**Computer Vision At Small Scale**

[[Asad iqbal](https://medium.com/@deasadiqbal?source=post_page-----8650519ef088--------------------------------)](https://medium.com/@deasadiqbal?source=post_page-----8650519ef088--------------------------------)

[Asad iqbal](https://medium.com/@deasadiqbal?source=post_page-----8650519ef088--------------------------------)

·

[Follow](https://medium.com/m/signin?actionUrl=https%3A%2F%2Fmedium.com%2F_%2Fsubscribe%2Fuser%2F7b3963afad1e&operation=register&redirect=https%3A%2F%2Fmedium.com%2F%40deasadiqbal%2Fcomputer-vision-at-small-scale-8650519ef088&user=Asad+iqbal&userId=7b3963afad1e&source=post_page-7b3963afad1e----8650519ef088---------------------post_header-----------)

7 min read

·

4 days ago

51

**Introduction**

Over the past decade, many projects involving computer vision (CV) have emerged, both in small scale proof-of-concept projects and bigger production applications. Typically:

* medical diagnosis help using radiography, biopsy and other medical images
* satellite imagery to analyse buildings, land use, etc.
* object detection and tracking in various contexts, like traffic estimation, waste estimation, etc.



Plastic waste identification in rivers by

[Surfrider Foundation](https://medium.com/u/a53e7e139732)

The go-to-method for applied computer vision is quite standardised:

* define the **problem** (classification, detection, tracking, segmentation), the **input data** (size and type of picture, field of view) and the **classes** (precisely what we are looking for)
* Annotate some pictures
* pick a network architecture, train — validate, get some statistics
* build the inference system and deploy it

By the end of 2023, the AI field is stormed by new success coming from generative AI: large language models (LLMs), and image generative models. It’s on everyone’s lips, does it change anything for small scale computer vision applications?

We’ll explore if we can leverage them to **build datasets**, leverage **new architectures and new pre-trained weights**, or **distilling knowledge** from big models.

**Small Scale Computer Vision**

What we are typically interested in here are applications that can be built and deployed at a relatively small scale:

* 💰 the cost of development should not be too high
* 💽 it should not require a monster infrastructure to train (think compute power and data scale)
* 🧑‍🔬 it should not require strong research skills, but rather apply existing techniques
* ⚡ the inference should be lightweight and fast, so that it could be embedded or deployed on CPU servers
* 🌍 The overall environmental footprint should be small (think compute power, general size of models / data, no specific hardware requirement)

This is clearly not the trend in AI these days, as we see models with billions of parameters starting to be standard in some applications. We hear a lot about these, still it’s important to remember that caring about smaller scale / footprint is critical, and that not all projects should follow the scale trends of Google, Meta, OpenAI or Microsoft. Even if they’re not in the spotlight, most interesting computer vision projects are actually at a much smaller scale than the ones making the headlines.

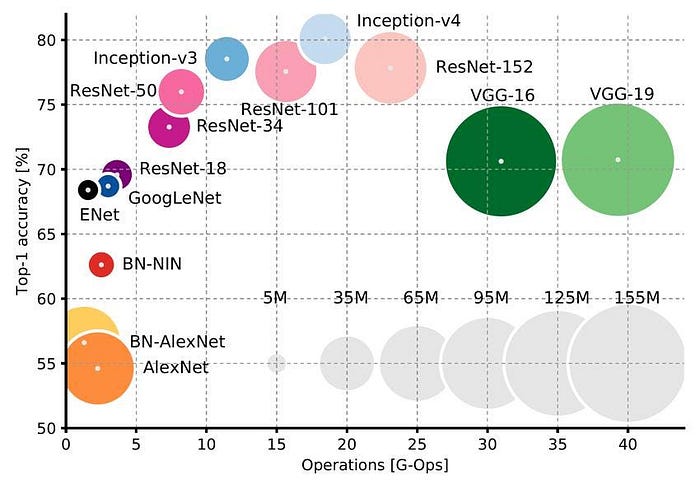
This does not mean that the impact of the application should be small or narrowed, just that we actively care about the development and inference costs.

With this in mind, can we still take advantage of recent developments in AI for our applications? Let’s first dive into the world of foundation models to understand the context.

**Foundation models in Computer Vision**

New Large Language Models (LLM) have been popular because you can easily use **foundation** **models** in your applications (many are open source, or usable through an API). Think about GPT, Bert, Llama as such models. A foundation model is a very large, generic neural network which is useful as a basis for most downstream tasks. It contains knowledge about a very broad range of topics, semantics, syntax, different languages, etc.

In Computer Vision we’ve been using such models for a while: it’s been standard in the last 10 years to use a neural network pre-trained on ImageNet (1 million labelled images) as a “foundation” model for a downstream task. You can build your neural network on top of it, and fine-tune it on your own data if needed.



