FULCRUM COURSE NOTES · 02

Computer Vision: Advanced

Detection, segmentation, transformers, generative models and video: the architectures behind modern vision systems, and where each one breaks.

FULCRUM

Utsav Poudel
Founder, Fulcrum

ABOUT FULCRUM

Who made this, and why it is free

Fulcrum is a volunteer-run non-profit. We help people from under-resourced regions learn artificial intelligence through one-on-one mentorship, research and publication supervision, project support, workshops, and a guest speaker series.

The premise is simple. Ability is spread evenly across the world; opportunity is not. What separates a capable student in Pokhara or rural Bihar from one at a well-funded lab is rarely intelligence. It is access to supervision, to reviewers, to equipment, and to the unwritten rules nobody writes down.

This deck is part of how we try to close that. It is free, it will stay free, and you may reuse and adapt it for your own teaching under CC BY 4.0. If it is useful to you, come and find us.

Find us

Website
fulcrumgo.github.io

More material
fulcrumgo.github.io/resources

Discord
discord.gg/gbQCGkupdJ

LinkedIn
linkedin.com/company/gofulcrum

Every deck we publish is free to download at fulcrumgo.github.io/resources

PART ONE

Beyond one label per image

Assumes the Foundations deck: convolution, receptive fields, and honest evaluation.

Object detection: two families

- **Two-stage** (R-CNN → Fast → Faster R-CNN): propose regions, then classify and refine each. More accurate, slower.
- **Single-stage** (YOLO, SSD, RetinaNet): predict boxes and classes in one pass over a grid. Faster, historically less precise on small objects.
- **Anchors** are prior boxes at fixed scales and aspect ratios; the network learns offsets from them.
- **Non-maximum suppression** removes duplicate boxes for the same object, a post-processing step, not a learned one.

RetinaNet's contribution was focal loss: down-weighting the flood of easy background examples that otherwise dominates a single-stage detector's gradient.

Segmentation architectures

- **Fully convolutional networks** replace the classifier head with convolutions, so the output stays a spatial map.
- **U-Net** adds skip connections between encoder and decoder, restoring the fine spatial detail that downsampling destroyed. Designed for biomedical images; still a strong default.
- **Mask R-CNN** bolts a mask branch onto Faster R-CNN to get instance segmentation.
- **Dilated convolution** enlarges the receptive field without losing resolution, useful when you cannot afford to downsample.

Why U-Net keeps winning

Medical datasets are small and detail matters at the boundary. Skip connections address exactly that. Matching an architecture to a data regime beats reaching for the newest model.

Residual connections, and why depth became possible

- Stacking layers naively makes networks *harder* to train. Deeper models were performing worse than shallower ones.
- **ResNet** (2015) has each block learn a residual `F(x)`, with output `F(x) + x`.
- That identity path gives gradients a clean route backwards, so hundreds of layers become trainable.
- The idea generalised well beyond vision. Residual connections are inside essentially every transformer too.

PART TWO

Attention arrives in vision

The convolutional prior is powerful. It is also an assumption, and assumptions can be paid for with data.

Vision Transformers

- **ViT** cuts an image into fixed patches, embeds each as a token, and runs a standard transformer encoder.
- It discards convolution's built-in assumption that nearby pixels belong together, so it needs more data, or strong augmentation and regularisation, to match a CNN.
- Given enough data it scales better than CNNs, which is why it dominates at large scale.
- **Swin Transformer** reintroduces locality with shifted windows, recovering efficiency and hierarchy.

Practical guidance: below roughly tens of thousands of images, a pretrained CNN is usually still the better starting point. Do not choose an architecture by publication date.

Self-supervised learning

Labels are the expensive part. Self-supervision removes the need for most of them.

- **Contrastive methods** (SimCLR, MoCo) pull two augmented views of one image together in embedding space and push different images apart.
- **Masked image modelling** (MAE) hides most patches and trains the model to reconstruct them.
- **DINO** trains without negatives and produces features whose attention maps segment objects with no segmentation labels at all.
- Pretrain on unlabelled data, then fine-tune on the small labelled set you can actually afford.

For under-resourced research groups this is the most important development of the last decade. Unlabelled images are free; annotation is not.

Generative models

- **GANs** pit a generator against a discriminator. Sharp results, but unstable training and prone to mode collapse.
- **VAEs** learn a probabilistic latent space. Stable, principled, and typically blurrier.
- **Diffusion models** learn to reverse a gradual noising process. They now dominate image generation on quality and controllability.
- Diffusion trades compute for quality: sampling takes many steps, though distillation techniques are closing that gap.

The intuition for diffusion

Destroying an image with noise is easy and needs no learning. So train a network to undo one small step of it, then run that reversal repeatedly from pure noise.

Vision-language models

- **CLIP** trains image and text encoders jointly so matching pairs land near each other in a shared space.
- That yields **zero-shot classification**: score an image against text prompts you invent at inference time, with no retraining.
- **Captioning and VQA** models attach a language model to a vision encoder to describe images and answer questions about them.
- These systems inherit their training data's biases wholesale, and the data is web-scraped, so audit before deploying anywhere consequential.

Video: time is not just another dimension

- **Optical flow** estimates per-pixel motion between frames, the classical foundation.
- **3-D convolutions** extend kernels across time; expensive but effective for short clips.
- **Two-stream** designs process appearance and motion separately, then fuse.
- **Tracking** couples per-frame detection with association across frames; the hard part is occlusion and identity switches, not detection.

Video multiplies your data volume, annotation cost and compute all at once. Be sure the problem genuinely needs temporal information before taking that on.

Deployment realities

- **Quantisation** to 8-bit typically gives large speedups for small accuracy loss.
- **Pruning** and **distillation** shrink models, distillation often preserves accuracy better.
- **Edge deployment** (ONNX Runtime, TensorRT, TFLite) matters when connectivity is unreliable, which is exactly the setting many of our mentees work in.
- Measure latency on the **target device**, not your laptop. The numbers differ by an order of magnitude.

Failure modes worth expecting

- **Distribution shift**, a model trained on one hospital's scanner degrades on another's. This is the rule, not the exception.
- **Shortcut learning**, the model finds a ruler in the image that correlates with the diagnosis, and uses that instead.
- **Adversarial fragility**, imperceptible perturbations flip predictions.
- **Dataset bias**, accuracy that is excellent on average and poor for the group you most needed it to serve.

Report per-group performance, not just aggregate. If you cannot, say so in the limitations section. Reviewers respect that far more than a suspiciously round number.

Choosing a research question that will survive review

- Novelty in architecture is the hardest and most crowded route. Novelty in **application, dataset, or careful evaluation** is often more publishable.
- A well-documented dataset from an under-represented setting is a real contribution.
- Reproducing a published result and reporting where it fails is a real contribution too, and more useful than most incremental papers.
- Ask early: what would change my mind about this hypothesis? If nothing would, it is not yet a research question.

ABOUT FULCRUM

Bring us the project

Fulcrum mentors work through exactly this: scoping, evaluating honestly, and getting to submission. Free, and open to anyone from an under-resourced region.

Fulcrum is a volunteer-run non-profit helping people from under-resourced regions learn AI through free mentorship, research supervision, and open material like this deck. Founded 2025 in Kathmandu, Nepal. Nothing we do costs anything.

Website
fulcrumgo.github.io

Every deck, free
fulcrumgo.github.io/resources

Apply for mentorship
discord.gg/gbQCGkupdJ

Partnerships
linkedin.com/company/gofulcrum