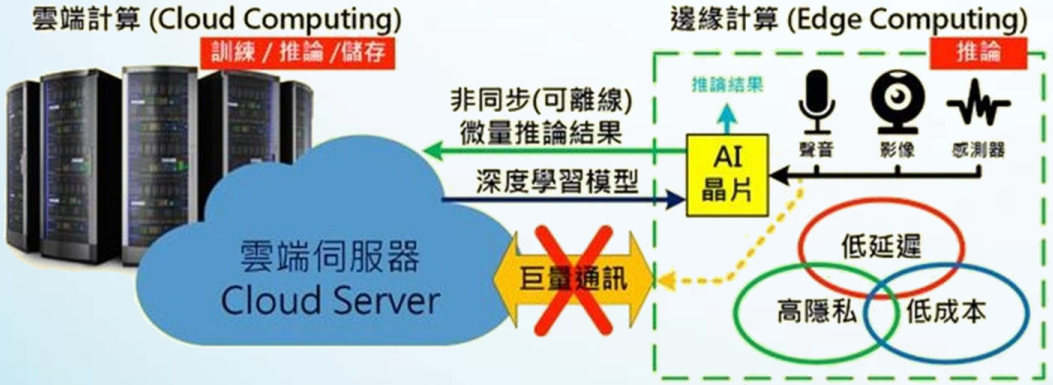
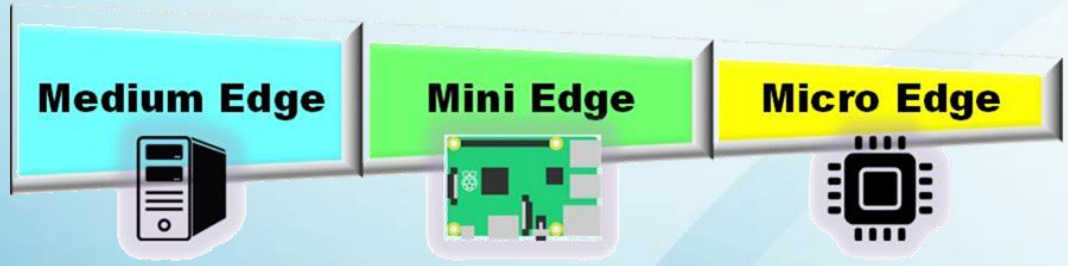
OmniXRI's Edge AI & TinyML 小學堂





【第3講】
資料集建置與標註

歐尼克斯實境互動工作室 (OmniXRI Studio)
許哲豪 (Jack Hsu)

簡報大綱

- 3.1. 資料集建置
- 3.2. 公開資料集
- 3.3. 資料集標註
- 3.4. 資料集迷思

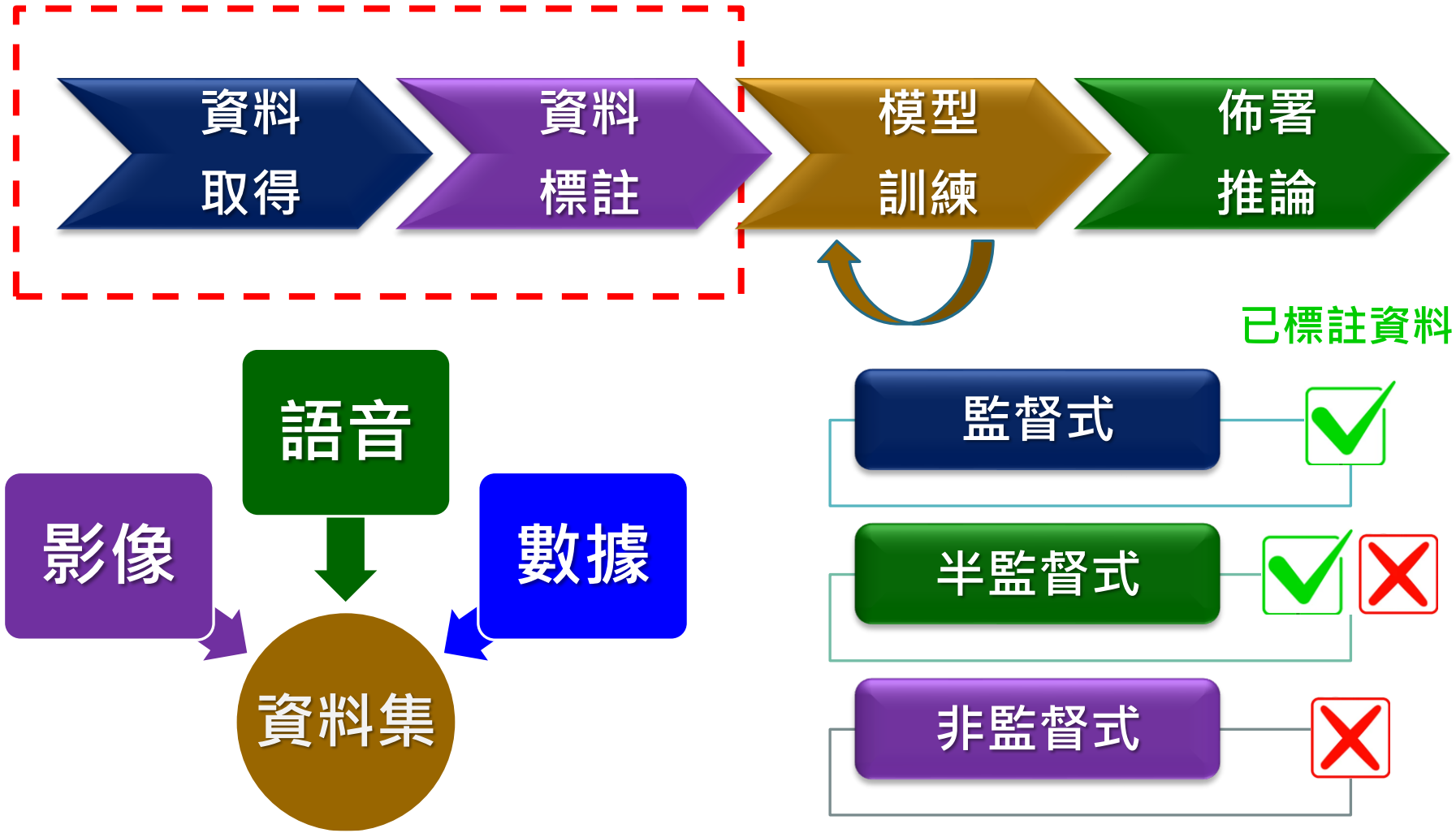
本課程完全免費，請勿移作商業用途！
歡迎留言、訂閱、點讚、轉發，讓更多需要的朋友也能一起學習。

完整課程大綱： https://omnixri.blogspot.com/2024/02/omnixris-edge-ai-tinyml-0.html
課程直播清單： https://www.youtube.com/@omnixri1784/streams

3.1.
資料集建置

- 工作流程
- 資料類型
- 資料取得
- 資料擴增

工作流程



資料類型 – 應用情境

電腦視覺(影像、影片)

- 影像分類
- 物件偵測
- 語義 / 實例分割
- 文字辨識(OCR）
- 生物辨別(人臉、指紋...)
- 圖像標題
- 影像生成、合成
- 行為(骨架)分析
- 軌跡預測
- 品質檢測(瑕疵、分級...)

自然語言

- 語音、文字互換
- 語音客服、助理
- 語義(文章)理解
- 語言翻譯

數據分析

- 時序預測
- 生產優化
- 回歸分析
- 推薦系統
- 資訊安全

資料型態 – 資料結構

格式化資料

- 一維時序資料（感測器）

  光電、溫度、濕度、壓力、

  聲音、速度、運動...
- 二維 / 三維影像資料

  可見光、紅外光、深度影像
- 影像序列資料（影片）

  原始、壓縮、串流
- 多維表單數據（資料庫）

非格式化資料

- 自然語言、文章
- 影像、影片內容
- 非定時定量收集的資料
- 偶發性資料

資料型態 – 感測器

影像來源：https://www.edntaiwan.com/20190201ta31-designers-guide-to-iiot-sensor-systems/

- A. 電阻、電容、電感式
- B. 光電式
  - 發射接收式、紅外熱幅射式
- C. 壓電式
- D. 電聲式
  - 超音波、麥克風
- E. 微機電式
  - 運動類、環境類
- F. 電磁波式
- G. 影像式

資料來源：https://omnixri.blogspot.com/2022/11/20221108.html

資料取得

資料集來源

- 公開資料集
- 私有資料庫
- 網路收集
- 自行拍攝取像
  - 專業相機（手機、相機、攝影機）
  - 網路攝影機（USB）
  - 開發板專用相機（CSI）
    - 可見光
    - 紅外線
- 感測器資料
  - 網路連接、MQTT

