

Lecture 1: Course Introduction

Welcome

- This course is BIODS 276 Advanced Topics in Computer Vision and Biomedicine (also CS 286)
- What we hope you will get out of this course:
 - Fluency in cutting edge computer vision models and research
 - Focus on vision and vision-language foundation models, from representation learners to diffusion and generative models
 - Understanding of existing applications and opportunities for future utility in biomedicine
 - Ability to think through considerations for real-world use, including model size and computation, training and inference settings, and training data

Today's agenda

- Brief overview and history of vision and multimodal (vision-language) foundation models
- Course logistics

Foundation models are now powering many applications involving images and videos



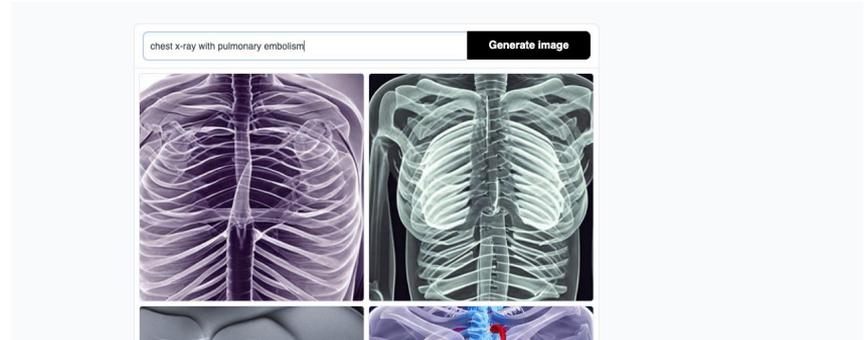
Segment Anything by Meta AI.
<https://segment-anything.com/>

Stable Diffusion Online

Stable Diffusion is a latent text-to-image diffusion model capable of generating photo-realistic images given any text input, cultivates autonomous freedom to produce incredible imagery, empowers billions of people to create stunning art within seconds.

Create beautiful art using stable diffusion ONLINE for free.

Get started →



Stable Diffusion.
<https://stablediffusionweb.com/>

Foundation models are now powering many applications involving images and videos

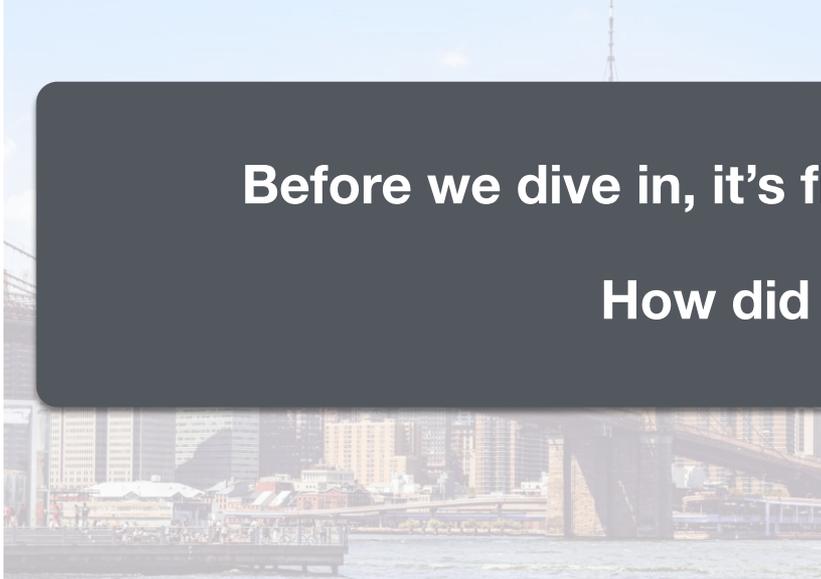
Stable Diffusion Online

Stable Diffusion is a latent text-to-image diffusion model capable of



Before we dive in, it's first helpful to understand...

How did we get here?

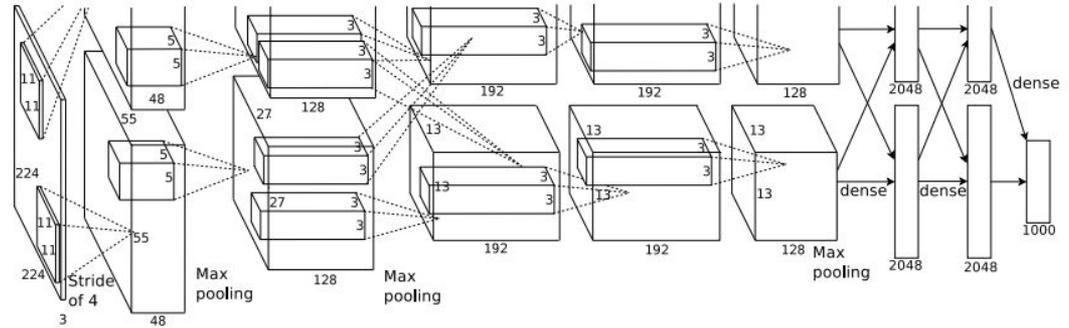
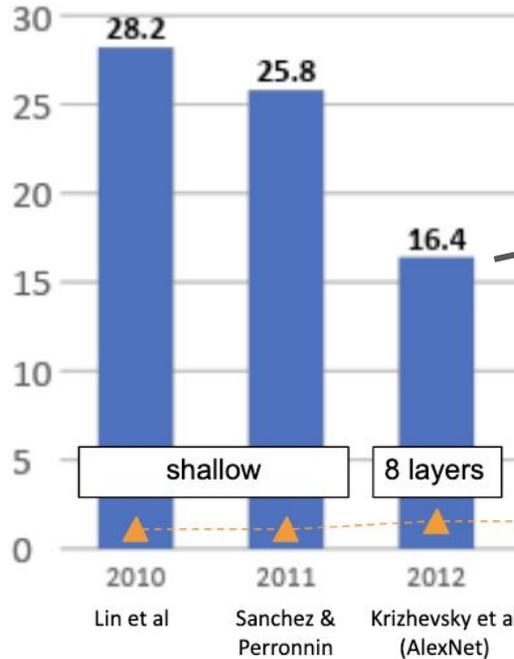


Segment Anything by Meta AI.
<https://segment-anything.com/>



Stable Diffusion.
<https://stablediffusionweb.com/>

2012: Deep learning emerges in AI

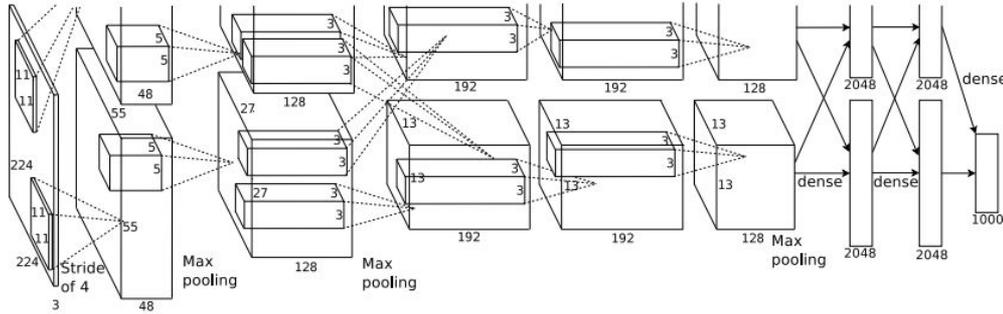


Krizhevsky et al. 2012. 8-layer “AlexNet”.

ImageNet Visual Recognition Challenge results.