For the last 10 years, the big question was the performance trade-offs of models on ImageNet

There are two main conceptual difference between ImageNet pre-trained networks and LLMs:

* the type of data we train it on: ImageNet relies on purely supervised learning: a large scale classification task, while LLM are generative models: they are trained in a self-supervised manner using raw text (the task is just to predict the next words).
* the adaptation of these foundation models to new tasks: ImageNet pre-trained network systematically requires a **new learning procedure** to be adapted to a new task. For LLMs, while it’s possible to fine-tune the models, the model is powerful enough to be used for a downstream task without any further training, just by **prompting** the model with the right information to make it useful for a new task.

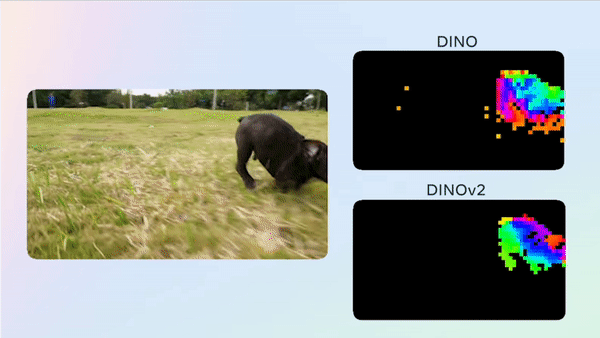
Most current Computer Vision applications such as classification, object detection, segmentation still use ImageNet pre-trained networks. Let’s review new models that are available or about to be, and could be of use for our Computer Vision tasks.

**New foundation models for computer vision: a short review**

In the world of Computer Vision, moving away from ImageNet, there’s been many examples of self-supervised networks, some of them being generative models (think GAN and more recently diffusion models). They are trained on just raw images, or image-text pairs (for instance an image and its description). They are sometimes called LVM (Large Vision Models).

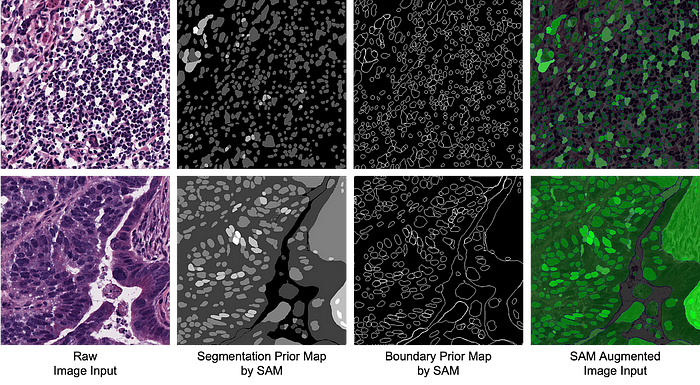
**(Weakly) supervised vision models** on extremely large amounts of data:

* [DINOv2](https://ai.meta.com/blog/dino-v2-computer-vision-self-supervised-learning/) (Meta) — A collection of large ViT (vision transformers, 1B parameters) explicitly aimed at being a good foundation model for Computer Vision, trained in a fully self-supervised manner.



Unsupervised depth estimation using DINOv2, with very good image understanding in the wild

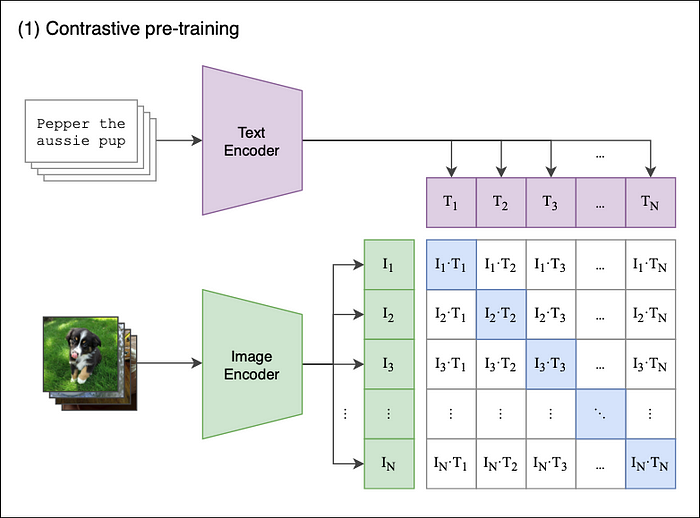
* [SAM Segment Anything](https://segment-anything.com/) (Meta) — a ViT working on high resolution images, specifically designed to be good at segmentation, and enabling zero-shot segmentation (no annotation required to produce new segmentation masks). SAM can be “finetuned*”*[cheaply using LoRA](https://github.com/hitachinsk/SAMed), reducing the amount of necessary training images drastically*.*Another use case is to use SAM as an additional input in [medical image segmentation](https://arxiv.org/abs/2304.11332).



