African Institute for Mathematical Sciences — Cameroon
Structured Master's, Data Science Stream · Computer Vision

Seeing Through Words

Week 3 · Day 3: CLIP, Vision–Language Models and Applications Across Domains

Yusuf Brima, Ph.D.

Introduction to Computer Vision · AIMS Cameroon, 2026–2027

Today

1. **Why connect images and language**
2. **CLIP**: learning a shared space from 400 million captions
3. **Zero-shot classification**, tested on Ugandan bean leaves
4. **Search, prompts and the limits of CLIP**
5. **Vision–language models**: from CLIP to multimodal assistants
6. **Applications across domains**: health, agriculture, earth observation, industry, robotics, conservation

By the end of today you should be able to

- explain contrastive training and compute its loss for a small batch;
- build a zero-shot classifier from text prompts;
- compare zero-shot, few-shot and fine-tuned approaches honestly;
- recognise CLIP's failure modes;
- map a real application to the right task, data and model.

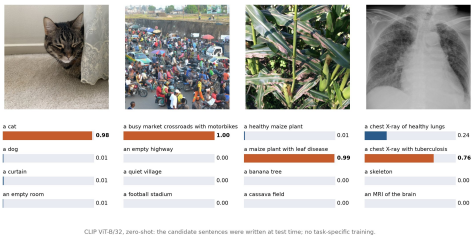
Where we are

So far

- Classifiers that know a fixed list of labels (Weeks 1–2).
- Captioning and text-guided detection, briefly (Week 2, Day 5).
- Diffusion models steered by text through a text encoder (yesterday).

Today

- The model behind that text encoder: **CLIP**.
- Then a tour of how everything in this course is used across domains.



CLIP on Week 1, Day 1: maize disease recognised from a sentence, with no training on maize.

Part 1

Why connect images and language?

The limits of a fixed label list

- An ImageNet classifier knows 1000 labels, chosen in 2009. “Moto-taxi”, “cassava mosaic” and “pothole” are not among them.
- Every new label means new labelled images and retraining.
- Labels are coarse: “dog” says nothing about “a muddy dog shaking itself by a river”.

Intuition

Language is a far richer label: open-ended, compositional (“a *yellow taxi in the rain*”), and already attached to billions of images on the web as captions and alt-text.

The idea

Learn image features and text features in **one shared space**, so that an image and a sentence describing it end up close together.

Vision and language: a short history

Year	Milestone
2014–15	Image captioning with a CNN encoder and a recurrent decoder (“Show and Tell”)
2015	Visual question answering (VQA) benchmark: answer questions about images
2019	Transformers that read image regions and words together (ViLBERT, LXMERT)
2021	**CLIP** and ALIGN: contrastive training on hundreds of millions of web pairs; zero-shot transfer
2022	Text-to-image diffusion (DALL · E 2, Stable Diffusion) built on CLIP-like text encoders
2023	Multimodal assistants: an LLM that can see (GPT-4V, LLaVA); SAM for promptable segmentation
2024–26	Open multimodal models, video understanding, vision–language–action models for robots

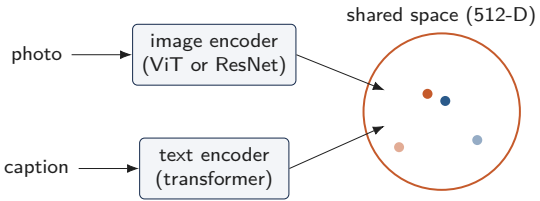
Intuition

The turning point was scale and supervision: instead of carefully labelled datasets, noisy captions from the whole web.

Part 2

CLIP: contrastive language–image pre-training

CLIP (2021): two encoders, one space

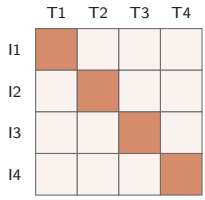


- Trained on **400 million** (image, caption) pairs collected from the web.
- Each encoder maps its input to a 512-number vector, normalised to length 1.
- Matching image and caption: vectors close together (high cosine similarity).

Intuition

CLIP never sees a class label. It learns what images are *about* by learning which sentences people wrote next to them.

Contrastive training: find your partner in the batch



similarities of 4 images with 4 captions

- Take a batch of N image–caption pairs (CLIP used $N = 32\,768$).
- Compute all $N \times N$ similarities.
- Loss: each image should pick its own caption among the N (a softmax over its row), and each caption its own image (a softmax over its column).
- The diagonal goes up; everything else goes down.