取像要領

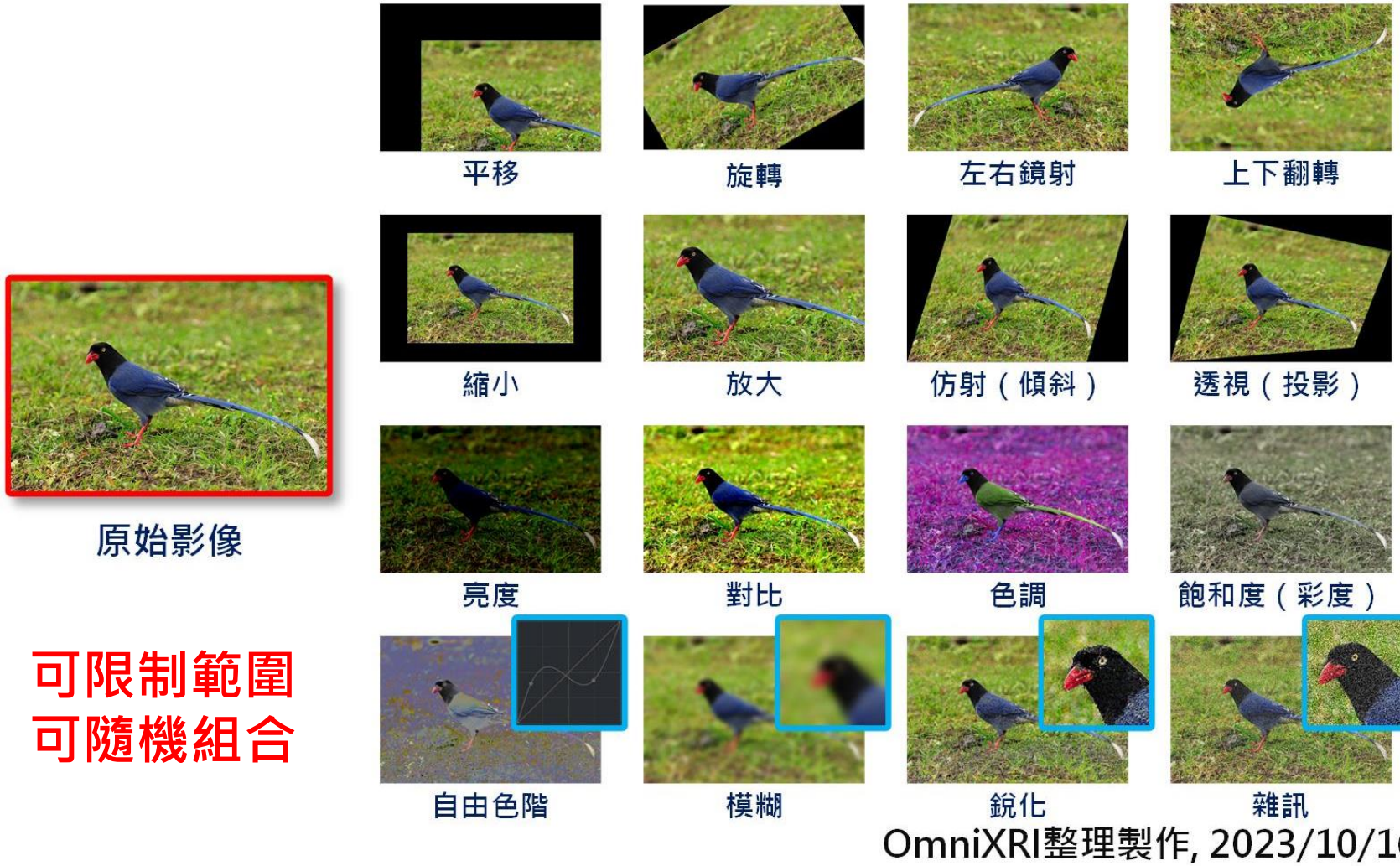
- 取像品質近似
  - 解析度、模糊度...
- 多視點
  - 遠近、角度、 物件正反面...
- 多光源差異
  - 環境亮度、陰影、光圈、快門...
- 多背景、品種

**** 建立足夠多樣性的資料集 ****

資料清洗

- 去冗、填空、修補、抑雜訊

資料擴增 – 影像處理式



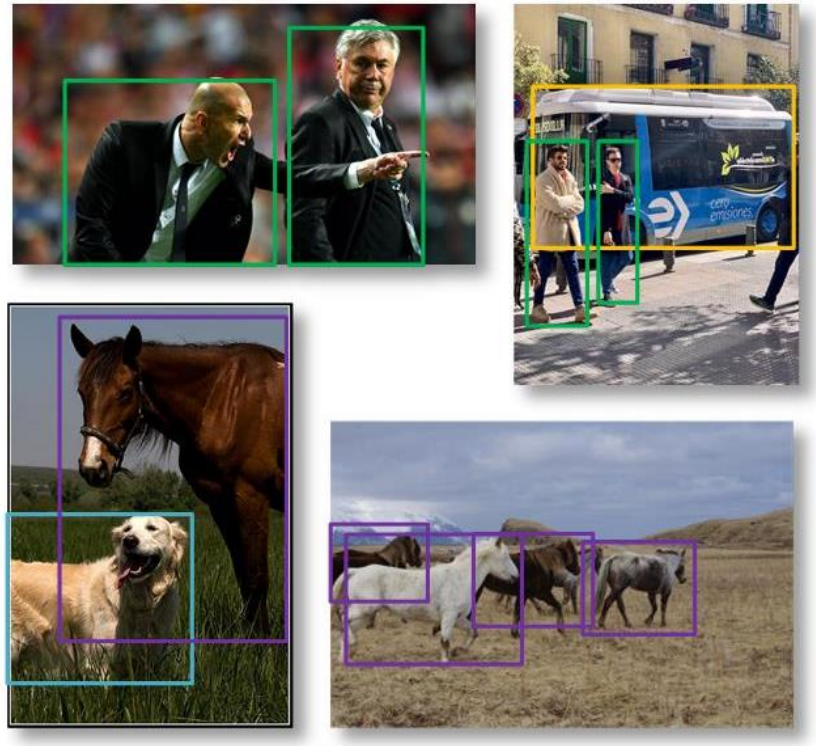
協同工具

OpenCV

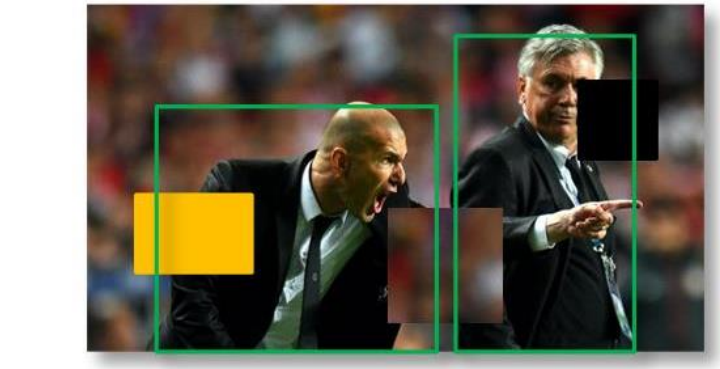
XnView

資料來源：https://omnixri.blogspot.com/2023/10/vmaker-edge-ai-10-ai.html

資料擴增 – 影像拼貼式



原始影像及物件框



適合物件偵測
可內嵌至訓練

馬賽克拼貼

OmniXRI整理製作, 2023/10/16

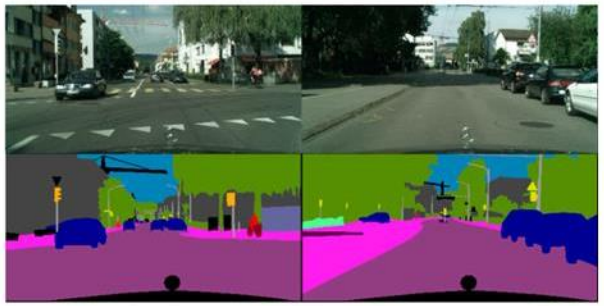
資料來源：https://omnixri.blogspot.com/2023/10/vmaker-edge-ai-10-ai.html

資料擴增 – 影像生成式

生成對抗網路
(Generative Adversarial Networks, GAN)

(Mixed & GAN Face)

立體渲染 / 數位孿生
(3D Rending / Digital Twins)

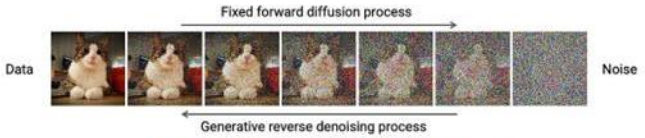


(3D Rendering & Object Segmentation)

(Background + 3D Models)

人工智慧生成內容
(AI Generated Content, AIGC)

(Real vs. Generated)



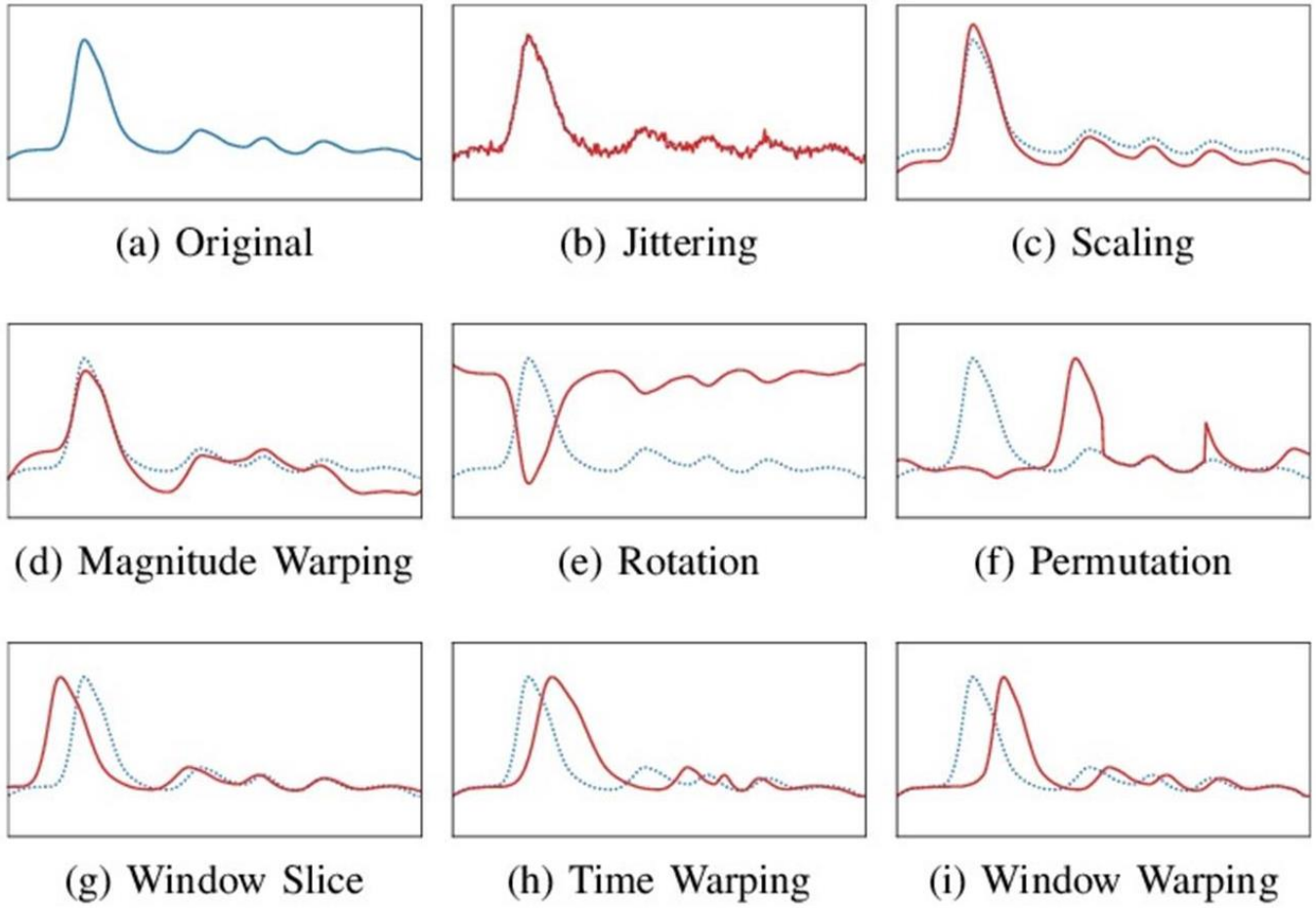
(Diffusion vs. Denoising)