Convergence of key ingredients of deep learning

Algorithms



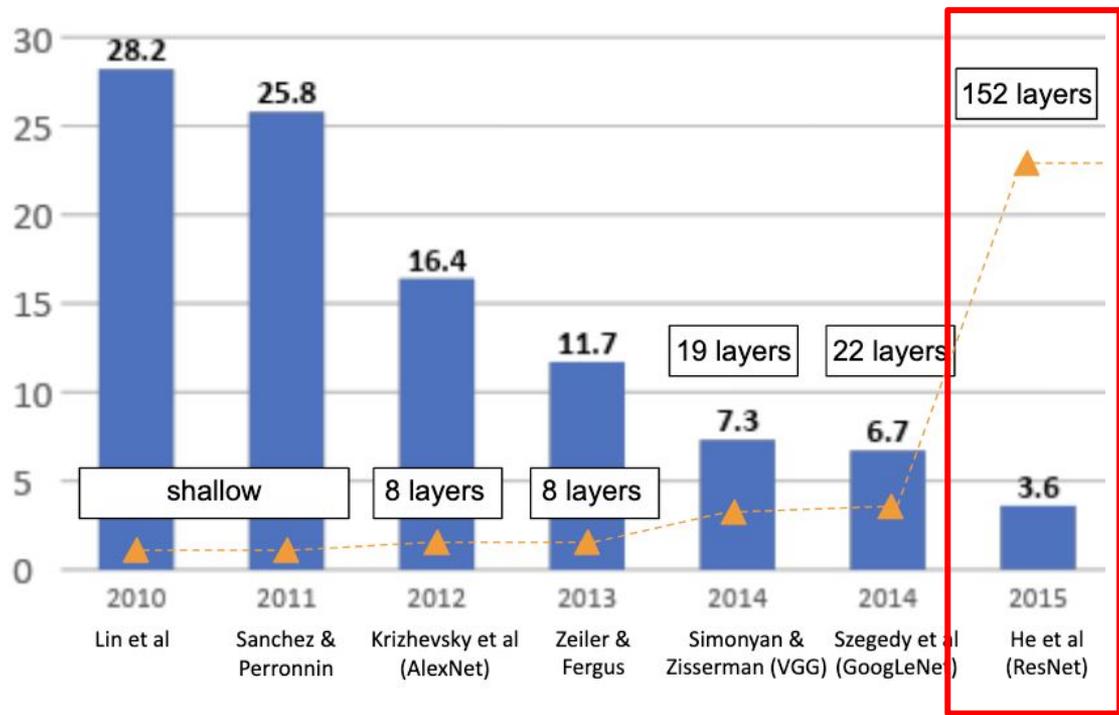
Compute



Data



2015: Very deep convnets and challenging vision tasks

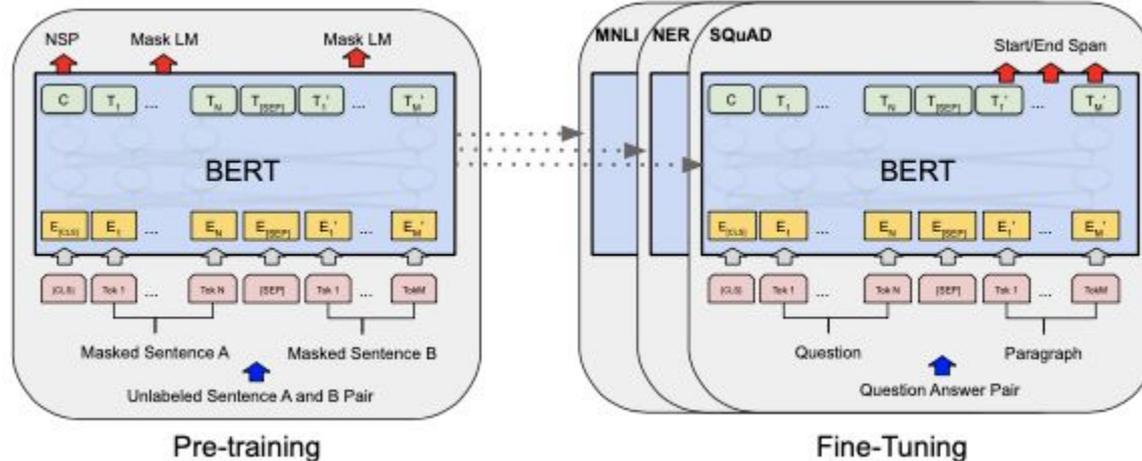


He et al. 2015. ResNet.



2018: Breakthroughs in deep learning for natural language processing (sequences)

Transformer neural network architectures (containing self-attention mechanism) and pre-training -> fine-tuning. State-of-the-art on 11 NLP benchmarks.



Vaswani et al. 2017. "Attention is all you need."

Devlin et al. 2018. BERT.

2020: Very large-scale text generation models

```
SYSTEM PROMPT (HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.
```

Transformer architectures become ubiquitous

Training these models on very large amounts of data -> rise of “foundation models”

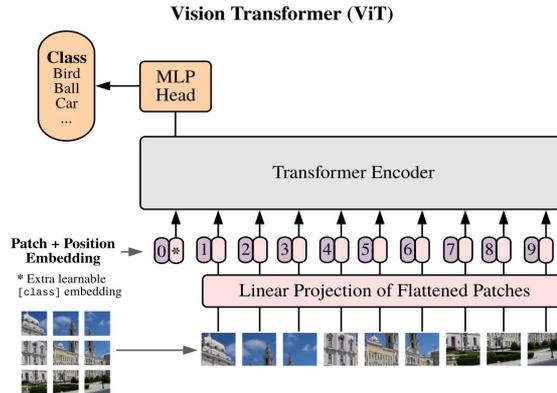
OpenAI GPT-3 (figure from GPT-2). Brown et al. 2020.

2021: Transformers & foundation models reach computer vision

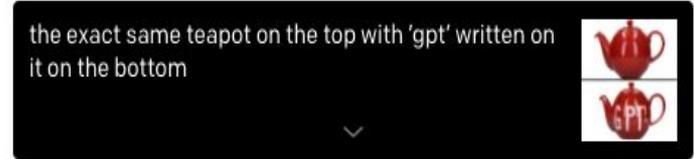
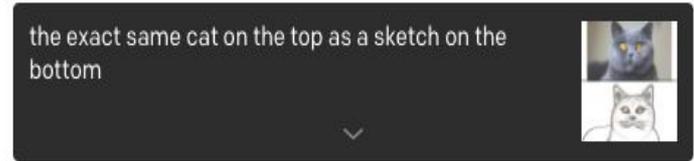
Transformer architecture can be applied to images as well!

Key idea: Convert image into sequence of patches (i.e. image tokens).

Can then benefit from Transformer architecture and its self-attention mechanism. Allows process of token encoding to incorporate contextual understanding (attention) from other tokens of the same (self) sequence.

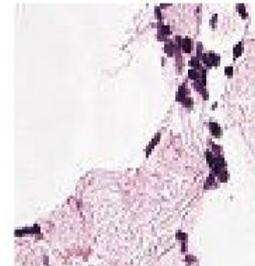


ViT. Dosovitsky et al. 2021



DALL-E. Ramesh et al. 2021.

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



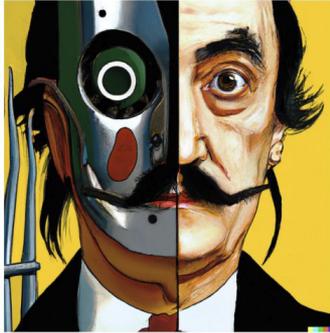
✗ this is a photo of lymph node tumor tissue

✓ this is a photo of healthy lymph node tissue

CLIP. Radford et al. 2021.

2022: Major improvements in text-to-image generation models

DALL-E 2
Ramesh et al. 2022.



vibrant portrait painting of Salvador Dalí with a robotic half face



a shiba inu wearing a beret and black turtleneck

Meta Make-a-Video.
Singer et al. 2022



Stable Diffusion Online

Stable Diffusion is a latent text-to-image diffusion model capable of generating photo-realistic images given any text input, cultivates autonomous freedom to produce incredible imagery, empowers billions of people to create stunning art within seconds.

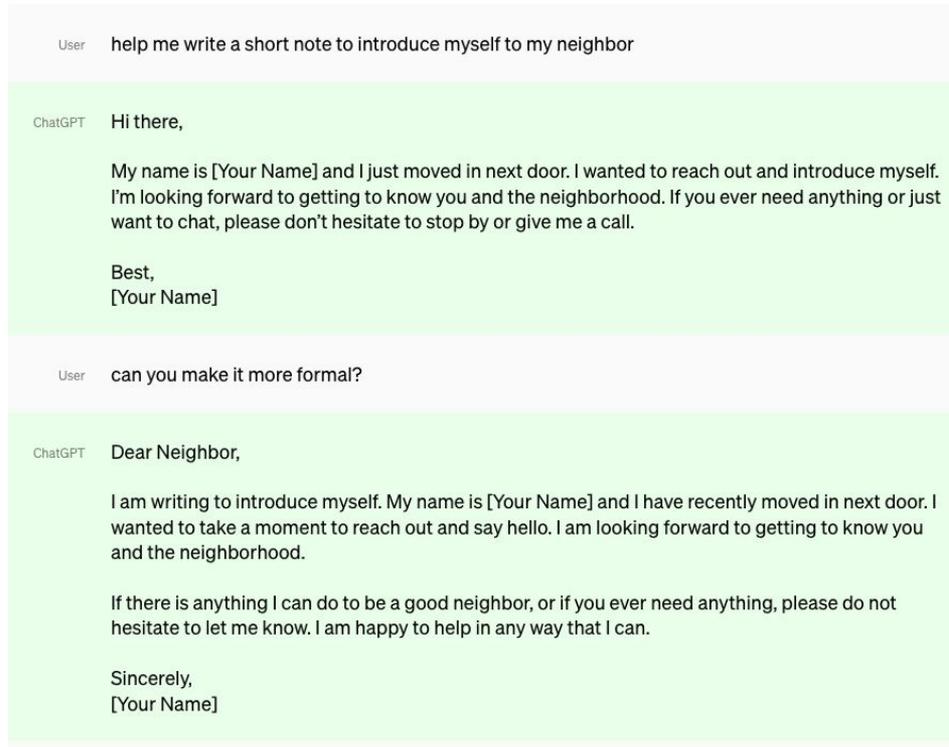
Create beautiful art using stable diffusion ONLINE for free.