Intuition

A matching game: shuffle 32 768 photos and their captions, and ask the model to reunite them. To win, it must understand both photos and sentences in detail.

The contrastive loss, worked

InfoNCE loss for image i

$$\mathcal{L}_i = -\log \frac{\exp(s_{ii}/\tau)}{\sum_j \exp(s_{ij}/\tau)}$$

s_{ij}: cosine similarity of image i and caption j; τ: a temperature (learned; about 0.01).

One image, three captions, $\tau = 0.1$

Similarities: own caption 0.30, others 0.20 and 0.10.
Scaled: $3, 2, 1$; softmax $= (0.665, 0.245, 0.090)$.
$\mathcal{L} = -\log 0.665 = 0.41$.

- It is Week 2's cross-entropy: the "classes" are the captions in the batch.
- The small temperature makes small differences in similarity count a lot.
- Bigger batches mean more wrong captions to beat: a harder, more informative game.

Practitioner's Note

Contrastive learning needs large batches and huge data: CLIP is expensive to train, cheap to use.

The full similarity matrix, worked

A batch of three, $\tau = 0.1$

	T1 goat	**T2** taxi	**T3** leaf
I1 goat	**0.32**	0.12	0.15
I2 taxi	0.10	**0.29**	0.08
I3 leaf	0.14	0.09	**0.18**

Row I3 scaled: $1.4, 0.9, 1.8$; softmax for the correct T3: $e^{1.8}/(e^{1.4} + e^{0.9} + e^{1.8}) = 6.05/12.56 = 0.48$; loss 0.73.

- Rows: each image chooses among captions. Columns: each caption chooses among images. CLIP averages the two losses.
- I3 has the weakest diagonal (0.18): the leaf is not yet well separated from the goat caption. Its gradient is largest; training works hardest on it.

Intuition

The loss does not ask "is this caption right?" in isolation, only "is it *more* right than the others in the batch?".

What lies in CLIP's space

- Images of the same thing, from different viewpoints, lighting or styles (photo, sketch, painting), land near each other and near the words that describe them.
- Texts with similar meaning land near each other.
- Distance in the space is a learned notion of *semantic* similarity.

Intuition

Week 1's dream was invariance: the same object despite changes in pose, light and background. CLIP learns a large part of that invariance from data, because the caption "a cat" stays the same while the pixels change wildly.

Why web captions are such a strong teacher

- **Scale**: 400 million pairs, against 1.3 million ImageNet images.
- **Variety**: photos, drawings, screenshots, diagrams, memes, from all over the web.
- **Rich labels**: captions mention objects, attributes, actions, styles, places and text in the image.
- **Free**: nobody had to label anything for CLIP.

Practitioner's Note

The same properties are the weaknesses: captions are noisy, sometimes wrong, biased towards what web users photograph and how they describe it, and mostly in English.

Intuition

Breadth over precision. CLIP knows a little about almost everything, and a lot about almost nothing in particular.

Part 3

Zero-shot classification

Building a classifier from words

1. Write one sentence per class: “a photo of a healthy bean leaf”, “a photo of a bean leaf with rust”, . . .
2. Encode each sentence with the text encoder: K vectors.
3. Encode the image.
4. Predict the class whose sentence is most similar to the image.

Zero-shot prediction

$$\hat{y} = \arg\max_{k} \cos\left(f_{\text{img}}(x),\ f_{\text{text}}(\text{prompt}_k)\right)$$

Intuition

The text vectors play the role of the last layer’s weights in Week 2’s classifier, but they are *written*, not trained. Changing the classes means changing sentences, not retraining.

CLIP on ImageNet: the result that surprised everyone

- Zero-shot CLIP (largest model) reached about 76% top-1 on ImageNet, matching a ResNet-50 trained on ImageNet's 1.28 million labelled images, without seeing any of them.
- On ImageNet variants (sketches, renditions, adversarially chosen photos), it held up far better than ImageNet-trained models.

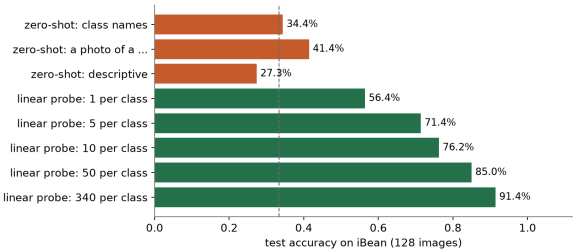
Intuition

An ImageNet model learns "what ImageNet's photographers photographed". CLIP learns from many more kinds of image, so a sketch of a banana is not a surprise to it.

