Syllabus

Books

1. TIF - Foundations of Computer Vision, Antonio Torralba, Phillip Isola and William T. Freeman. This book is free and goes through the very latest applications of deep learning for computer vision such as diffusion models and others.

2. BISHOP - Deep Learning - Foundations and Concepts, by C Bishop and H Bishop. This book can be accessed and viewed online from the book's website.

3. SZELINSKI - Computer Vision: Algorithms and Applications, 2nd Edition. This book is free to download for personal use. It may serve as an alternative to the TIF book for some of the topics covered in this course.

Planned Schedule

Lecture	Title	Details
1	**Introduction and the foundations of vision**	We start with an introduction to Computer Vision for the general application area of agents with egomotion. Throughout this course we will assume that a monocular or stereo camera is mounted on an agent that can in general move in a 3D environment and focus on its perception system assuming only the presence of camera sensing. In this lecture we explain we review prerequisites on programming (Python) as well as linear algebra, probability theory and basics on how cameras work. With the help of the TAs & other tutorial videos, we also ensure that students have setup a programming environment necessary for the projects and assignments of the course. Reading: selected pages from the course web site. **Reading:** Selected pages from the course website.
Part I: Detection and Segmentation		
2	**Introduction to Statistical Learning**	The computer vision system is now dissected into its parts with the very first part being featurization. We introduce the end to end prediction problem using simpler learning architectures and subsequently fully connected neural architectures. Our focus here is to understand how prediction can be engineered by applying the maximum likelihood optimization principle in the regression and classification tasks that we meet throughout this course. **Reading:** BISHOP Chapter 4, Chapter 5
3	**Dense Neural Networks**	We now apply the principles of statistical learning in the design of dense neural networks that use the cross entropy loss function. These dense layers are everywhere as building blocks from object detection to transformers and we will learn how to train and regularize them. **Reading:** BISHOP Chapter 6.
4	**Convolutional Neural Networks (CNNs)**	We then introduce CNNs with their innate ability to efficiently learn spatial hierarchies of features through backpropagation. We treat

Lecture	Title	Details
		simple tasks such as image classification and then quickly dive into architectures that were particularly made for image featurization such as Residual Networks (ResNets) explaining why they are so popular especially for real-time perception . **Reading:** BISHOP Chapter 10.
5	**Object Detection**	As a first task in scene understanding, we now design object detectors, initially from CNNs, that identify and locate objects of interest. We treat two main architectures: `YOLO` and `Faster R-CNN`. `YOLO` is known for its speed and efficiency while `Faster R-CNN` focusing on higher detection accuracy often resulting in better precision and recall in complex scenes. **Reading:** SZELINSKI Chapter 6.
6	**Semantic Segmentation**	Many computer vision applications require far finer granularity than a bounding box around the object(s). Here we expand on the task of CNN-based object detectors to include heads that are able to label the specific pixels of the object that occupy the scene as well as expand on panoptic segmentation that labels everything in the scene **Reading:** SZELINSKI Chapter 6.
7	**Vision Transformers (ViT)**	At this point, we introduce transformer-based architectures that will be the basis for more advanced tasks later on in this course. We focus on Vision Transformers (ViT) that leverage a self-attention mechanism to model global dependencies within an image treating the image as a sequence of patches. We understand that comparatively to CNNs, ViT-models suffer from increased latency inhibiting real-time applications relative to CNN counterparts while for non real-time setting they improve performance on tasks requiring an understanding of the whole image context. **Reading:** BISHOP Chapter 12 and TIF Chapter 26.
8	**Object Tracking**	In this lecture we focus entirely on video streams and on the requirements of many computer video applications such as video

Lecture	Title	Details
		surveillance to track objects within the geometrical boundaries of a single camera. We look at various architectures, that can correct the reflexive nature of earlier build CNN and ViT detectors to track an object despite challenges such as occlusion, motion blur, and changes in appearance. **Reading:** Course web site and notes.
Part II: Vision Language Models (VLMs)		
9	**Contrastive Learning**	In Part I we developed models and systems that can perceive the environment purely from visual information. In this part we explore the interplay between vision and language pretraining that has dramatically improved our ability to reason when presented with multiple modalities simultaneously. Discriminative (contrastive) representation learning principles are coupled with OpenAI's CLIP's ability to relate images and text is a key starting point in this exploration. **Reading**: CLIP paper and TIF Chapter 51.
10	**From Assignment to Generation**	We build on the *retrieval* abilities of CLIP and add the *generative* abilities of text decoders and query transformers to enable tasks such as image captioning, Visual Question Answering (VQA) and others. We treat the Bootstrapped Language-Image Pretraining (BLIP/2) model as a baseline architecture before transitioning to LLaVA a richer category of models that enable open ended tasks with prompting. **Reading:** Course notes on BLIP/2.
11	**Prompted Vision Models**	In this conclussive lecture on VLMs we revisit pure visual tasks such as segmentation via Meta's Segment Anything Model (SAM) and present them as *workers* receiving *multimodal* prompts from VLMs that act as_planners_. **Reading:** Course notes based on SAM paper.
Part III: Generative Vision Models		

Lecture	Title	Details
12	**Representing Scenes as Neural Radiance Fields**	In the last few years we have seen significant advances in the ability of models to generate realistic scenes. We are now at a point in this course where we have all the tools available to us to study the first generative method that, amazingly, creates 3D scenes from a set of 2D images. We cover concepts such as volume rendering and others, paired with fully connected neural networks to achieve photorealistic generation of 3D scenes. **Reading:** Course notes based on NeRF paper.
13	**Diffusion Models and DALL-E**	In this final lecture, we look at diffusion, inspired from thermodynamics, as a general modeling approach, called physics-inspired learning, that does conditional image generation given a textual description. In essence we are trying to reverse the image captioning task we have treated earlier and in this exciting last lecture we will combine the textual representation learning of the CLIP model we have seen earlier with a conditional diffusion process to create photorealistic images guided by our prompts. **Reading:** Course notes on DALL-E/2 and Stable Diffusion.