Get started →



Stable Diffusion.
<https://stablediffusionweb.com/>

Late 2022: ChatGPT reaches public mainstream



User can you make it more formal?

ChatGPT Dear Neighbor,

I am writing to introduce myself. My name is [Your Name] and I have recently moved in next door. I wanted to take a moment to reach out and say hello. I am looking forward to getting to know you and the neighborhood.

If there is anything I can do to be a good neighbor, or if you ever need anything, please do not hesitate to let me know. I am happy to help in any way that I can.

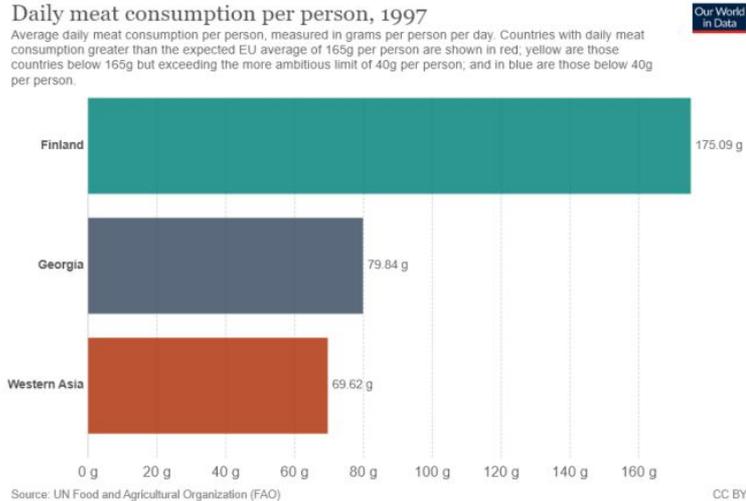
Sincerely,
[Your Name]

OpenAI ChatGPT
<https://openai.com/blog/chatgpt>

2023: Advance in multimodal (e.g. vision-language) generative models

GPT-4 visual input example, Chart Reasoning:

User What is the sum of average daily meat consumption for Georgia and Western Asia?
Provide a step-by-step reasoning before providing your answer.



OpenAI 2023
<https://openai.com/research/gpt-4>

GPT-4 To find the sum of average daily meat consumption for Georgia and Western Asia, we will follow these steps:

1. Identify the average daily meat consumption for Georgia.
2. Identify the average daily meat consumption for Western Asia.
3. Add the values from steps 1 and 2.

Step 1: From the provided chart, we see that the average daily meat consumption for Georgia is 79.84 grams per person per day.

Step 2: From the chart, we see that the average daily meat consumption for Western Asia is 69.62 grams per person per day.

Step 3: Add the values from steps 1 and 2.

$79.84 \text{ g (Georgia)} + 69.62 \text{ g (Western Asia)} = 149.46 \text{ g}$

So, the sum of average daily meat consumption for Georgia and Western Asia is 149.46 grams per person per day.

Late 2023: Excitement over next-generation text-to-video models

Many AI video generation tools emerge, including Runway Gen-2, Pika Labs, Stable Diffusion Video, and others.

OpenAI Sora is also released in early 2024.



Figure credit: Stable Diffusion Video, Stability AI.
<https://github.com/Stability-AI/generative-models>

2024: Heavy competition among industry generative AI products

Arena (battle) Arena (side-by-side) Direct Chat Leaderboard About Us

Chatbot Arena (formerly LMSYS): Free AI Chat to Compare & Test AI Chatbots

Blog | GitHub | Paper | Dataset | Twitter | Discord | Kaggle Competition

New Launch! Jailbreak models at RedTeam Arena.

<https://lmarena.ai/>

Google DeepMind

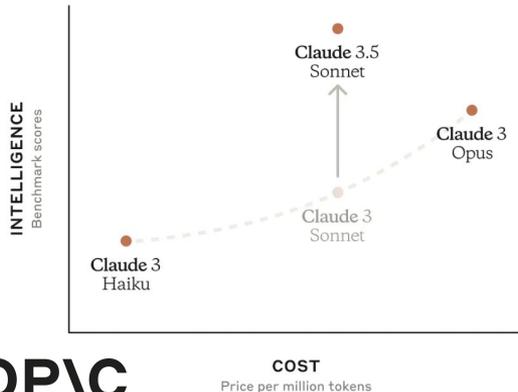
Gemini models Ultra Pro Flash Nano

Gemini Pro

Our best model for general performance across a wide range of tasks

Try Gemini Advanced Build with Gemini

<https://deepmind.google/technologies/gemini/pro/>



ANTHROPIC

<https://www.anthropic.com/news/claude-3-5-sonnet>

The screenshot shows the OpenAI chat interface with a dropdown menu for model selection. The menu is titled 'o1-preview' and lists several options: 'GPT-4o' (Great for most tasks), 'o1-preview' (Uses advanced reasoning, selected), 'o1-mini' (Faster at reasoning), 'More models', and 'Temporary chat' (toggle switch).

OpenAI

<https://openai.com/o1/>

Two major classes of vision foundation models

Representation Learners

Train on huge amounts of data

Objective: learn powerful generalist feature extractors that can be used for downstream computer vision tasks

- Vision Only (e.g. DINOv2)
- Vision-Language (e.g. CLIP)

Generative Models

Train on huge amounts of data

Objective: learn to generate images and/or text

- Text -> Image (e.g. StableDiffusion, DALL-E)
- Text, Image -> Text (e.g. Flamingo, GPT4-V)

Two major classes of vision foundation models

Representation Learners

Train on huge amounts of data

Objective: learn powerful generalist feature extractors that can be used for downstream computer vision tasks

- Vision Only (e.g. DINOv2)
- Vision-Language (e.g. CLIP)

Generative Models

Train on huge amounts of data

Objective: learn to generate images and/or text

- Text -> Image (e.g. StableDiffusion, DALL-E)
- Text, Image -> Text (e.g. Flamingo, GPT4-V)

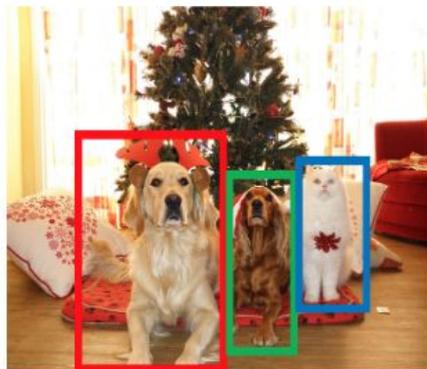
Computer vision spans a range of visual AI tasks