OmniXRI整理製作, 2023/10/16

須耗費大量算力，無法即時產生。

資料來源：https://omnixri.blogspot.com/2023/10/vmaker-edge-ai-10-ai.html

資料擴增 – 時序資料



Src: https://arxiv.org/abs/2004.08780

- 抖動(Jittering)
- 縮放(Scaling)
- 振幅扭曲(Magnitude Warping)
- 信號反轉(Rotation)
- 排列置換(Permutation)
- 視窗切片(Window Slice)
- 時間扭曲(Time Warping)
- 視窗扭曲(Window Warping)

OmniXRI整理製作, 2023/10/16

資料來源：https://omnixri.blogspot.com/2023/10/vmaker-edge-ai-10-ai.html

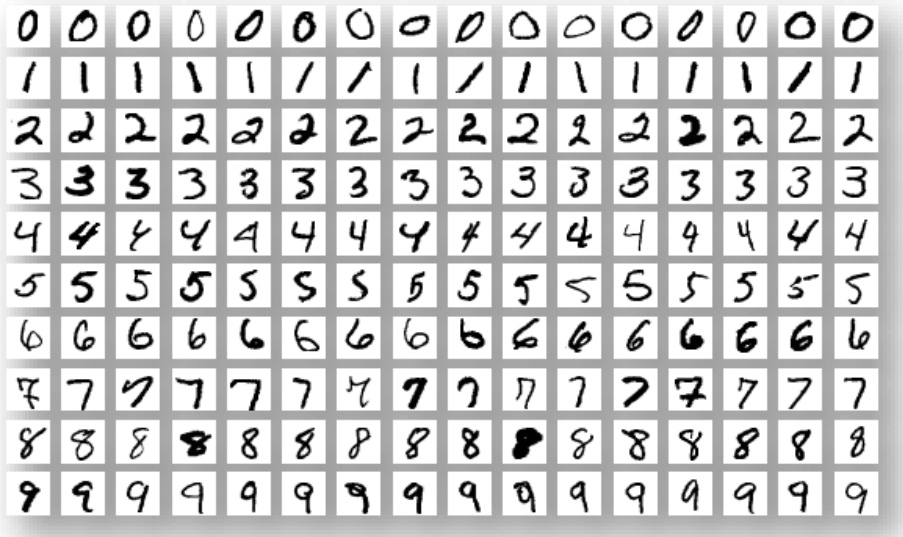
3.2.
公開資料集

- 常見影像資料集
- 常見聲音資料集
- 其它公開資料集

常見影像資料集 – MNIST / CIFAR-10

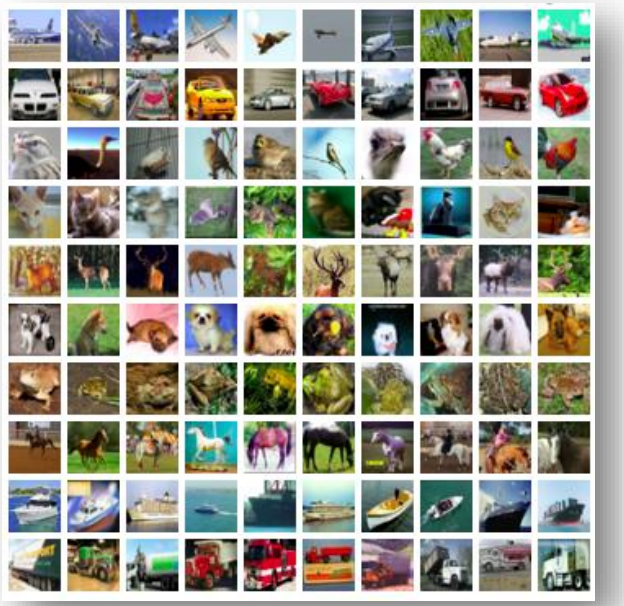
MNIST 手寫數字

- 28*28灰階影像
- 共分十類，數字 0 ~ 9
- 6萬張訓練集，1萬張測試集



CIFAR-10 彩色影像

- 32*32 彩色影像
- 共分十類，飛機、 貓、狗...
- 6萬張，每類6千張影像



常見影像資料集 – Pascal VOC

Pattern
Analysis
Statical Modeling and
Computa**t**ional
Learning
Visual
Object
Classes

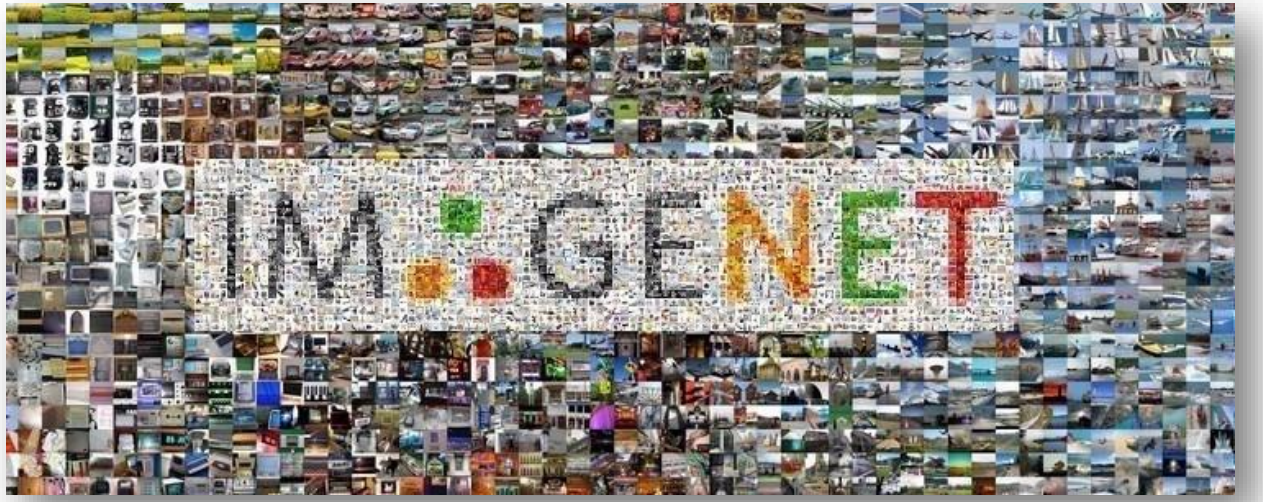
Pascal VOC (2005 ~ 2012)

共有1萬7千多張影像，分為20類，標註內容包括影像分類、物件偵測、語義分割等。其標註資料主要採用**XML**格式。

http://host.robots.ox.ac.uk/pascal/VOC/

常見影像資料集 – ImageNet

ImageNet
Large
Scale
Visual
Recognition
Challenge



ImageNet (2010 ~ 2017)

共有超過1400萬張影像，透過Amazon Mechanical Turk外包協助下進行手動標註，包含2萬多個類別，超過100萬個物件偵測邊界框(Bounding Box)被標註。

http://image-net.org/

常見影像資料集 – Microsoft COCO

Common

Objects

In

Contex

MS COCO (2015 ~)

共有超過32萬張影像，包含91(**80**)個類別，超過250萬個物件測邊界框被標註。其標註資料主要採用**JSON**格式。

https://cocodataset.org/

常見聲音資料集 – Speech / ESC-50

Google Speech Command

常用30多種英文命令，數字、上下左左右前後、開關等。

Word	Number of Utterances
Backward	1,664
Bed	2,014
Bird	2,064
Cat	2,031
Dog	2,128
Down	3,917
Eight	3,787
Five	4,052
Follow	1,579
Forward	1,557
Four	3,728
Go	3,880
Happy	2,054
House	2,113
Learn	1,575
Left	3,801
Marvin	2,100
Nine	3,934
No	3,941
Off	3,745
On	3,845
One	3,890
Right	3,778
Seven	3,998
Sheila	2,022
Six	3,860
Stop	3,872
Three	3,727
Tree	1,759
Two	3,880
Up	3,723
Visual	1,592
Wow	2,123
Yes	4,044
Zero	4,052

https://www.tensorflow.org/datasets/catalog/speech_commands

ESC-50 環境音資料集

生活中常見五大類聲音

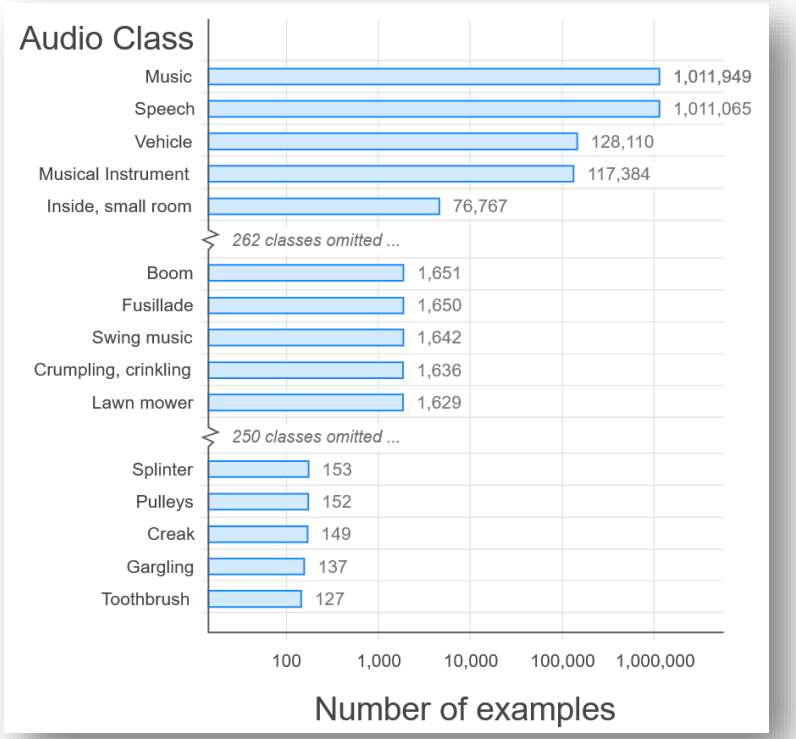
Animals	Natural soundscapes & water sounds	Human, non-speech sounds	Interior/domestic sounds	Exterior/urban noises
Dog	Rain	Crying baby	Door knock	Helicopter
Rooster	Sea waves	Sneezing	Mouse click	Chainsaw
Pig	Crackling fire	Clapping	Keyboard typing	Siren
Cow	Crickets	Breathing	Door, wood creaks	Car horn
Frog	Chirping birds	Coughing	Can opening	Engine
Cat	Water drops	Footsteps	Washing machine	Train
Hen	Wind	Laughing	Vacuum cleaner	Church bells
Insects (flying)	Pouring water	Brushing teeth	Clock alarm	Airplane
Sheep	Toilet flush	Snoring	Clock tick	Fireworks
Crow	Thunderstorm	Drinking, sipping	Glass breaking	Hand saw