Practitioner's Note

The headline numbers are for everyday categories. On specialised data (satellite, medical, our bean leaves) zero-shot CLIP is much weaker; always measure on your own task.

A real test: Ugandan bean leaves (Week 2, Day 2)



CLIP ViT-B/32, the 128 test leaves; guessing gives 33%.

- **Zero-shot**, class names only: 34.4%.
- "A photo of . . . ": 41.4%.
- Descriptive prompts (spots, pustules, healthy green): 27.3%, *worse than guessing*.
- **Linear probe** on CLIP features: 56.4% with 1 image per class, 76.2% with 10, 91.4% with all.

Intuition

Zero-shot barely beats guessing here: bean diseases are far from CLIP's web captions, and our careful descriptions made it worse. Yet one labelled leaf per class already beats every prompt: the image features are good; the text side does not know these diseases.

The linear probe, explained

- Encode every training image once with CLIP's image encoder: 512 numbers each.
- Train a logistic-regression classifier on these vectors (Week 2's "frozen features").
- Seconds on a laptop; no gradient through CLIP.

Our numbers

Labelled leaves per class → test accuracy: 1: 56.4%; 5: 71.4%; 10: 76.2%; 50: 85.0%; all (about 340): 91.4%.

Intuition

The probe asks a different question from zero-shot: not "which sentence matches?" but "where do the classes sit in CLIP's space?". CLIP's features already separate the leaves well; its *text* side just does not know the right names for them.

Practitioner's Note

With one image per class, results vary a lot with which image you happen to pick. Report the average over several random draws.

Prompt engineering

- CLIP was trained on captions, not single words: "a photo of a {class}" usually beats "{class}".
- Describing *what the class looks like* sometimes helps when the class name is rare on the web, but not always: on our beans, descriptive prompts scored 27.3%, the worst of the three.
- **Prompt ensembles**: average the text vectors of several phrasings ("a photo of", "a close-up of", "a field photo of").

Practitioner's Note

Choose prompts on a *validation* set, never on the test set: with enough phrasings you can always find one that looks good on 128 test images by chance.

Intuition

Prompt engineering is feature engineering in words. It helps, but it is fragile: small wording changes can swing accuracy by several points.

Combining text and a few labels

- Use the zero-shot text vectors as the *starting* weights of the linear classifier, then train on a few labelled images.
- Or average: class prototype = mix of the text vector and the mean image vector of the labelled examples.
- **Prompt learning** (CoOp, 2022): learn the prompt's word vectors themselves from a few labelled images.

Intuition

Text gives a prior ("rust looks like orange pustules"); images refine it ("this is what rust looks like *in these photos*"). With very few labels, the prior helps; with many, the images dominate.

Zero-shot, few-shot or fine-tuned?

	Zero-shot CLIP	Linear probe on CLIP	Fine-tuned network
Labels needed	none	a few per class	hundreds or more
Compute	encode once	seconds	minutes to hours
Beans result	41.4% (best prompt)	76.2% (10 per class)	95% (Week 2)
Best for	exploring, common concepts	most small projects	deployment, unusual domains

Practitioner's Note

For a capstone: start with zero-shot as a free baseline, add a linear probe with whatever labels you have, and fine-tune only if the probe falls short. Report all three.

Zero-shot classification in code

```
from transformers import CLIPModel, CLIPProcessor
name = "openai/clip-vit-base-patch32"
model = CLIPModel.from_pretrained(name).eval()
proc  = CLIPProcessor.from_pretrained(name)

prompts = ["a photo of a healthy bean leaf",
           "a photo of a bean leaf with rust
    pustules",
           "a photo of a bean leaf with angular
    brown spots"]
inputs = proc(text=prompts, images=image,
              return_tensors="pt", padding=True)
with torch.no_grad():
    probs = model(**inputs).logits_per_image.
     softmax(-1)
```

- About 600 MB to download once; runs on a laptop CPU in well under a second per image.
- `logits_per_image` are the scaled cosine similarities; softmax turns them into probabilities over *your* prompts.
- The probabilities only compare the prompts you gave: an image of a goat still gets 100% spread over three leaf diseases.

Open-vocabulary detection and segmentation

- Replace a detector's fixed class head with CLIP-like text vectors: each box is scored against any phrase (OWL-ViT, Grounding DINO).
- Combine with SAM (Week 2, Day 4): text $\rightarrow$ boxes $\rightarrow$ masks.
- Lets you search for "a goat", "a moto-taxi with three passengers" or "a plastic bag" without training a new detector.