Classification



CAT

Object Detection



DOG, DOG, CAT

Segmentation



DOG, DOG, CAT

Computer vision spans a range of visual AI tasks

Keypoint / pose estimation



Computer vision spans a range of visual AI tasks



Images: Recognize **objects**



Dog
Cat
Fish
Truck



Videos: Recognize **actions**



Swimming
Running
Jumping
Eating
Standing

Computer vision spans a range of visual AI tasks

Image generation

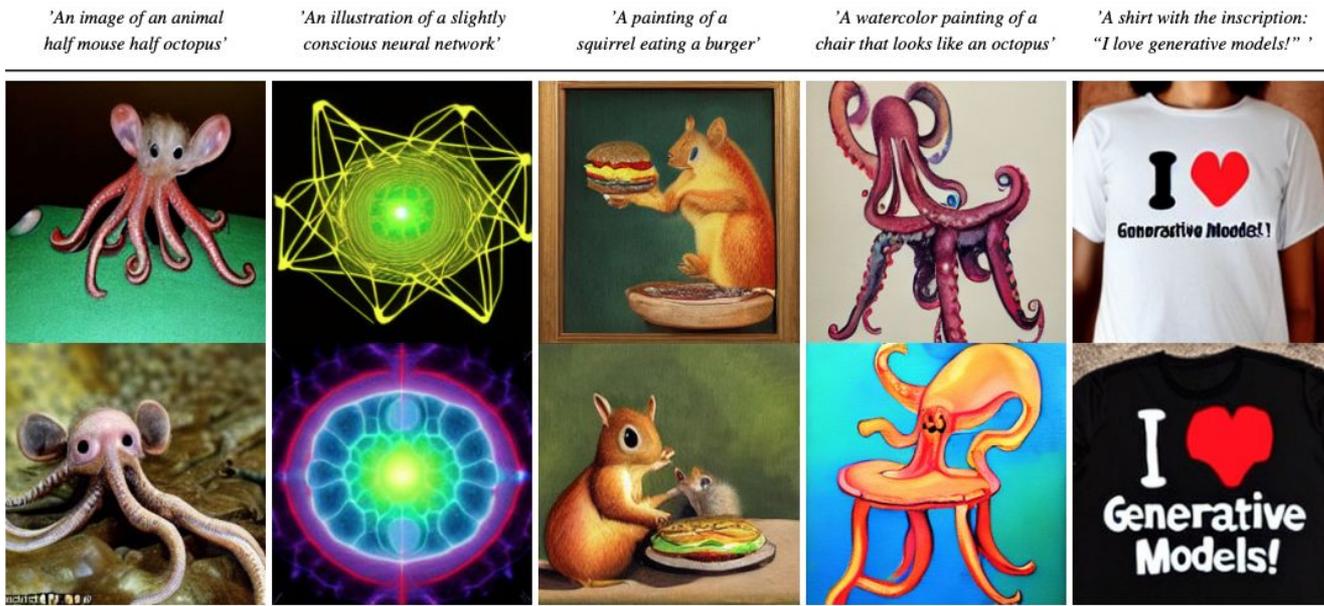


Figure Credit: Rombach et al. 2022

Computer vision spans a range of visual AI tasks

Image generation

These are some of the major types of outputs, but there are many more

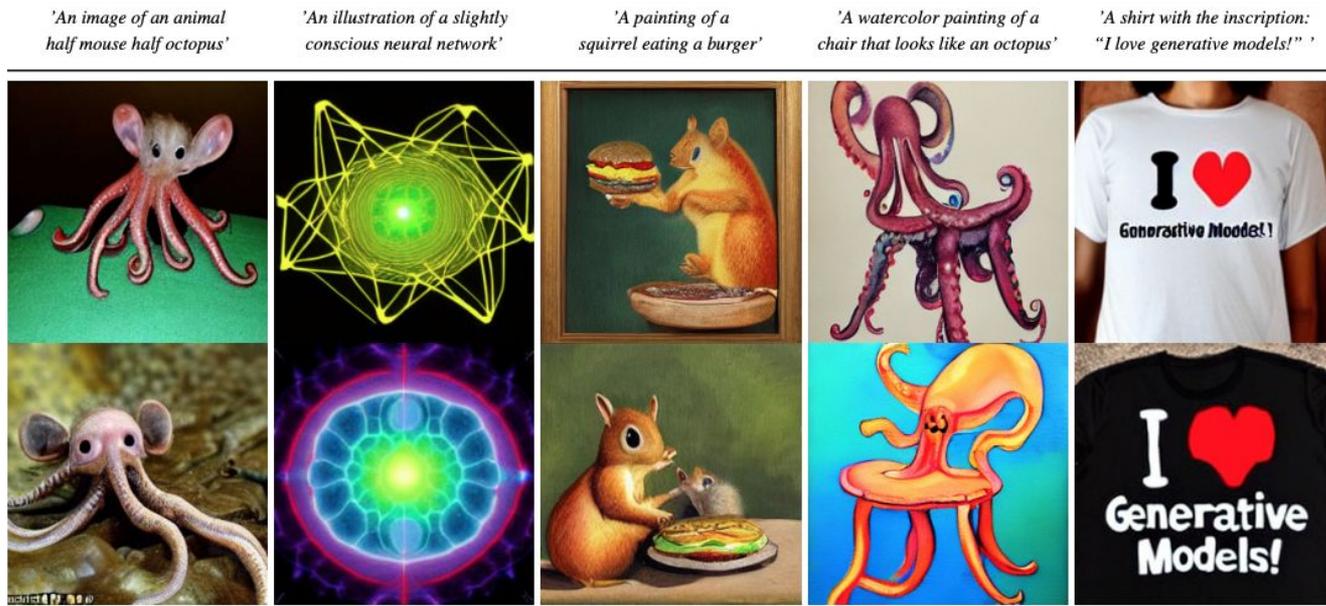


Figure Credit: Rombach et al. 2022

Two major classes of vision foundation models

Representation Learners

Train on huge amounts of data

Objective: learn powerful generalist feature extractors that can be used for downstream computer vision tasks

- Vision Only (e.g. DINOv2)
- Vision-Language (e.g. CLIP)

Generative Models

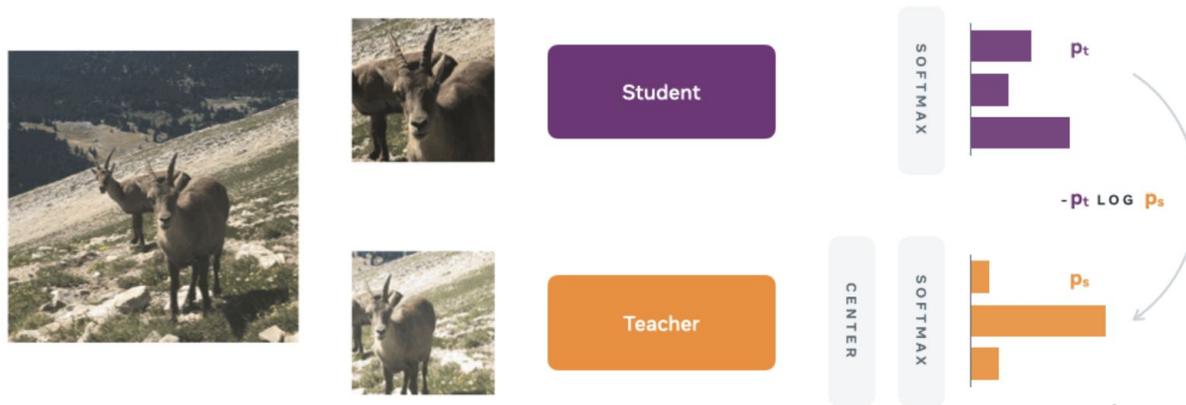
Train on huge amounts of data

Objective: learn to generate images and/or text

- Text -> Image (e.g. StableDiffusion, DALL-E)
- Text, Image -> Text (e.g. Flamingo, GPT4-V)

DINOv2: Transformer-based representation learner trained on 142 million images

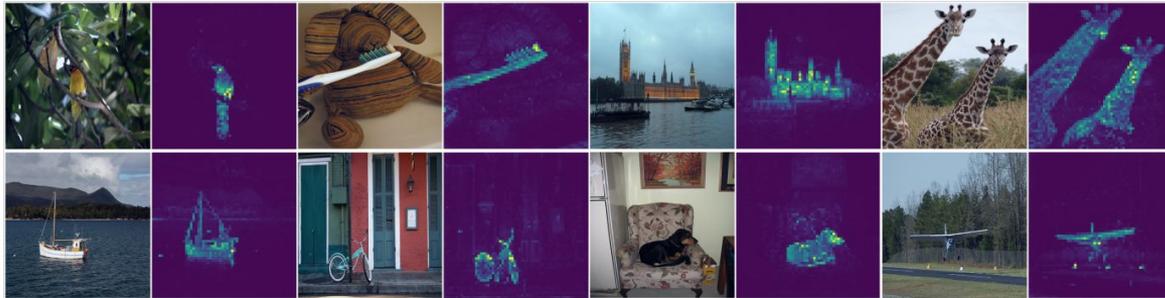
- DINO (self-Distillation with NO labels) is trained through self-supervised learning (no external label curation needed)
- Student-teacher self-supervised training objective: At a high level, images (teacher) should produce similar feature representations to randomly cropped portions of the image (student)



Caron et al. 2021, Oquab et al. 2023

DINOv2: Transformer-based representation learner trained on 142 million images

- Once trained, the DINO neural network backbone can be used as the “pre-trained” feature extractor for many downstream tasks

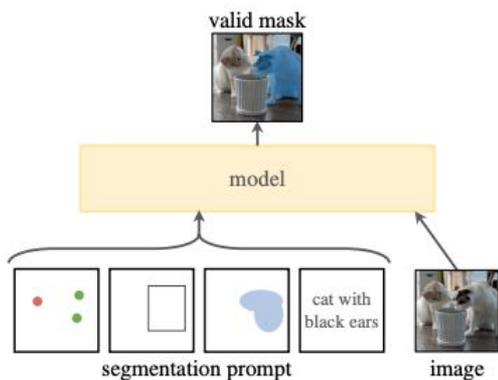


- DINOv2 made further implementation refinements and trained on 142 million images
- An example of a vision “**foundation model**”: a model trained on a very large amount of data that can be subsequently used or tuned for diverse tasks. DINOv2 was demonstrated to be a state-of-the-art backbone for many tasks.