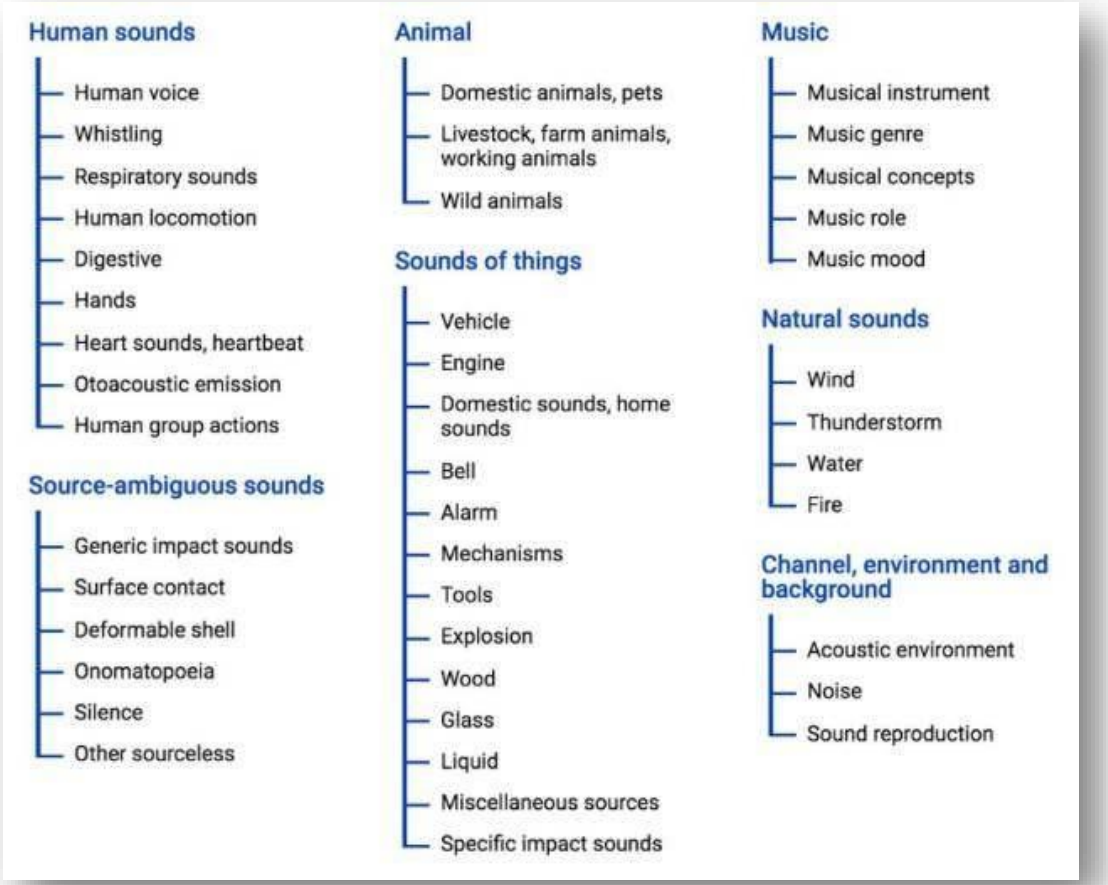
https://github.com/karolpiczak/ESC-50

常見聲音資料集 – AudioSet

Google AudioSet

210萬已標註影片，5800小時聲音，
527類聲音分類。



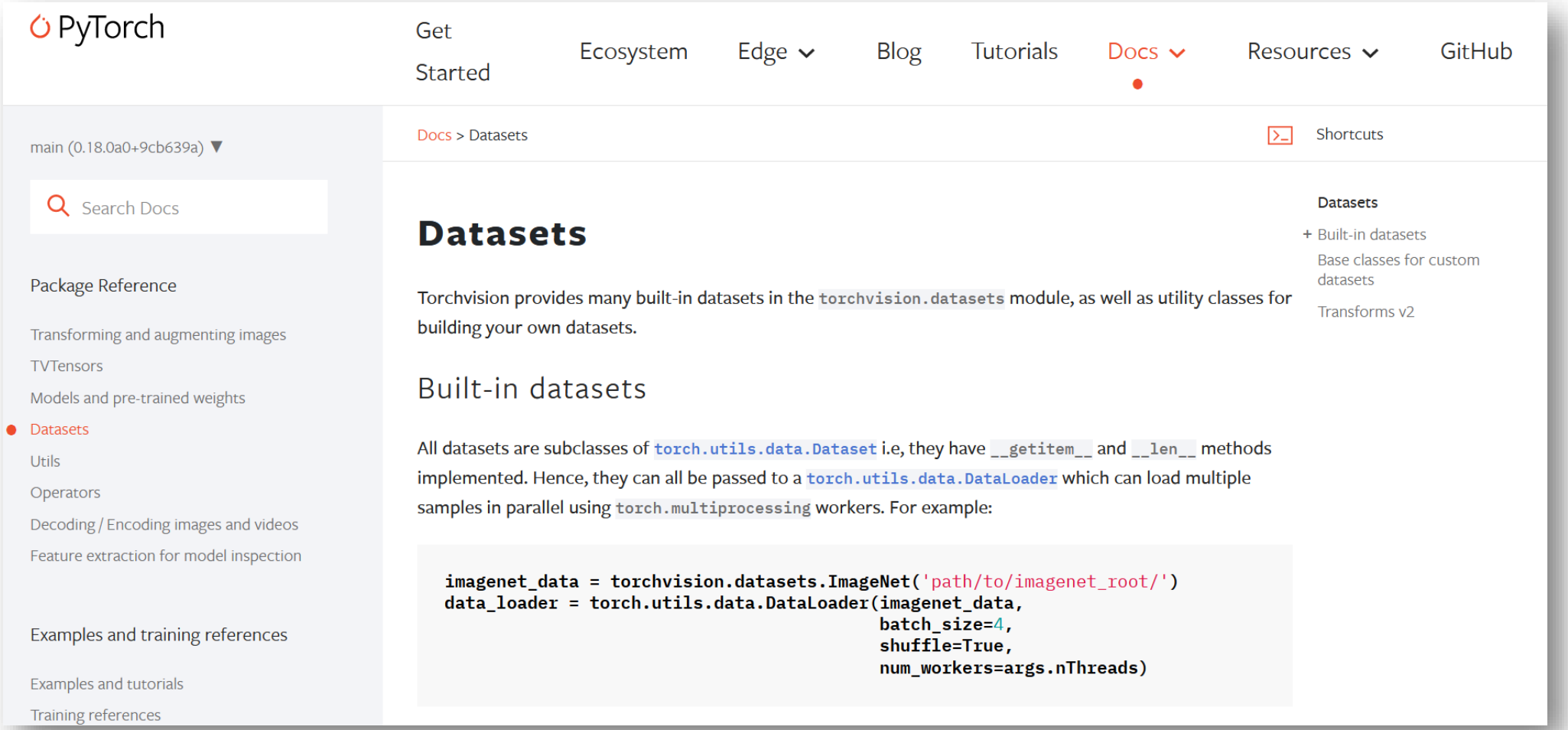


https://research.google.com/audioset/index.html

其它公開資料集 – TensorFlow

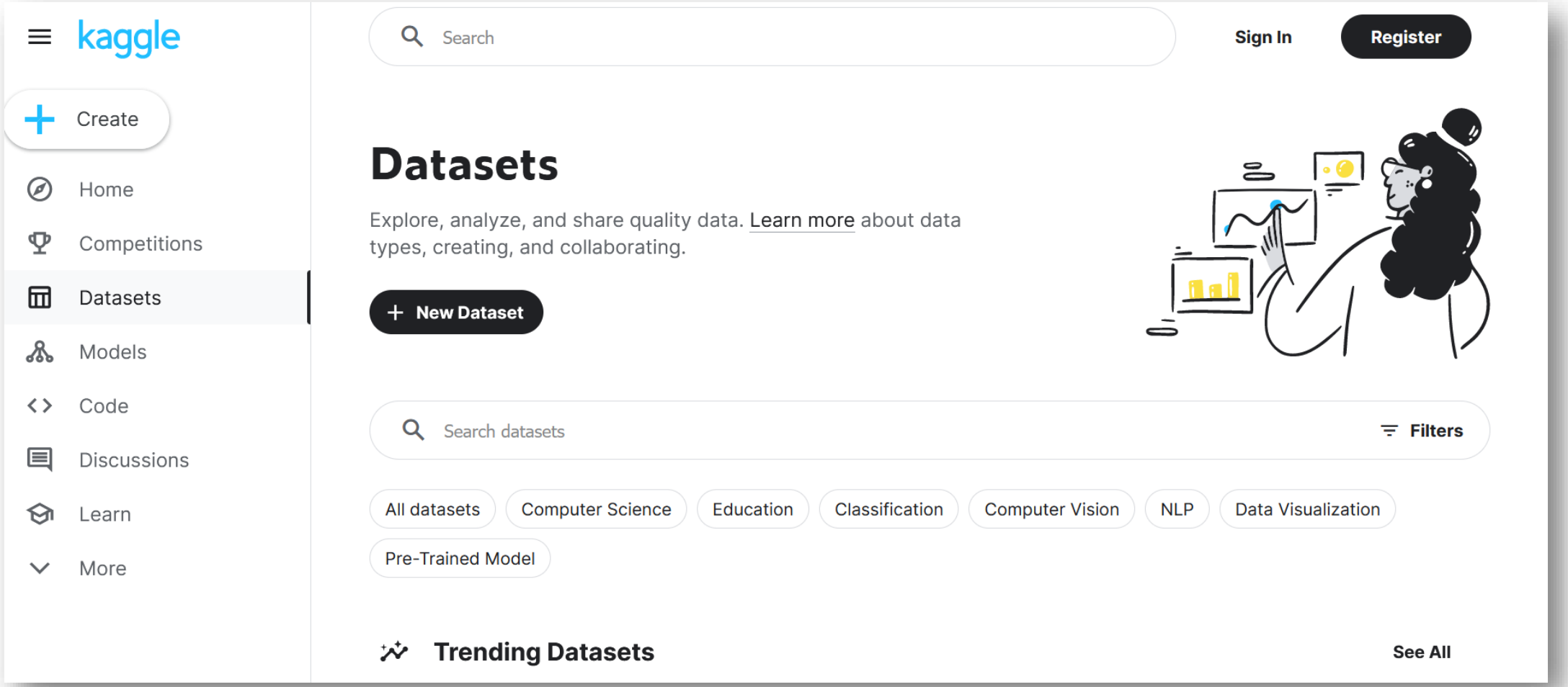
https://www.tensorflow.org/datasets?hl=zh-tw