Practitioner's Note

Week 2, Day 5 showed OWL-ViT finding some queries and missing others. Open-vocabulary detectors are excellent for *bootstrapping labels*: run them, correct their boxes by hand, then train a fast specialist detector on the corrected labels.

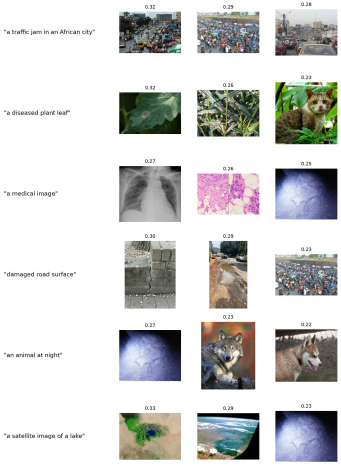
Intuition

Language turns detection into search: the class list becomes whatever you can describe.

Part 4

Search and the limits of CLIP

Searching images with words



Twenty photographs from this course, encoded once. Each query sentence is encoded and the three most similar photos are returned (similarity above each).

- No labels, no training: just distances in CLIP's space.
- Works for queries nobody planned for, such as "an animal at night".
- Fails honestly too: the cat among leaves comes third for "a diseased plant leaf"; the night-time fox for "a medical image". With only 20 photos, the best few may still be poor matches.

Practitioner's Note

Text-to-image search powers photo libraries, e-commerce and archive search. For a large collection, store the image vectors in a vector index and search millions in milliseconds.

Image-to-image search and deduplication

- Query with an *image* instead of text: find the most similar photos in a collection.
- **Deduplication**: near-identical images (same scene, re-saved, cropped) have very similar vectors; remove them before splitting train and test.
- **Data cleaning**: images far from all others are often mislabelled or broken.

Practitioner's Note

Duplicates across train and test inflate accuracy (Week 2's leakage). A five-minute CLIP deduplication pass on a capstone dataset is cheap insurance: compute all pairwise similarities and inspect pairs above about 0.95.

Cost

10 000 images $\rightarrow$ 50 million pairs, but as one $10\,000 \times 512$ matrix product: about a second.

When CLIP is fooled: a typographic attack



RyanDannyChacko, CC BY-SA 4.0

A hand-written label reading “DOG” placed on the photo of the cat.

- Original: “cat” with probability 1.00.
- With the label: “dog” 0.55, “cat” 0.45.

Intuition

CLIP learned that words in an image often say what the image is about (signs, product labels, memes). It reads the text and partly believes it. A clear strength for reading signs becomes a weakness here.

Other known limits

- **Counting**: "three people" vs. "five people" are barely distinguished.
- **Spatial relations**: "a cup on a book" vs. "a book on a cup".
- **Fine-grained and specialist classes**: crop diseases, parasites, X-ray findings.
- **Negation**: "no people" is close to "people".
- **Bias**: associations from web text, including harmful stereotypes about people.

Intuition

CLIP's text encoder sees a sentence roughly as a bag of meaningful words. Word order, numbers and "not" carry little weight, because predicting which caption goes with which image rarely depends on them.

Practitioner's Note

Never use CLIP's similarity to classify *people* into social categories. Its associations reflect web text, not facts about individuals.

CLIP's similarity scores are not probabilities

- Raw cosine similarities for matching pairs are typically only about 0.25–0.35; non-matching about 0.1–0.2.
- The softmax probability depends entirely on which prompts you offered: add a fourth class and every probability changes.
- A "98%" zero-shot prediction can be confidently wrong if the right answer was not among the prompts.

Intuition

CLIP answers "which of *these* fits best?", never "is this one right?". Always include an "other" or "none of these" option when the real world might contain something else.

Practitioner's Note

Calibrate a threshold on a validation set if you need to reject images (Day 4 discusses calibration).

Languages and places

- CLIP's captions are overwhelmingly English; French works partly; most African languages, such as Fulfulde, Ewondo or Duala, barely at all.
- Scenes, crops, foods and clothing common in Cameroon are less frequent in its data, so its similarities are less reliable for them.
- Multilingual CLIP variants exist, trained or distilled with many more languages.

Intuition

A model's "world" is its training data. Zero-shot works well for concepts the web talks about a lot, in the languages the web uses most.