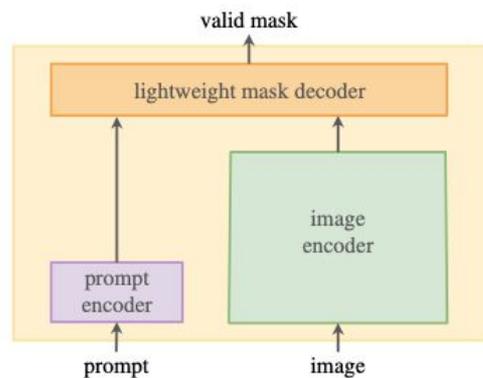
Caron et al. 2021, Oquab et al. 2023

Segment Anything Model (SAM): A foundation model targeted for segmentation

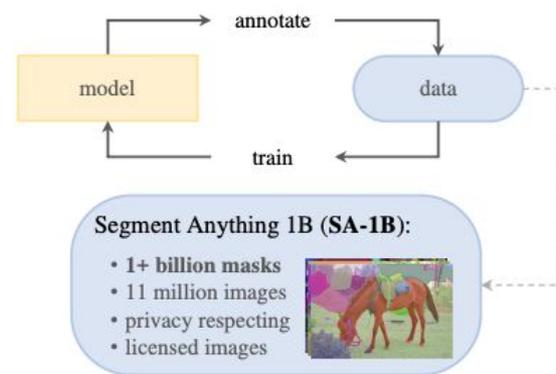
- Foundation Model for promptable segmentation, based on a Transformer encoder-decoder architecture
- Trained on 1 billion masks from 11 million images, using the model in a data collection loop



(a) **Task:** promptable segmentation



(b) **Model:** Segment Anything Model (SAM)



(c) **Data:** data engine (top) & dataset (bottom)

Kirrilov et al. 2023.

Segment Anything Model (SAM): A foundation model targeted for segmentation

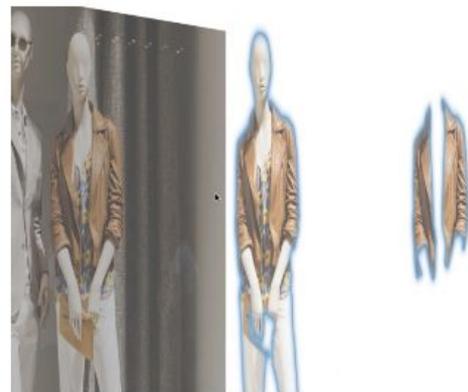
- Allows segmentation based only from prompts, without needing to provide training examples of target classes (zero-shot segmentation)



Prompt it with interactive points and boxes.



Automatically segment everything in an image.



Generate multiple valid masks for ambiguous prompts.

Meta AI Segment Anything Model (SAM)
<https://segment-anything.com/>

Kirrilov et al. 2023.

Two major classes of vision foundation models

Representation Learners

Train on huge amounts of data

Objective: learn powerful generalist feature extractors that can be used for downstream computer vision tasks

- Vision Only (e.g. DINOv2)
- Vision-Language (e.g. CLIP)

Generative Models

Train on huge amounts of data

Objective: learn to generate images and/or text

- Text -> Image (e.g. StableDiffusion, DALL-E)
- Text, Image -> Text (e.g. Flamingo, GPT4-V)

Two major classes of vision foundation models

Representation Learners

Train on huge amounts of data

Objective: learn powerful generalist feature extractors that can be used for downstream computer vision tasks

- Vision Only (e.g. DINOv2)
- Vision-Language (e.g. CLIP)

Generative Models

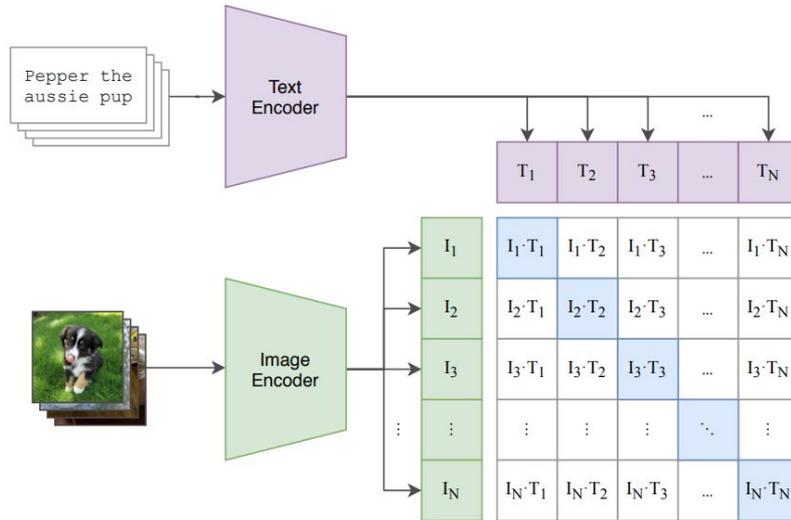
Train on huge amounts of data

Objective: learn to generate images and/or text

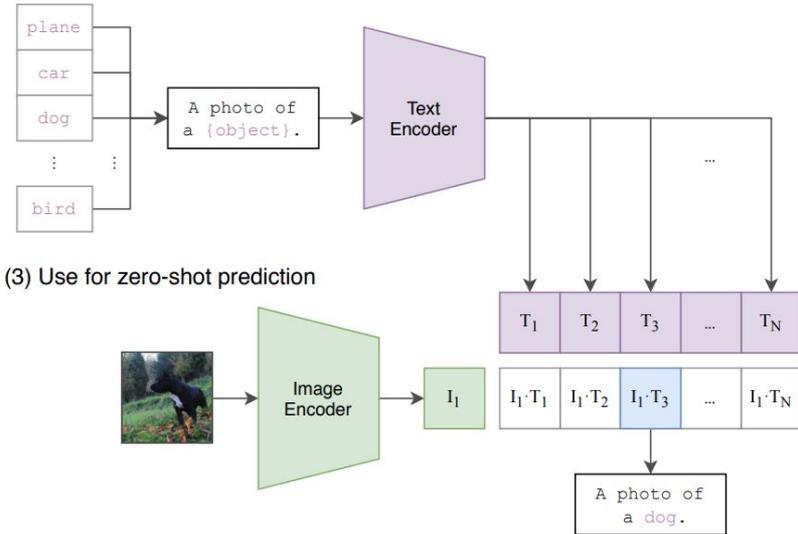
- Text -> Image (e.g. StableDiffusion, DALL-E)
- Text, Image -> Text (e.g. Flamingo, GPT4-V)

CLIP learns a joint representation space for images and text

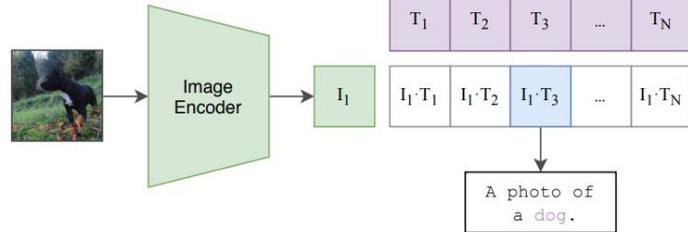
(1) Contrastive pre-training



(2) Create dataset classifier from label text



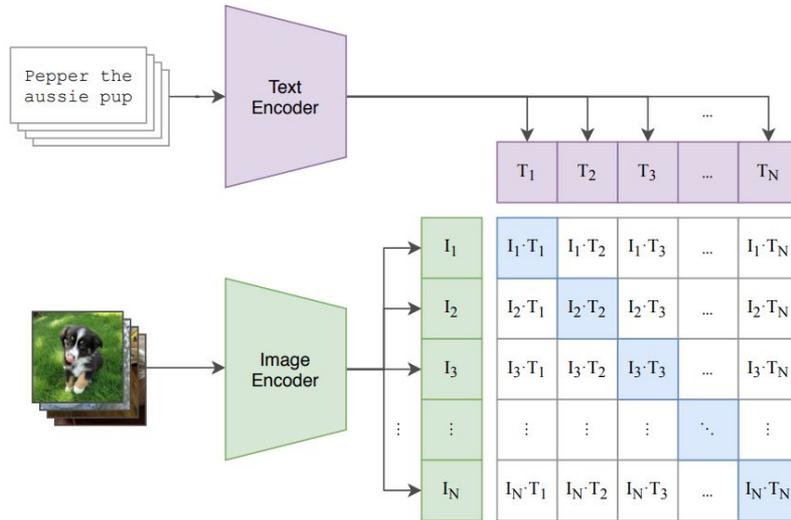
(3) Use for zero-shot prediction



Radford et al. 2021.