其它公開資料集 – PyTorch



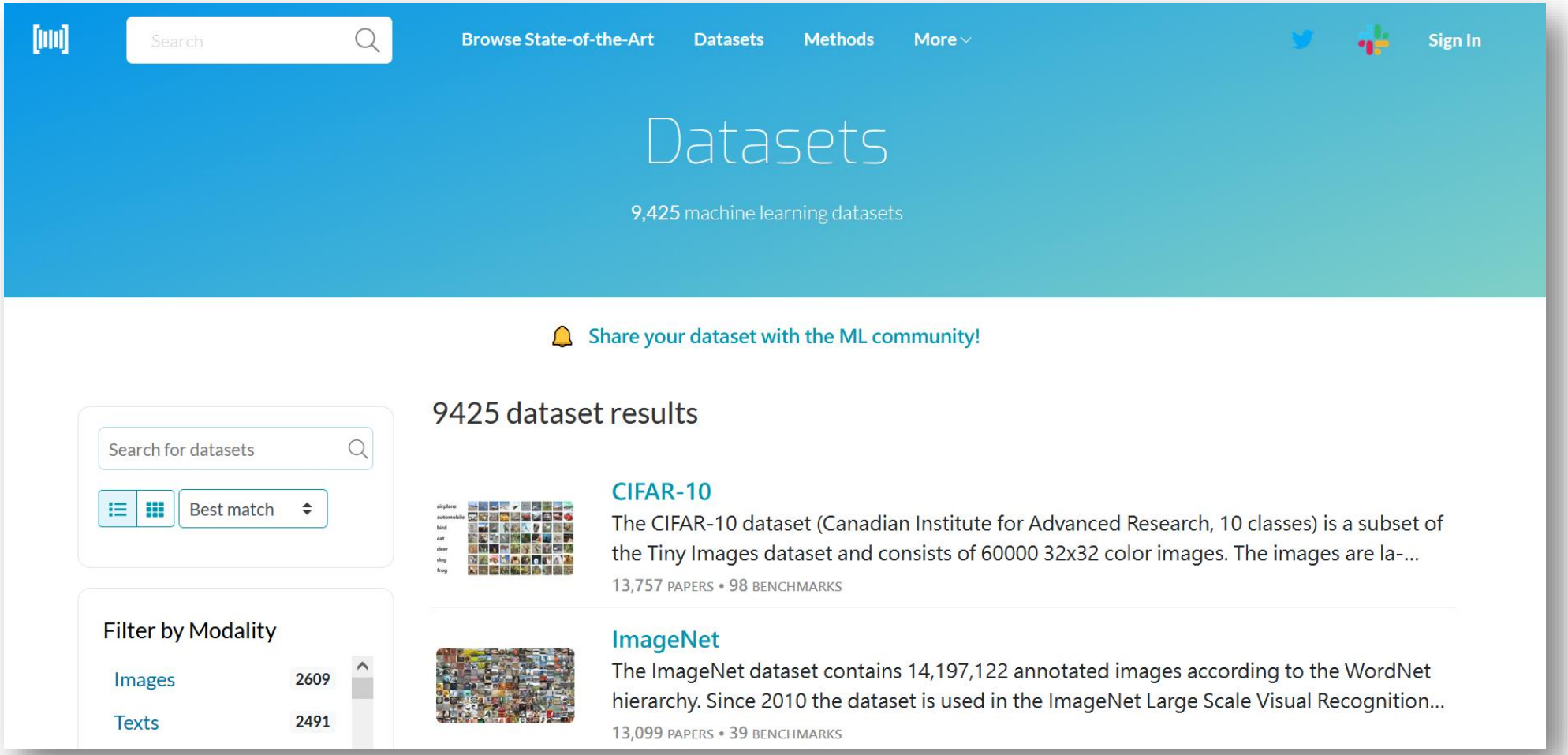
https://pytorch.org/vision/main/datasets.html