Practitioner's Note

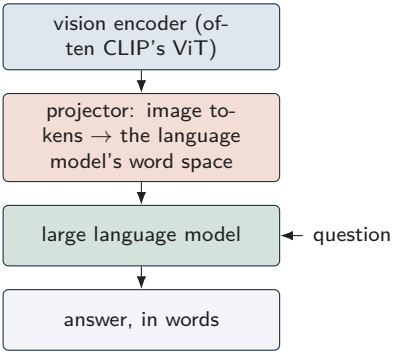
Building local evaluation sets (images from Cameroon, prompts in local languages) is valuable research in its own right, and often publishable.

Part 5

Vision–language models

From matching to generating text

- **CLIP** scores image–text pairs: it can *choose* a description, not *write* one.
- **Captioners** (BLIP, Week 2, Day 5) generate a sentence from an image.
- **Multimodal assistants** (2023 onwards) answer questions, follow instructions and hold conversations about images.



Intuition

The image becomes a few hundred "visual words" that the language model reads alongside the question, exactly as it reads text tokens.

How multimodal assistants are trained

1. **Start** from a pre-trained vision encoder (CLIP) and a pre-trained language model.
2. **Align**: train the small projector on image–caption pairs, so image tokens "speak the language" of the LLM.
3. **Instruction-tune**: train on (image, question, good answer) examples, so the model follows requests.

Intuition

Transfer learning again (Week 2, Day 2), at the largest scale: two huge pre-trained models, glued with a small learned connector and taught to cooperate.

Practitioner's Note

These models can describe, explain and reason about images in many languages. They also hallucinate fluently (Week 2, Day 5). Test them on your images and questions before relying on any answer.

What vision–language models are good at, and not

Often good

- Describing everyday scenes and objects.
- Reading printed text (signs, labels, documents).
- Answering questions that need general knowledge.
- Explaining charts and diagrams.

Often weak

- Precise counting and measurement.
- Small details and small objects.
- Specialist images: microscopy, X-rays, satellite bands.
- Saying "I don't know" when unsure.

Practitioner's Note

A detector that counts parasites (Week 2, Day 3) and a segmenter that measures lesion area (Week 2, Day 4) are often more trustworthy than asking a general assistant "how many?" or "how much?". Combine them: specialist models measure, the assistant explains.

Testing a vision–language model honestly

1. Collect 50–100 images from your setting, with known answers.
2. Ask the same questions in the same wording; also rephrase.
3. Score answers against the truth; note hallucinations separately.
4. Ask about things *not* in the image ("what colour is the dog?" when there is no dog).

Intuition

Fluency is not accuracy. A model that answers every question confidently, including impossible ones, is telling you about its training, not about your image.

Practitioner's Note

Report the rate of "confident wrong answers", not just accuracy. In medicine and agriculture, a wrong answer stated confidently is worse than "I am not sure".

Part 6

Applications across domains

A framework for any application

1. **Decision**: what will someone do with the output?
2. **Task**: classify, detect, segment, measure, describe, generate?
3. **Data**: where from, how labelled, how will the test set differ from training?
4. **Model**: zero-shot $\rightarrow$ frozen features $\rightarrow$ fine-tuning; classical baseline.
5. **Deployment**: device, connectivity, speed, who checks errors?

Intuition

The next slides walk through domains with this lens. The models change little between domains; the decisions, data and risks change a lot.

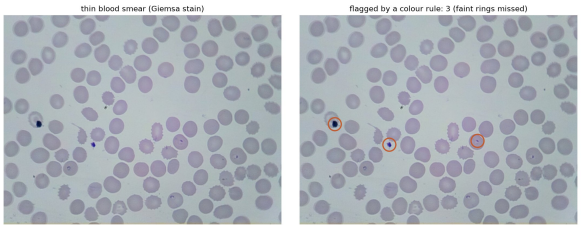
What changes between domains

Aspect	Examples
Image type	RGB phone photo; grey X-ray; 13-band satellite tile; stained microscopy; thermal
Image size	224^2 crops vs. $10\,000^2$ satellite scenes or gigapixel pathology slides
Labels	cheap (crowd) vs. expert-only (radiologist, pathologist, agronomist)
Cost of errors	a missed defect vs. a missed tumour vs. a wrong species count
Shift	new hospital, new season, new camera, new country
Rules	medical-device regulation, data protection, export of data

Intuition

Before choosing a model, fill in this table for your project. It tells you which week of this course matters most for you.