CLIP learns a joint representation space for images and text

(1) Contrastive pre-training



Radford et al. 2021.

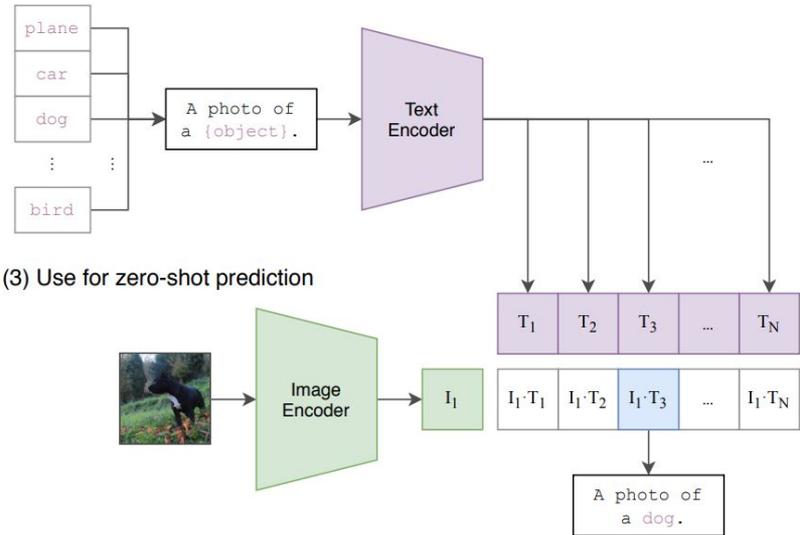
Learn text and image encoders (i.e. feature extractors) through a self-supervised “contrastive” objective: images and their matching text captions should produce more similar features than non-matching ones

Trained on 400M image-text pairs: generated by searching for image-text pairs on the web, where text comes from a base query list of 500,000 queries comprising all words occurring at least 100 times in the English version of Wikipedia

CLIP learns a joint representation space for images and text

Due to the contrastive training objective, image and text encoders output features in a shared representation space and the trained CLIP model can be used for **zero-shot classification**, i.e., perform N-way classification without showing the model paired examples of (input, class) for any of the N classes.

(2) Create dataset classifier from label text



Radford et al. 2021.

Two major classes of vision foundation models

Representation Learners

Train on huge amounts of data

Objective: learn powerful generalist feature extractors that can be used for downstream computer vision tasks

- Vision Only (e.g. DINOv2)
- Vision-Language (e.g. CLIP)

Generative Models

Train on huge amounts of data

Objective: learn to generate images and/or text

- Text -> Image (e.g. StableDiffusion, DALL-E)
- Text, Image -> Text (e.g. Flamingo, GPT4-V)

Two major classes of vision foundation models

Representation Learners

Train on huge amounts of data

Objective: learn powerful generalist feature extractors that can be used for downstream computer vision tasks

- Vision Only (e.g. DINOv2)
- Vision-Language (e.g. CLIP)

Generative Models

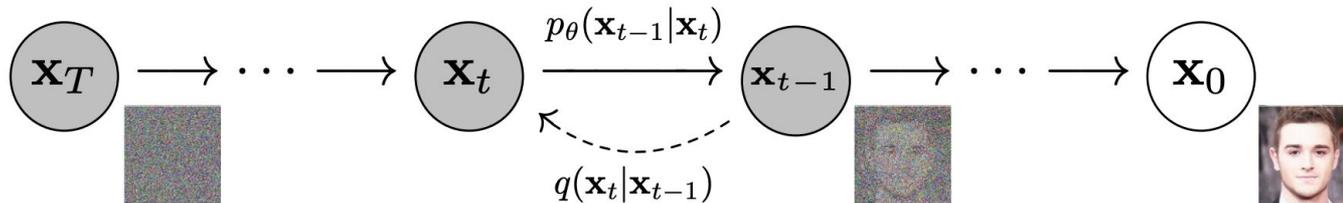
Train on huge amounts of data

Objective: learn to generate images and/or text

- Text -> Image (e.g. StableDiffusion, DALL-E)
- Text, Image -> Text (e.g. Flamingo, GPT4-V)

Diffusion models: key ingredient of recent advances in high-quality image generation

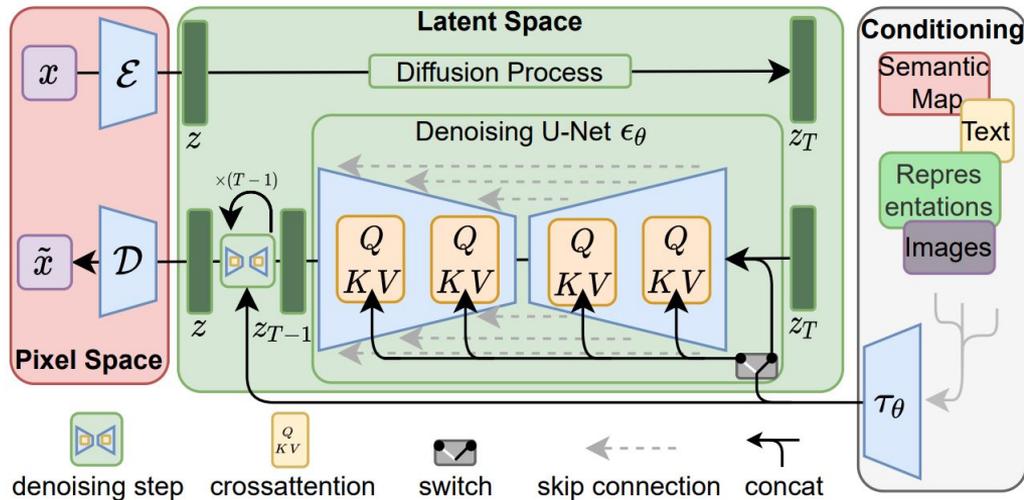
High-level intuition: Progressively add noise to an image, and train a model that learns the reverse process of denoising. Once trained, the reverse process can be applied to transform random noise to images from the data distribution.



Ho et al. 2020.

Diffusion models: key ingredient of recent advances in high-quality image generation

Follow-up improvements move the diffusion process to operate over latent feature representations instead of the raw image, and incorporate **conditioning based on features from text, other images, etc.**



Rombach et al. 2022.

Stable Diffusion: example of text -> image generative model (+ other types of conditioning!)

Text -> image model trained on LAION: open-source dataset of 400M image-text pairs

Condition on provided input images to remove objects



Text-to-Image Synthesis on LAION. 1.45B Model.

'An image of an animal half mouse half octopus'

'An illustration of a slightly conscious neural network'

'A painting of a squirrel eating a burger'

'A watercolor painting of a chair that looks like an octopus'

'A shirt with the inscription: "I love generative models!"'



Rombach et al. 2022.

Proliferation of diffusion-based generative models trained on large data (including video!) and image generation products

Google Imagen.
Saharia et al. 2022.



Meta Make-a-Video.
Singer et al. 2022



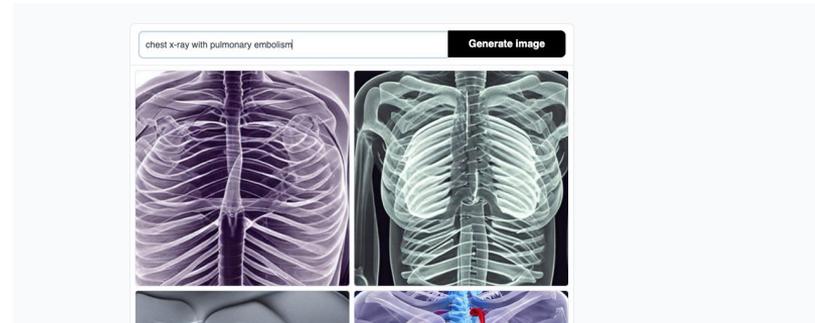
Sora by OpenAI.
<https://openai.com/sora>

Stable Diffusion Online

Stable Diffusion is a latent text-to-image diffusion model capable of generating photo-realistic images given any text input, cultivates autonomous freedom to produce incredible imagery, empowers billions of people to create stunning art within seconds.