其它公開資料集 – Kaggle



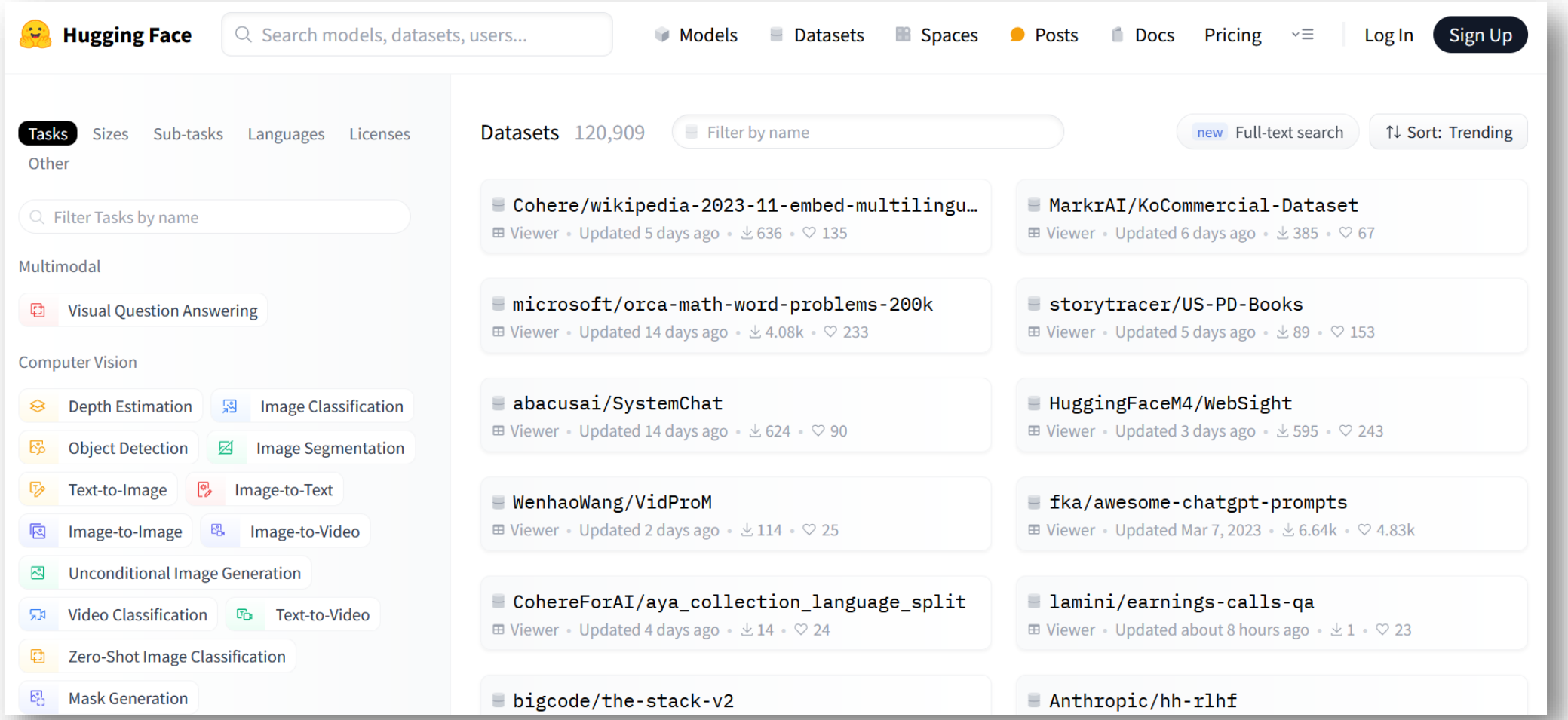
https://www.kaggle.com/datasets

其它公開資料集 – Papers with Code



https://paperswithcode.com/datasets

其它公開資料集 – Hugging Face



https://huggingface.co/datasets

其它公開資料集 – 台灣相關

國網中心資料集平台

https://scidm.nchc.org.tw/

政府資料開放平台

https://data.gov.tw/

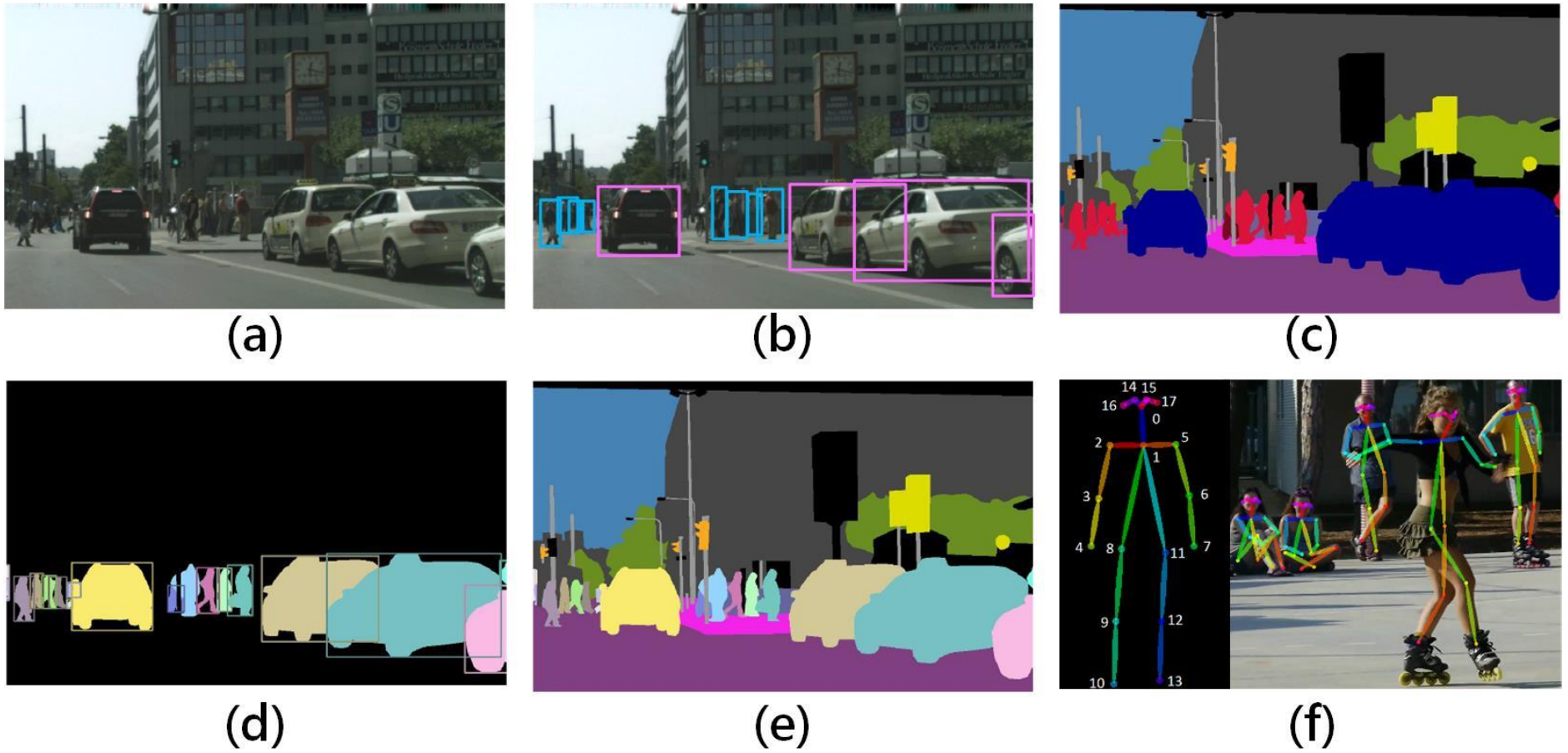
AI CUP 教育部全國大專校院人工智慧競賽

https://www.aicup.tw/

3.3.
資料集標註

- 影像標註類型
- 影像標註格式
- 影像標註工具

影像標註類型 – 依工作內容

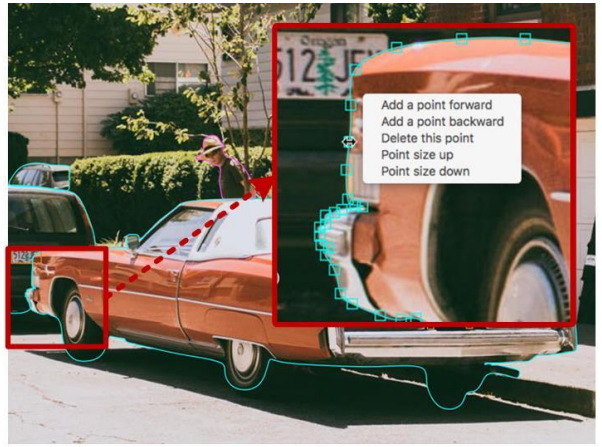


(a)原始影像/影像分類，(b)物件偵測，(c)語義分割，(d)實例分割，(e)全景分割，(f)人體姿態（關鍵點）。

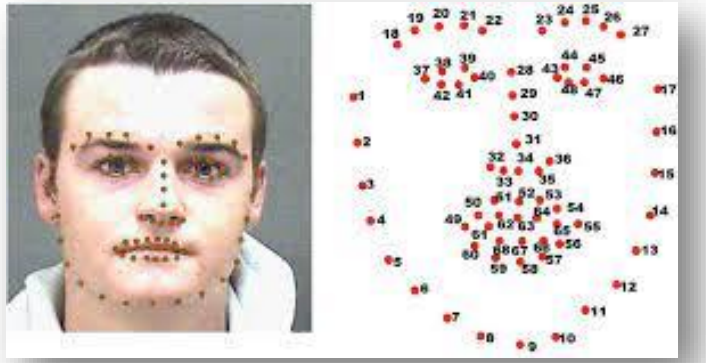
影像標註類型 – 依標註外形

(a)

(b)



(c)



(a) Face Landmark

(d)

(e)

(a)點/線，(b)矩形，(c)多邊形/貝茲曲線，(d)自由筆刷，(e)超像素。

影像標註類型 – 依時序內插

較適合直線移動物件，
可上下、左右移動。

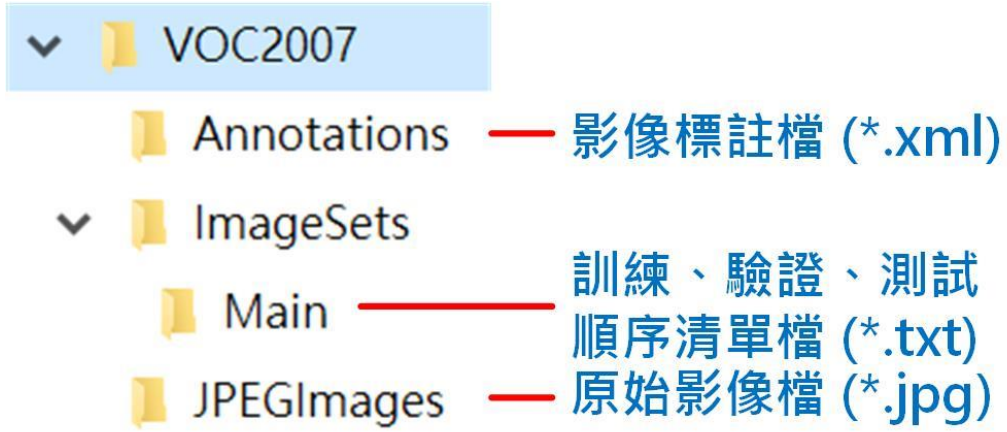
適用影片中移動體標註，
遇不等速移動需手動調整。

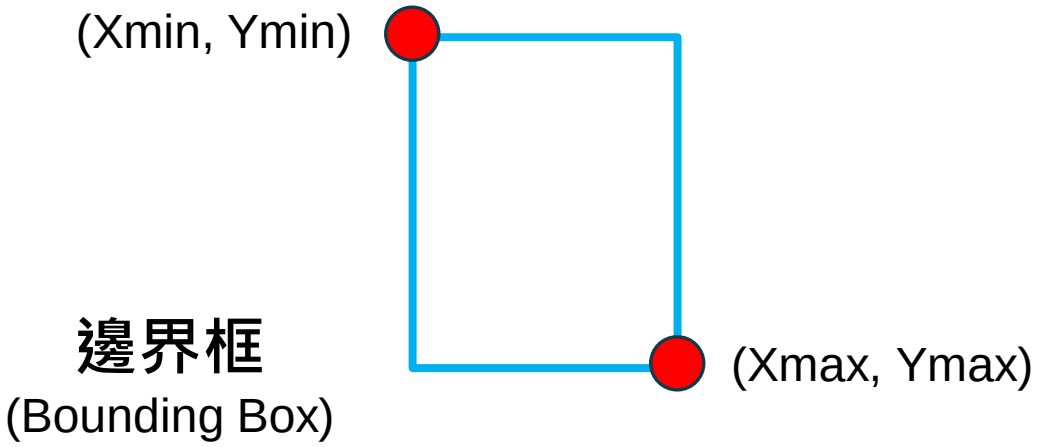
資料來源：https://youtu.be/sA_H3EGzO60

影像標註格式

- 影像/聲音/數據 單分類
  - 將所有資料標註結果集中在一個文字檔（如CSV）
  - 將不同分類資料置於不同檔案夾，以檔案夾名稱做為標註結果。
- 物件偵測（邊界框）
  - VOC (*.xml)
  - COCO (*.json)
  - YOLO (*.csv)
- 影像分割(語義/實例/全景)
  - 像素點標註
    - 輸出為和輸入尺寸相同之影像
    - 非破壞性未壓縮(已壓縮)影像
    - 256色調色盤影像
    - Run Length Encoder編碼文字檔
  - 封閉多邊形
- 姿態估測 / 人臉地標(Landmark)
  - 固定節點座標(2D x,y, 3D x,y,z)

影像標註格式 – Pascal VOC (xml)





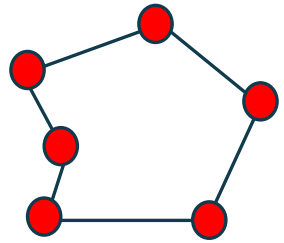

```
<annotation>
    <folder>JPEGImages</folder>
    <filename>img_1020.jpg</filename> 影像名稱
    <path>C:\VOC\JPEGImages\img_1020.jpg</path>
    <source>                                  影像路徑
        <database>Unknown</database>
    </source>
    <size>
        <width>640</width>
        <height>480</height>   影像尺寸
        <depth>3</depth>       色彩通道數
    </size>
    <segmented>0</segmented>
    <object>
        <name>tomato</name> 物件名稱
        <pose>Unspecified</pose>物件姿態
        <truncated>0</truncated> 影像是否被截切
        <difficult>0</difficult> 影像是否困難辨識
        <bndbox>
            <xmin>360</xmin>   物件框左
            <ymin>227</ymin>   上角座標
            <xmax>395</xmax>   物件框右
            <ymax>276</ymax>   下角座標
        </bndbox>
    </object>
    <object>
        ⋮   其它物件資訊
    </object>
</annotation>
```

第一個物件

影像標註格式 – MS COCO (json)