Biomedicine: microscopy



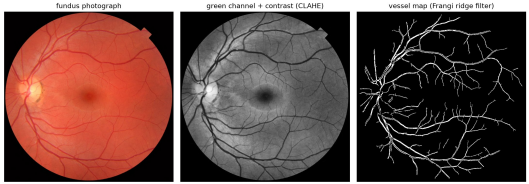
OpenFlexure Project, CC BY-SA 4.0

- **Decision**: treat for malaria? estimate parasite density.
- **Task**: detection and counting (Week 2, Day 3); classical colour rules as a baseline (Week 1).
- **Data**: thousands of annotated cells exist (NIH); local smears differ in stain and microscope.
- **Deployment**: cheap microscopes plus phones (OpenFlexure); offline.

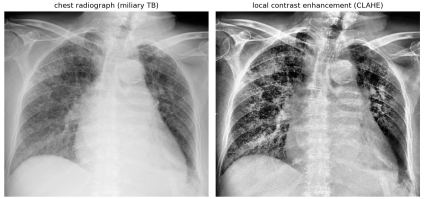
Practitioner's Note

Sensitivity matters most: a missed infection is worse than a false alarm that a technician re-checks. Choose thresholds accordingly.

Biomedicine: radiology and ophthalmology



scikit-image sample image (M. Häggström, CC0)



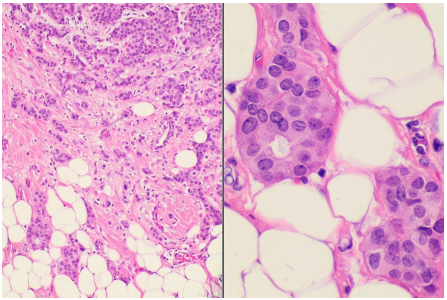
B. Herreros et al., CC BY 4.0

- **Retina**: screening for diabetic retinopathy from fundus photos; vessel segmentation (Week 1 classical; U-Net, Week 2).
- **Chest X-ray**: TB and pneumonia triage where radiologists are scarce.
- **Key risk**: models trained in one hospital fail in another (capstone topic P4).

Practitioner's Note

Medical AI is regulated as a medical device in many countries. Research prototypes must not be presented as diagnostic tools.

Biomedicine: pathology



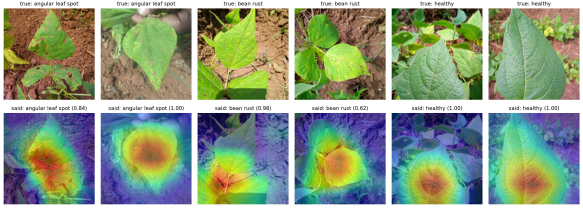
Mikael Häggström, CC0

- **Decision**: grade a tumour; find metastases in lymph nodes.
- **Task**: classify and segment tissue in whole-slide images of billions of pixels, processed as thousands of tiles.
- **Data**: labels from pathologists are slow and costly; weak labels (one diagnosis per slide) are common.
- **Shift**: staining varies between laboratories: colour normalisation (Week 1, Day 2) and augmentation help.

Practitioner's Note

Pathology foundation models, pre-trained on millions of tiles, are now openly available; a linear probe on their features is a strong baseline.

Agriculture



- **Decision**: treat, remove, or monitor a plant; estimate yield.
- **Tasks**: disease classification (Week 2, Day 2), severity segmentation (Week 2, Day 4), fruit counting, weed detection.
- **Data**: iBean and iCassava from Makerere (Uganda); field photos vary with season, light and phone.
- **Deployment**: offline phone apps; extension officers in the loop.

Intuition

The hard part is rarely the network (95% on iBean in minutes); it is collecting test photos from other farms and seasons, and making the output useful to a farmer.

Agriculture: what makes field images hard



Chroanch, CC BY-SA 4.0



Dwight Sipler from Stow, MA, USA, CC BY 2.0

- Clutter: other leaves, soil, hands, shadows.
- Symptoms of different diseases (and of nutrient deficiency or drought) can look alike.
- Several problems on one leaf at once.
- Datasets photographed on plain backgrounds (PlantVillage) transfer poorly to the field.

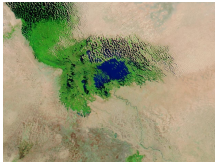
Practitioner's Note

Collect test images the way users will take them: their phones, their fields, their lighting. Add an "unsure, ask an expert" outcome.

Earth observation



NASA, Public domain



NASA MODIS Rapid Response, Public domain

- **Decisions**: plan land use, respond to floods, monitor deforestation and lakes.
- **Tasks**: land-cover classification and segmentation, change detection, building and road mapping.
- **Data**: free Sentinel-2 (10 m, every few days) and Landsat archives; labels are scarce.
- **Specifics**: many spectral bands, huge images (tiling), clouds, and coordinates that must be preserved.