A generic model (SAM) gives a segmentation prior, which is not good enough by itself, but which helps with final segmentation

**Vision-language foundation models trained on image-text pairs**:

* [CLIP](https://openai.com/research/clip) (OpenAI) — alignment of images and short-descriptions, well suited to low-shot classification and used in practice as a foundation model for various downstream CV tasks
* [Scaling Open-Vocabulary Object Detection](https://arxiv.org/abs/2306.09683) (Google)



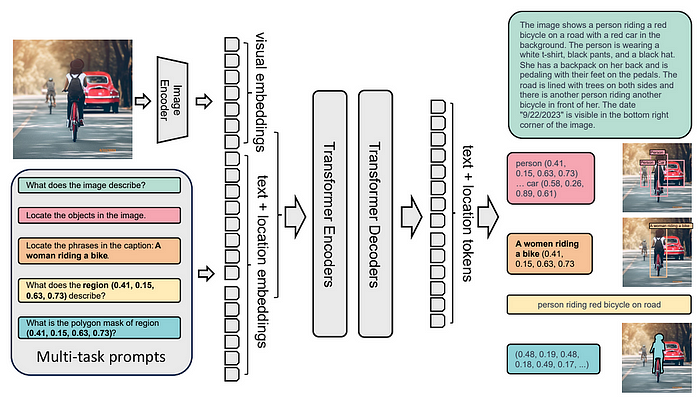
CLIP Contrastive learning procedure, to match Image Features with text encoder features

**Large Generative Models**, which are now multimodal (including large language models capable in their architecture to understand complex text):

* [StableDiffusion](https://stability.ai/stable-diffusion)
* [Dall-E](https://openai.com/dall-e-3) (OpenAI)

**Vision-specialised multi-task large models**:

* [Florence-2: unified Computer Vision](https://arxiv.org/pdf/2311.06242.pdf) (Microsoft)



Florence-2: a multi-task vision-language model, with both semantic and spatial understanding

**The big bad models**, closed source, only available through APIs: Large multi purpose models, not centred on Vision but demonstrating outstanding vision capabilities, with generation capabilities as well:

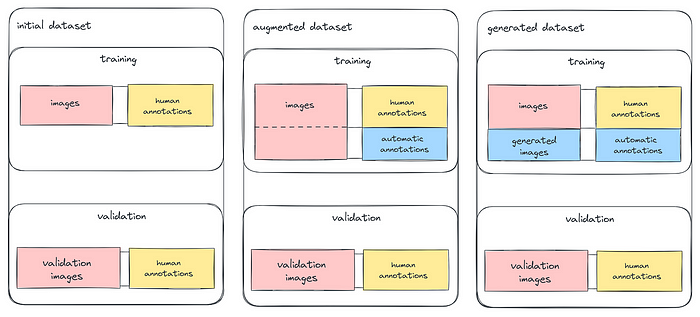
* [GPT-4V](https://cdn.openai.com/papers/GPTV_System_Card.pdf) (OpenAI)
* [Gemini](https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf) (Google)
* Note that there are many open source, smaller mutli-purpose vision+text chat models being developped as well, for instance [LlaVA](https://llava-vl.github.io/" \t "_blank).

All these models are strong foundation models which cover many vision domains and would be good at discriminative or generative tasks in many contexts. Still, how can they be leveraged in our specific, small-scale context?

**Building training datasets**

A pragmatic idea using these new models is to keep our standard training pipeline, for instance using a widely used [Yolo](https://github.com/ultralytics/ultralytics) detector, but improve our dataset by either generating new training images and/or generating annotations. The process is as follows:

* A standard dataset consists of an annotated set of training and validation images
* An**augmented dataset** would use a strong general purpose model to add automatic annotations:  
  1) new annotations to unlabelled images ⇒ this requires a model already suited to the task. You may use a very large general purpose model, carefully fed with examples or prompts, to do zero-shot annotations, or even fine-tune the very large model from your existing human annotations.  
  2) adding a new layer of information to current annotations, for instance adding automatically segmentation annotations from bounding box information using SAM



Standard training, augmented dataset and generated dataset

* A **generated dataset** would consist of generated images alongside with their annotations. You build a careful prompt consisting of images and/or text to generate thousands of images and their annotations. You may directly use an API to generate these annotated images (the cost should be small compared to finding good images and collecting human annotations).

It’s critical to keep the validation set separated from the generated or augmented set as you want to measure actual performance on carefully human labelled data. This means that in practice, we still need to perform *some* manual labelling on real images, even if we chose new generative techniques or foundation models.