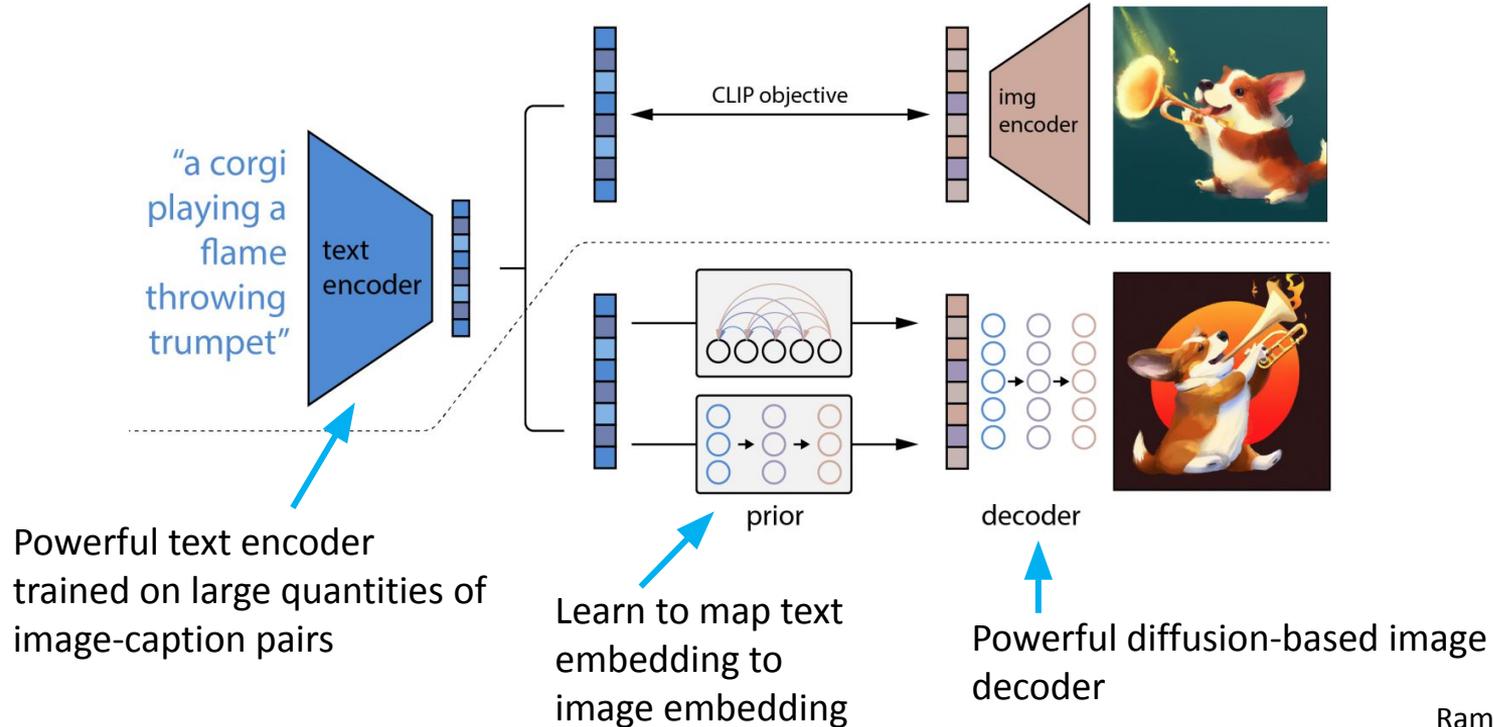
Create beautiful art using stable diffusion ONLINE for free.

Get started →



Stable Diffusion.
<https://stablediffusionweb.com/>

DALL-E: related family of OpenAI models that focus on high quality text-conditioned image generation



Ramesh et al. 2022.

DALL-E: related family of OpenAI models that focus on high quality text-conditioned image generation

DALL-E 3 is trained on large amounts of detailed, synthetically generated captions to enable high-fidelity and high-detail generation



<https://openai.com/dall-e-3>

Two major classes of vision foundation models

Representation Learners

Train on huge amounts of data

Objective: learn powerful generalist feature extractors that can be used for downstream computer vision tasks

- Vision Only (e.g. DINOv2)
- Vision-Language (e.g. CLIP)

Generative Models

Train on huge amounts of data

Objective: learn to generate images and/or text

- Text -> Image (e.g. StableDiffusion, DALL-E)
- Text, Image -> Text (e.g. Flamingo, GPT4-V)

Two major classes of vision foundation models

Representation Learners

Train on huge amounts of data

Objective: learn powerful generalist feature extractors that can be used for downstream computer vision tasks

- Vision Only (e.g. DINOv2)
- Vision-Language (e.g. CLIP)

Generative Models

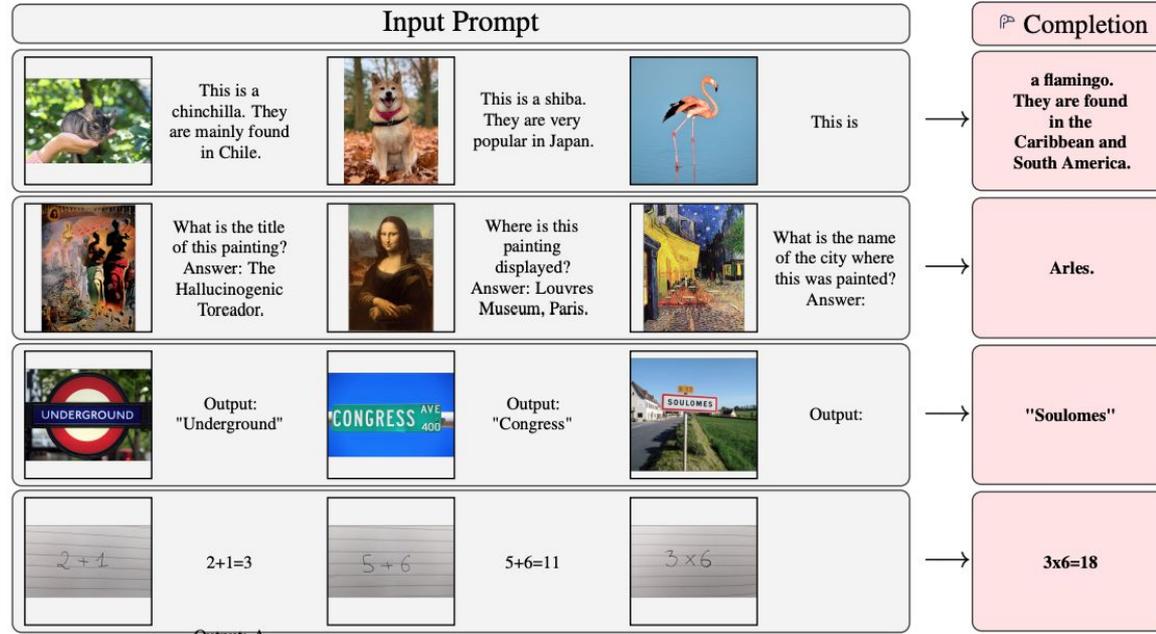
Train on huge amounts of data

Objective: learn to generate images and/or text

- Text -> Image (e.g. StableDiffusion, DALL-E)
- Text, Image -> Text (e.g. Flamingo, GPT4-V)

Extending large language models (LLMs) to take interleaved text and images as input

DeepMind Flamingo



Illustrates same “zero-shot” and “few-shot” capabilities as LLMs, i.e. can solve new tasks out-of-the-box (no need for traditional model fine-tuning on a training dataset), with only zero or few training examples provided in the prompt

Alayrac et al. 2022

Extending large language models (LLMs) to take interleaved text and images as input

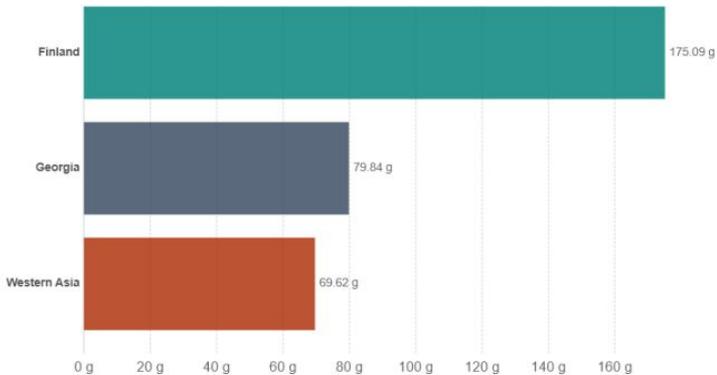
OpenAI GPT-4

GPT-4 visual input example, Chart Reasoning:

User What is the sum of average daily meat consumption for Georgia and Western Asia?
Provide a step-by-step reasoning before providing your answer.

Daily meat consumption per person, 1997

Average daily meat consumption per person, measured in grams per person per day. Countries with daily meat consumption greater than the expected EU average of 165g per person are shown in red, yellow are those countries below 165g but exceeding the more ambitious limit of 40g per person, and in blue are those below 40g per person.