Practitioner's Note

Foundation models pre-trained on satellite imagery (rather than ImageNet) are increasingly available; try them before training from scratch.

Earth observation: more than three colours

- Sentinel-2 records 13 spectral bands, including near-infrared, where healthy vegetation reflects strongly.
- Classic index from two bands, still widely used:

$$\mathrm{NDVI} = \frac{\mathrm{NIR} - \mathrm{Red}}{\mathrm{NIR} + \mathrm{Red}}$$

- Networks take all bands as input channels: the first convolution simply has 13 input channels instead of 3 (Week 2, Day 1).

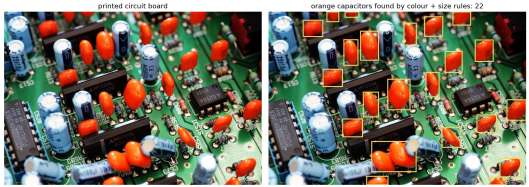
NDVI

Healthy crop: NIR 0.50, Red 0.08: $\mathrm{NDVI} = 0.42/0.58 = 0.72$.
Bare soil: NIR 0.30, Red 0.25: $0.05/0.55 = 0.09$.
Water: NIR 0.02, Red 0.05: $-0.03/0.07 = -0.43$.

Intuition

Week 1, Day 2: a "colour" is just a measurement per band. Satellites measure bands our eyes cannot see.

Industrial inspection and infrastructure

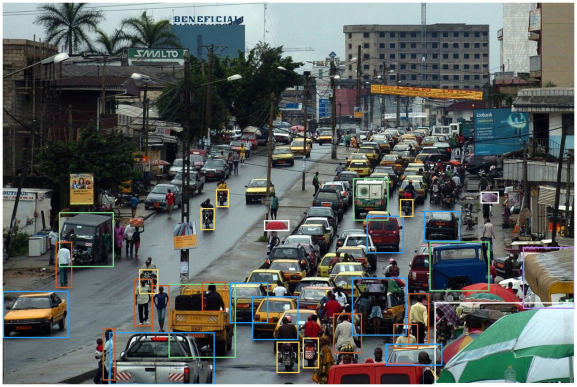


Ashley Pomeroy, CC BY 4.0

Sunolafjagtenben-hur, CC0

- **Decisions**: reject a part; send a crew to repair a road or wall.
- **Tasks**: anomaly detection (Day 1), defect detection and segmentation, measurement.
- **Data**: many good examples, few defects; controlled lighting in factories, uncontrolled outdoors.
- **Deployment**: fast, on the line; false alarms have a cost (Day 1's worked example).

Transport and cities



Happiraphael, CC BY-SA 4.0

- **Decisions**: plan roads and public transport; target road-safety campaigns.
- **Tasks**: detection and tracking (Week 2, Day 3), counting, helmet use (capstone P12), pothole mapping (P5).
- **Data**: local classes (moto-taxis, minibuses) missing from COCO.
- **Risks**: surveillance and privacy (Day 4).

Robotics and autonomous systems



Guaka, CC BY-SA 4.0

- Robots must perceive *to act*: where is the object, how far, how to grasp it, where is free space?
- **Tasks**: depth estimation (above), segmentation of free space, object pose, visual localisation and mapping (SLAM, built on Week 1's features).
- **Constraints**: real time, on-board compute, safety.
- Vision–language–action models (2023 onwards) let robots follow spoken instructions.

Intuition

For a robot, a wrong output is not a wrong label but a wrong movement. Robustness and uncertainty (Day 4) become safety requirements.

Conservation and ecology



Dodro, CC BY 4.0

- **Camera traps**: filter empty frames; identify species; count individuals (capstone P9).
- **Drones and satellites**: count large animals, map habitats, detect illegal logging.
- **Individual identification**: stripes, spots and ear notches as natural "fingerprints".
- **Data**: LILA BC hosts large public camera-trap collections, including from Africa.

Practitioner's Note

Split by camera location, not by image: animals and backgrounds repeat within a site, and a random split overestimates accuracy badly.

Documents, culture and accessibility

- **OCR and document understanding**: digitising archives, forms, receipts and handwritten records; land registries; school records.
- **Cultural heritage**: restoring old photographs (Day 2's inpainting), cataloguing museum collections with CLIP search.
- **Accessibility**: reading text aloud, describing scenes, recognising banknotes (capstone P10).

Intuition

Some of the highest-impact uses of vision are not glamorous: turning paper into searchable data, or helping someone read a medicine label.

Practitioner's Note

OCR for local languages and scripts, and handwriting in French and English as written locally, are under-served and worth building datasets for.