COCO 資料格式

```
{
   "info": info,
   "images": [image],
   "annotations": [annotation],
   "licenses": [license],
}

info{
   "year": int,
   "version": str,
   "description": str,
   "contributor": str,
   "url": str,
   "date_created": datetime,
}

image{
   "id": int,
   "width": int,
   "height": int,
   "file_name": str,
   "license": int,
   "flickr_url": str,
   "coco_url": str,
   "date_captured": datetime,
}

license{
   "id": int,
   "name": str,
   "url": str,
}

annotation{
   "id": int,
   "image_id": int,
   "category_id": int,
   "segmentation": RLE or [polygon],
   "area": float,
   "iscrowd": 0 or 1,
   "bbox": [x,y,width,height],
}

categories[{
   "id": int,
   "name": str,
   "supercategory": str,
}]
```



Segmentation : [polygon] 封閉多邊形

```
"segmentation":
[[510.66,423.01,...,510.45,423.01]],
"area": 702.1057499999998,
"iscrowd": 0,
```

Segmentation : [RLE] 持續長度編碼

```
"segmentation": {"counts":
[20736,2,453,5,452,9,447,13,444,...,5,34552],
 "size": [457,640]
},
"area": 3074,
"iscrowd": 1,
```

[白數量, 黑數量,白數量, …]
[3, 4, 3, …]

影像標註格式 – YOLO (txt)

VOC 物件偵測資料格式

```
<size>   影像尺寸
        <width>353</width>     影像寬度
        <height>500</height>   影像高度
        <depth>3</depth>       色彩深度(通道數)
</size>
<segmented>0</segmented>   是否分割
<object>   物件（目標）訊息
        <name>dog</name>                物件名稱（20分類）
        <pose>Left</pose>               拍攝角度（前後左右及未定）
        <truncated>1</truncated>        是否被遮擋或截斷
        <difficult>0</difficult>        是否難以檢測
        <bndbox>  物件（目標）外框
                <xmin>48</xmin>         物件左上角X座標
                <ymin>240</ymin>        物件左上角Y座標
                <xmax>195</xmax>        物件右下角X座標
                <ymax>371</ymax>        物件右下角Y座標
        </bndbox>
</object>
```

YOLO 物件偵測資料轉換

X = (xmin +(xmax-xmin)/2) * 1.0 / width

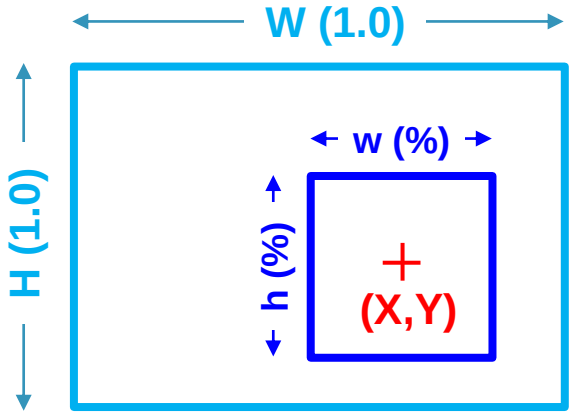
Y = (ymin +(ymax-ymin)/2) * 1.0 / height

W = (xmax-xmin) * 1.0 / width

H = (ymax-ymin) * 1.0 / height

邊界框位置
以左上表示

邊界框位置
以中心表示



YOLO 物件標註檔案格式(*.txt)

[分類編號ID] [物件中心X座標] [物件中心Y座標] [物件寬度佔比W] [物件高度佔比H]

0 0.344192634561 0.611000000000 0.416430594901 0.262000000000

OmniXRI Oct. 2020 整理繪製

影像標註格式 – 旋轉物件標註 (xml)

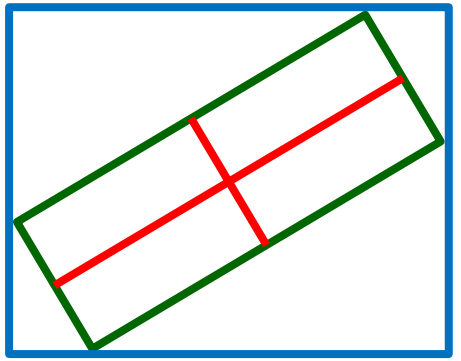
正常邊界框標註

旋轉邊界框標註

Oriented Object Detection (OOD)

```
<object>
    <name>car</name>
    <pose>Unspecified</pose>
    <truncated>0</truncated>
    <difficult>0</difficult>
    <robndbox>
        <cx>343.0944</cx>
        <cy>715.3419</cy>
        <w>123.0101</w>
        <h>59.9565</h>
        <angle>0.459784</angle>
    </robndbox>
    <extra/>
</object>
```

正常邊界框標註



旋轉邊界框標註

資料來源：https://blog.roboflow.com/yolov5-for-oriented-object-detection/

影像標註工具 – LabelImg

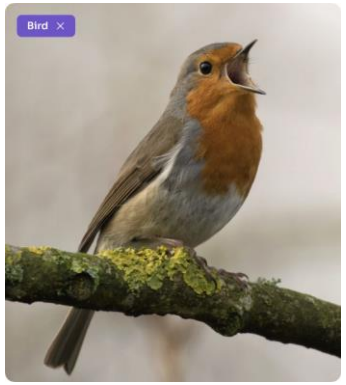
GITHUB https://github.com/tzutalin/labelImg
最簡單安裝方式在Python 3.x環境下使用pip安裝
pip3 install labelImg （建議在Python虛擬環境下安裝、執行）
啟動標註程式
labelImg

- 開源圖形化介面標註工具由Python和QT5所開發
- 支援多種平台安裝（Windows, Linux, Mac）
- 主要提供物件框位置、大小及標籤標註功能
- 提供PASCAL VOC(xml)、YOLO(txt)格式輸出

影像標註工具 – Intel (OpenCV) CVAT

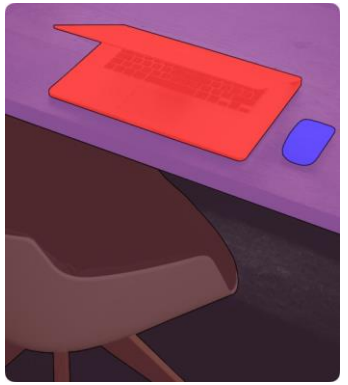
Intel Computer Vision Annotation Tool

- 支援多種標註格式輸出入(CVAT, VOC, YOLO, COCO, Tfrecord, MOT, LabelMe …)
- 可輸出 PNG格式影像多分類分割遮罩(mask)檔
- 可線上標註或以Docker於本地端安裝伺服器
- 支援影像分類、物件偵測、影像分割等任務。
- 標註方式：點、折線、方框、多邊形。
- 支援影片自動標註（補間追蹤特定物件）功能，節省標註時間。
- 僅適用Google Chrome瀏覽器。
- Github: https://github.com/opencv/cvat
- 線上版: https://www.cvat.ai/

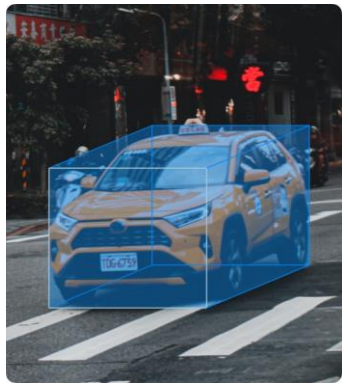


影像分類


物件偵測


影像分割


影片(內插)標註

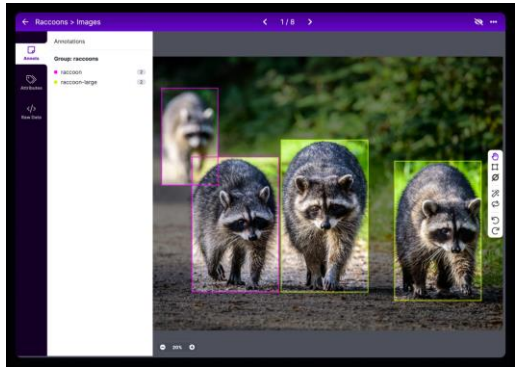

立體盒


骨架

影像標註工具 – Roboflow

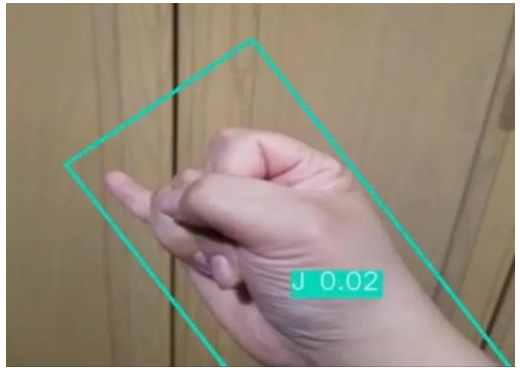
線上標註工具https://roboflow.com/

- 提供多樣標註方式。
- 支援多種儲存格式。
- 可協助資料擴增完成資料集。
- 線上工作可支援多人協作。
- 提供常用模型及小型測試資料集。



物件邊界框

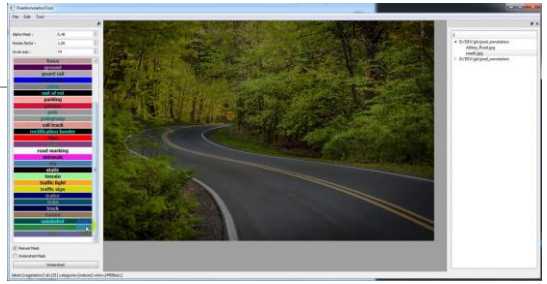
封閉多邊形



旋轉邊界框

SAM 自動分割生成外框

影像標註工具 – 其它



【影像分類/物件偵測類】

- RectLabel https://rectlabel.com（商業付費版）
- Labelme https://github.com/wkentaro/labelme
- OpenCV SuperAnnotate Desktop https://opencv.org/superannotate-desktop（部份免費）
- labelbox https://github.com/Labelbox/Labelbox
- Microsoft VoTT https://github.com/microsoft/VoTT
- VGG Image Annotator (VIA) http://www.robots.ox.ac.uk/~vgg/software/via