Source: UN Food and Agricultural Organization (FAO)

CC BY

OpenAI 2023

<https://openai.com/research/gpt-4>

GPT-4 To find the sum of average daily meat consumption for Georgia and Western Asia, we will follow these steps:

1. Identify the average daily meat consumption for Georgia.
2. Identify the average daily meat consumption for Western Asia.
3. Add the values from steps 1 and 2.

Step 1: From the provided chart, we see that the average daily meat consumption for Georgia is 79.84 grams per person per day.

Step 2: From the chart, we see that the average daily meat consumption for Western Asia is 69.62 grams per person per day.

Step 3: Add the values from steps 1 and 2.

$$79.84 \text{ g (Georgia)} + 69.62 \text{ g (Western Asia)} = 149.46 \text{ g}$$

So, the sum of average daily meat consumption for Georgia and Western Asia is 149.46 grams per person per day.

Two major classes of vision foundation models

Representation Learners

Train on huge amounts of data

Objective: learn powerful generalist feature extractors that can be used for downstream computer vision tasks

- Vision Only (e.g. DINOv2)
- Vision-Language (e.g. CLIP)

Generative Models

Train on huge amounts of data

Objective: learn to generate images and/or text

- Text -> Image (e.g. StableDiffusion, DALL-E)
- Text, Image -> Text (e.g. Flamingo, GPT4-V)

Two major classes of vision foundation models

Representation Learners

Train on huge amounts of data

Objective: learn powerful generalist feature extractors that can be used for downstream computer vision tasks

- Vision Only (e.g. DINOv2)
- Vision-Language (e.g. CLIP)

Generative Models

Train on huge amounts of data

Objective: learn to generate images and/or text

- Text -> Image (e.g. StableDiffusion, DALL-E)
- Text, Image -> Text (e.g. Flamingo, GPT4-V)

In this class, we will:

- Acquire deep technical understanding of seminal works in each of these classes of models
- Explore ongoing directions of research to further advance these models
- Alternate discussion of computer vision methodology with examples of real world application in biomedicine

Course schedule

Lecture 1	Sep 23 (Mon)	Course Introduction
Lecture 2	Sep 25 (Wed)	Deep Learning for Image Data: From CNNs to Foundation Models
Lecture 3	Sep 30 (Mon)	Vision Representation Learners
Lecture 4	Oct 2 (Wed)	Vision Representation Learners in Biomedicine
Lecture 5	Oct 7 (Mon)	Vision-Language Representation Learners
Lecture 6	Oct 9 (Wed)	Vision-Language Representation Learners in Biomedicine
Lecture 7	Oct 14 (Mon)	Vision Diffusion and Generative Models
Assignment	Oct 16 (Wed)	Assignment 1 due
Lecture 8	Oct 16 (Wed)	Vision Generative Models in Biomedicine
Lecture 9	Oct 21 (Mon)	Vision-Language Generative Models
Project	Oct 23 (Wed)	Proposal due
Lecture 10	Oct 23 (Wed)	Vision-Language Generative Models in Biomedicine
Lecture 11	Oct 28 (Mon)	A Deeper Dive into GPUs and Compute

Lecture 12	Oct 30 (Wed)	Guest Lecture: Advanced Vision Models in Biomedicine
Lecture 13	Nov 4 (Mon)	Discussion: Recent Advances in Vision Representation Learners
Lecture 14	Nov 6 (Wed)	Discussion: Recent Advances Vision Generative Models
Lecture 15	Nov 11 (Mon)	Discussion: Recent Advances in Vision-Language Representation Learners
Assignment	Nov 13 (Wed)	Assignment 2 due
Lecture 16	Nov 13 (Wed)	Discussion: Recent Advances in Vision-Language Generative Models
Lecture 17	Nov 18 (Mon)	Discussion: Recent Advances in Computing Efficiency
Project	Nov 20 (Wed)	Milestone due
Lecture 18	Nov 20 (Wed)	Guest Lecture: Advanced Vision Models in Industry
	Nov 25-29	Thanksgiving break
Lecture 19	Dec 2 (Mon)	Guest Lecture: Responsible and Ethical Deployment of Advanced Vision Models
Lecture 20	Dec 6 (Fri)	Course Conclusion: The Next Frontiers of Computer Vision and Biomedicine

Course Logistics

Lectures: MW 10:30-11:50am

- Most days: Alway Building, Room M114
- Except on 10/9 in Alway Building, Room M108
- And on 10/21 and 11/4 at CCSR Building Room 4205

Course materials will be hosted on website: <http://biods276.stanford.edu>

Teaching team

Instructor



Serena Yeung
syyeung@stanford.edu
OH: 10-12am Tuesdays,
except Oct 1 at 9:30-10:30
Location: Packard 361

Assistant Instructor



Xiaohan Wang
xhanwang@stanford.edu
OH: 4-6pm Thursdays
Location: Packard 339

Teaching Assistant



James Burgess
jmhb@stanford.edu
OH: 4-6pm Mondays
Location: Packard 339
except 10/28 at Packard
364

Office hours will start week 2

Prerequisites

This course is considered an advanced course and students should be comfortable with deep learning and computer vision at the level of CS231N or BIODS220.

Ed Platform

- Will be used for **all** course communications.
- All students should be automatically added, please check
 - Periodically added from axess, may be a delay if you recently enrolled
 - Email staff mailing list biods276-aut2024-staff@lists.stanford.edu if you do not yet have access
- If it is a personal matter (e.g., OAE), please make a private post to the teaching team or instructor.
- Note: we will not be using canvas in this course. Communications will be through Ed, and grading will be through gradescope.

Grading

- Sign up for Gradescope through the “logistics tab”
- Breakdown:
 - Assignment 1: 20%
 - Assignment 2: 20%
 - Discussion Presentation: 15%
 - Course project: 40%
 - Class participation: 5%

Assignments

- There will be two homework assignments in the class, each worth 20% of the final grade.
 - For each assignment, students will write a report in response to a prompt chosen from a course-provided list, e.g., read and compare / contrast two related papers.
 - Topics will be related to and reinforce the material covered in class.
 - A1 (Representation learners): Out Wed 10/3, due Wed 10/16.
 - A2 (Generative models): Out Wed 10/30, due Wed 11/13.
-
- In this class, all deadlines refer to 11:59pm PST on the stated day.
 - Collaboration policy: All assignments must be completed individually. Study groups are allowed, but each student must produce an independent assignment and write the names of their group members on the assignment.

Discussion Presentation

- There will be a number of class sessions dedicated to student presentation and discussion of recent advances related to the course topics.
- Each student will sign up to present a recent paper (from a course-provided list) during these class sessions, as well as to ask questions in other sessions.
- More details will be provided next week. This will be worth 15% of the final grade.

Project

- The course project will be worth 40% of the final grade. The goal is to gain hands-on experience interacting with some of the large vision and vision-language models discussed in class. The project will include a project proposal, milestone report, and final presentation and report.
- There will be several options of project type to choose from:
 - **Red teaming analysis.** Conduct a "red teaming" analysis of at least two large vision or vision-language models, considering a biomedical use case or motivation.
 - **Implement an agentic system.** Implement an agentic system that incorporates large vision or vision-language models to address a specific biomedical problem.
 - **Explore a technical innovation.** Explore and experimentally assess the effectiveness of a novel technical approach or innovation to enhance the capabilities of existing vision or vision-language models, considering a biomedical use case or motivation.

Project (cont.)

- The intent of all project types is to gain deeper understanding of the capabilities, limitations, and opportunities for future advancement of these models.
- Can work in groups of 1-2. (Grades will be calibrated by group size)
- Will release detailed project guidelines and suggestions, and discuss in lecture next week.
- Graded components:
 - Proposal: Due Wed 10/23.
 - Milestone: Due Wed 11/20.
 - Project advising session: after the milestone, details TBD.
 - Final report due: Wed 12/11.
 - Final project presentation: In person, during the final exam period for this course (Mon 12/9, 3:30-6:30pm)

Class participation

- In-class attendance is an important part of the learning experience for this advanced, discussion-centric course
- Students are expected to attend at least 15 class sessions during the course (out of 20 in the schedule). Attendance will be recorded through a sign-in sheet at each class, and satisfactory attendance will comprise 5% of the course grade.
- Any additional absence beyond this will need to be under extenuating circumstances (such as a medical absence justified by a doctor's note), and depending on the circumstances students may be asked to write a one-page make-up report on a provided prompt.

Late days

- Can be used on A1, A2, project proposal, project milestone report.
- Cannot be used on project final presentation, or final project report.
- 3 late days total, 2 max for any assignment.
- Grades will be deducted by 25% for each additional late day.

Next time

- Neural networks for image data: From CNNs to Transformers