Applications at a glance

Domain	Main tasks	Course tools	Main risk
Microscopy	detect, count	YOLO, colour rules	missed positives
Radiology	classify, segment	transfer learning, U-Net	hospital-to-hospital shift
Agriculture	classify, segment severity	CLIP probe, fine-tuning, SAM	field and season shift
Earth observation	segment, change detection	U-Net, multispectral input	clouds, label scarcity
Industry	anomaly, defect detection	autoencoders, YOLO	false-alarm rate
Cities	detect, track, count	YOLO + tracking	privacy, surveillance
Robotics	depth, pose, free space	depth models, SLAM	safety
Conservation	filter, classify, count	classifiers, detectors	site leakage
Documents	OCR, layout	detectors, VLMs	local scripts and languages

Intuition

The same toolbox everywhere; the last column changes, and it is where most real projects succeed or fail.

From prototype to deployment

1. **Prototype**: a notebook with a good test score.
2. **Pilot**: the real users, real devices, real images; measure again.
3. **Integrate**: into a workflow; who sees the output, who checks errors?
4. **Monitor**: accuracy drifts as cameras, seasons and populations change.
5. **Maintain**: re-label, retrain, re-validate.

Intuition

Most of the work in a real vision system happens after the model is "finished". Reported accuracy is a starting point, not a guarantee.

Practitioner's Note

Day 4 covers monitoring, documentation (model cards, datasheets) and the human side of deployment.

Choosing your capstone approach with today's tools

1. **Hour 1**: zero-shot CLIP with three prompt styles on a validation split: a free baseline.
2. **Hour 2**: CLIP-feature deduplication; check the split for leakage.
3. **Hour 3**: linear probe on CLIP or DINO features; plot accuracy against labels per class.
4. **Then**: fine-tune (Week 2) only if the probe falls short.

Intuition

Cheap experiments first. Each step tells you whether the next, more expensive one is worth doing, and gives you a row in your results table.

Practitioner's Note

Encode your images once and save the features to disk: every later experiment then runs in seconds, even on a laptop CPU.

Part 7

Wrapping up

Key takeaways

1. **CLIP**: image and text encoders trained contrastively on 400M pairs; one shared space.
2. **Zero-shot**: classes become sentences; useful with no labels; weak on specialist images (41.4% at best on beans, where guessing gives 33%).
3. **Linear probe on CLIP features**: a few labels go a long way (76.2% with 10 per class).
4. CLIP reads text in images, struggles with counting, relations and negation, and reflects its data's biases.
5. Multimodal assistants = vision encoder + projector + LLM; powerful and prone to hallucination.
6. Across domains, the toolbox is shared; decisions, data shift and risks differ.

Before the next lecture

Practise (tutorial with the tutors)

- Zero-shot CLIP on your capstone images: try three prompt styles on a validation split.
- Add a linear probe; plot accuracy against labels per class.
- Build a text search over your own photos.

Capstone: main model trained by today; tomorrow is for evaluation, failure analysis and slides, with code frozen by the evening.

Next lecture: responsible vision

- When images change: robustness, measured
- Can we trust the confidence?
- Fairness and hidden subgroups
- Privacy, consent, surveillance, deepfakes
- Documentation, regulation, deployment
- The road ahead

Credits

CLIP ViT-B/32: OpenAI, via Hugging Face. iBean dataset: Makerere AI Lab and NaCRRI, Uganda (MIT licence). All CLIP experiments were run for this lecture on a laptop CPU with `Week3/tools/make_week3_figures.py`; other figures reused from Weeks 1–2.

Photographs from Wikimedia Commons:
`cat_leaves`: RyanDannyChacko, CC BY-SA 4.0
`malaria_smear`: OpenFlexure Project, CC BY-SA 4.0
`histo_breast`: Mikael Häggström, CC0
`maize_blight`: Chroanch, CC BY-SA 4.0
`tomato_blight`: Dwight Sipler from Stow, MA, USA, CC BY 2.0
`xray_tb`: B. Herreros et al., CC BY 4.0
`lake_chad_apollo7`: NASA, Public domain
`lake_chad_2024`: NASA MODIS Rapid Response, Public domain
`pcb`: Ashley Pomeroy, CC BY 4.0
`crack_wall`: Sunolafjagtenben-hur, CC0
`douala_deido`: Happiraphael, CC BY-SA 4.0
`douala_2014`: Guaka, CC BY-SA 4.0
`camera_trap_fox`: Dodro, CC BY 4.0