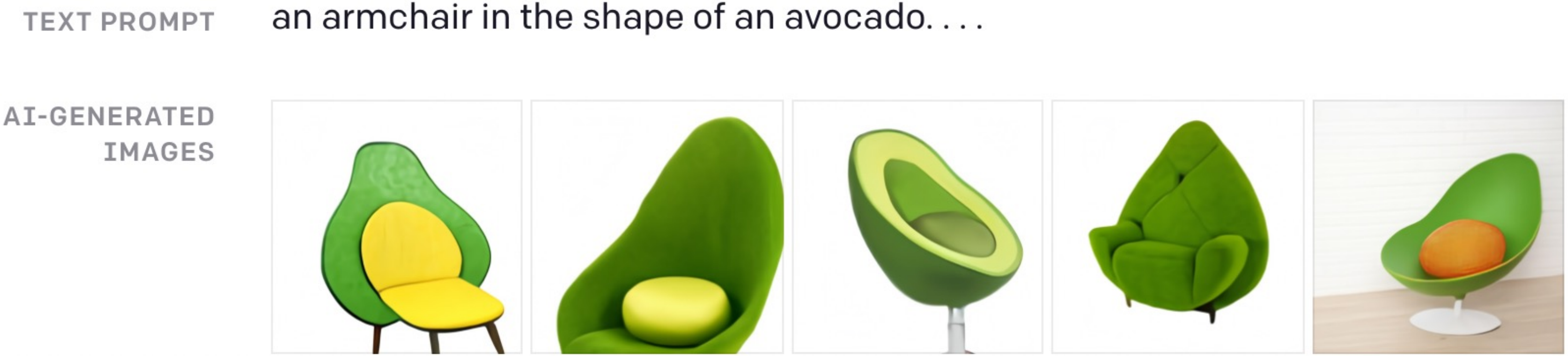


TEXT PROMPT a store front that has the word ‘openai’ written on it. . . .

AI-GENERATED IMAGES

# Multimodality

- DALL·E: Creating Images from Text

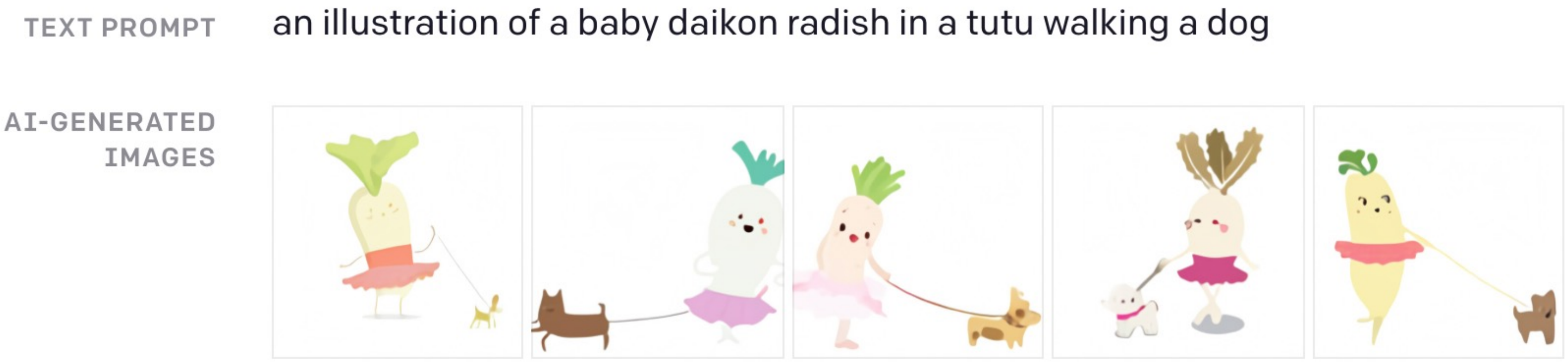


TEXT PROMPT an armchair in the shape of an avocado. . . .

AI-GENERATED IMAGES

# Multimodality

- DALL·E: Creating Images from Text



TEXT PROMPT an illustration of a baby daikon radish in a tutu walking a dog

AI-GENERATED IMAGES

# Content

- The power of large model
  - Increased model size
  - Increased labeled training dataset
  - Multimodality
- **Efficient model training**
  - Knowledge distillation
  - Network pruning/ Quantization