【影像分割類】

- PixelAnnotationTool https://github.com/abreheret/PixelAnnotationTool
- superpixels-segmentation https://github.com/Labelbox/superpixels-segmentation
- semantic-segmentation-editor https://github.com/Hitachi-Automotive-And-Industry-Lab/semantic-segmentation-editor
- Segment Anything Model

https://segment-anything.com/

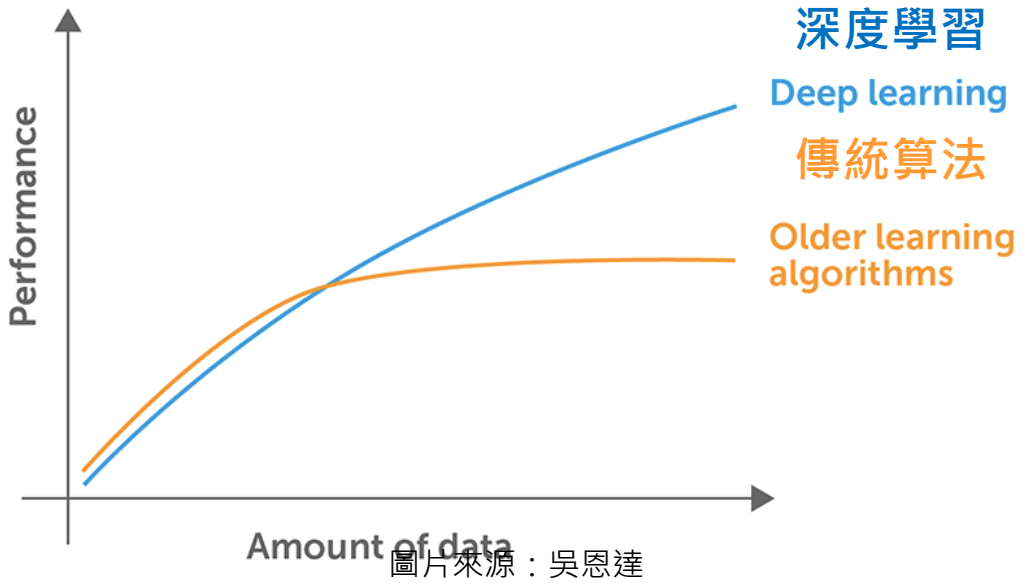
3.4.
資料集迷思

- 資料增長
- 標註水準
- 子集不均
- 自動聚類

資料增長

* AI 會自我學習只要一直提供資料就會變更厲害？

資料增長 vs. 訓練成果



圖片來源：吳恩達

擴增資料集、反覆訓練

- 監督式（分類）
- 非監督式（聚類）
- 遷移式（監督型加速式）

免資料集、明確獎懲規則

- 增強式（遊戲、閉路控制）

少量資料集、反覆訓練

- 生成對抗式（GAN）

不想學習圖書館再多書也沒用

標註水準

* 數據越多學習訓練效果越好？

監督式

- 標註品質
  - 標註工具（便捷、管理）
  - 普通人 vs. 專家
- 資料多樣性
  - 資料重新採集
  - 資料衍生擴增
  - 收集公開資料集
  - 對抗生成
- 子類平均度

高品質標註資料越多越好

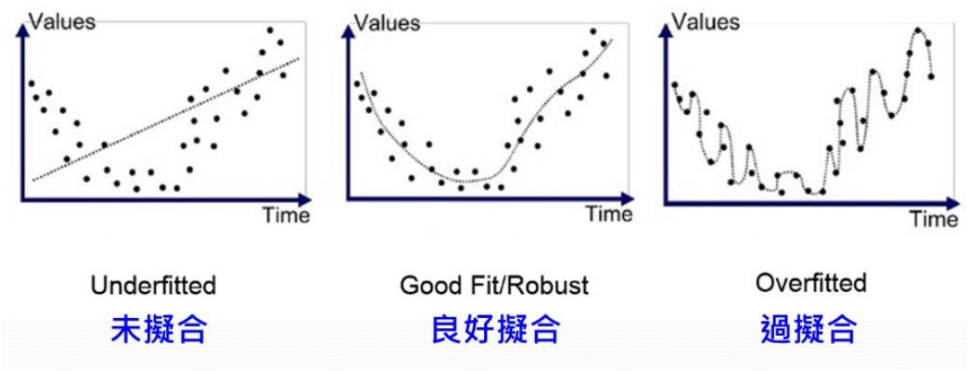
非監督式

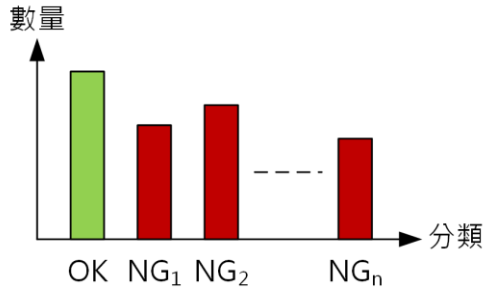
- 增強式─獎懲規則驅動免資料集（如遊戲）
- 關連式無標註型資料集越多越好（如語音、翻譯）
- 高維資料聚類
  - 格式化資料及樣態種類
  - 資料降維、特徵提取
  - 計算複雜度（運算速度）
  - 記憶體容量

資料在精不在多

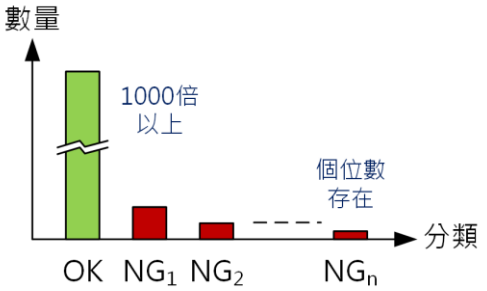
子集不均

* 只要大量良品就能自動學習特徵做為不良品檢測？





理想樣態分佈



真實樣態分佈

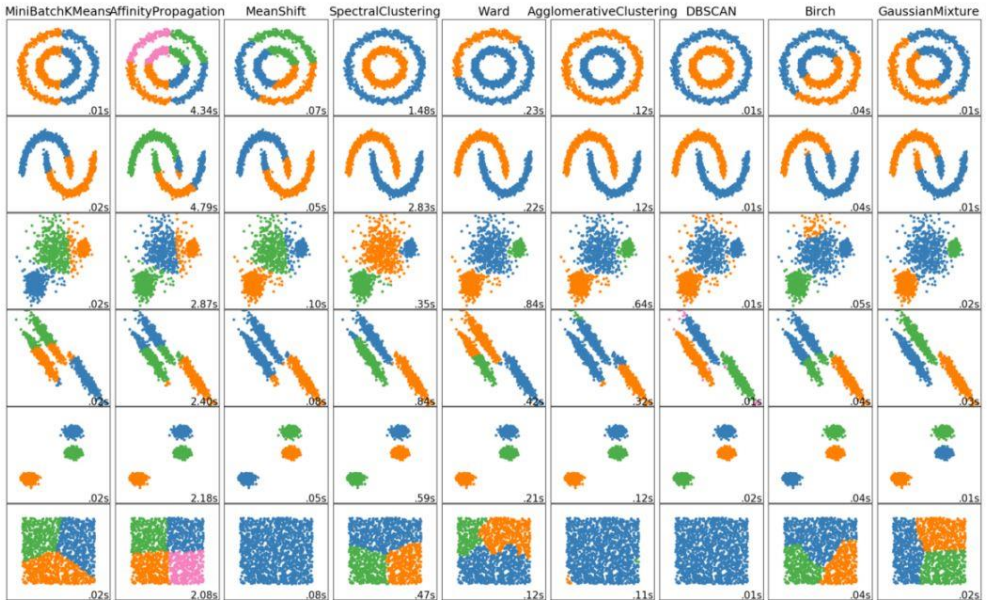
- 人類採特徵法學習只須少量樣本，深度學習採暴力法（填鴨死記）須巨量樣本。
- 異常偵測法(Anomaly Detection)採資料統計法找出合理分佈區間，再挑出離群樣本，適用資料維度較小的的案例。

自動提取特徵，工程耗大

自動聚類

* 提供巨量未標註資料就能自動找出分類規則關連？

常見聚類算法及分類結果



（二維資料分類）

圖片來源：https://www.plob.org/article/12370.html

翻譯問題 (LLM已攻克？)

- 表面上為非監督式學習，但實質上訓練素材（語料）為人類產生之標註資料。

影像（語音）分類問題

- 高維（超稀疏）資料自動聚類困難

半監督式

- 部份有標註，部份自動產生標註，再由人決定標註正確性，加快標註時效。

困難重重，有待突破

小結

資料集建置

- 依據需求使用公開資料集或自行收集相關資料集，並保持多樣性及進行初步整理清洗，不足處再加以擴增。

公開資料集

- 可從多種管道中取得不同型態已標註之資料集，電腦視覺、自然語言、格式數據皆有對應資料集可供參考。

資料集標註

- 了解資料標註方式，資料存放格式及基本工具操作。

資料集迷思

- 資料在精不在多，標註品質決定訓練成果，子集不均需調整，自動聚類仍不成熟。

參考文獻

- 許哲豪，臺灣科技大學資訊工程系「人工智慧與邊緣運算實務」（2021~2023）

https://omnixri.blogspot.com/p/ntust-edge-ai.html

- 許哲豪，【AI HUB專欄】如何建立精準標註的電腦視覺資料集

https://omnixri.blogspot.com/2020/10/ai-hub_16.html

- 許哲豪，【課程簡報分享】AI萬能？導入AI的八大迷思剖析

https://omnixri.blogspot.com/2019/08/aiai.html

- 許哲豪，【vMaker Edge AI專欄 #10】訓練AI模型資料不足怎麼辦？聊聊資料集擴增手法

https://omnixri.blogspot.com/2023/10/vmaker-edge-ai-10-ai.html

延伸閱讀

- Intel, CVAT Computer Vision Annotation Tool | OpenVINO™ toolkit | Ep. 51 | Intel Software (Youtube)

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

- Intel, CVAT Auto Annotation | OpenVINO™ toolkit | Ep. 52 | Intel Software (Youtube)

https://www.youtube.com/watch?v=jbqOa8DX7Jg

- MakerPro, 【Roboflow標記工具】新手也能輕鬆上手，打造專業AI資料集的線上指南！

https://makerpro.cc/2024/01/roboflow-marking-tool-is-easy-to-use/

歐尼克斯實境互動工作室
(OmniXRI Studio)
許哲豪 (Jack Hsu)

Facebook : Jack Omnixri
FB社團： Edge AI Taiwan邊緣智能交流區
電子信箱：omnixri@gmail.com
部落格：https://omnixri.blogspot.tw
開　源：https://github.com/OmniXRI

YOUTUBE 直播： https://www.youtube.com/@omnixri1784/streams