

# Deep Learning: Modern Artificial Vision

Jaume Ivars Grimalt

March 2025



# Contents

Acknowledgements	iii
Preface	v
<b>I Foundations of Modern Computer Vision</b>	<b>1</b>
<b>1 What is Computer Vision?</b>	<b>3</b>
1.1 Basic Concepts . . . . .	3
1.2 The Role of Deep Learning in Computer Vision . . . . .	3
<b>2 Convolutional Neural Networks (CNNs)</b>	<b>5</b>
2.1 The Basics of Convolution . . . . .	5
2.2 Pooling Layers . . . . .	5
2.3 Wide Convolutions and Point-wise Convolutions . . . . .	5
2.4 Activation Functions . . . . .	5
<b>3 Optimization in Deep Learning</b>	<b>7</b>
3.1 Gradient Descent and Backpropagation . . . . .	7
3.2 Optimizers . . . . .	7
3.3 Challenges in Optimization . . . . .	7
<b>4 Building Blocks of Modern CNNs</b>	<b>9</b>
4.1 Batch Normalization . . . . .	9
4.2 UpSampling and Transposed Convolutions . . . . .	9
4.3 Gradient Flow Through Layers . . . . .	9
<b>II Core Architectures and Building Blocks</b>	<b>11</b>
<b>5 Backbone Networks</b>	<b>13</b>
5.1 CNN-based Backbones . . . . .	13

---

5.1.1	ResNet . . . . .	13
5.1.2	ResNeXt . . . . .	13
5.1.3	EfficientNet . . . . .	13
5.1.4	MobileNet . . . . .	13
5.2	Introduction to Transformers in Vision . . . . .	13
5.2.1	Vision Transformer . . . . .	13
5.3	Advanced Transformer Backbones . . . . .	13
5.3.1	Swin Transformer . . . . .	13
<b>6</b>	<b>Neck Networks</b>	<b>15</b>
6.1	FPNs . . . . .	15
6.2	PANs . . . . .	15
6.3	BiFPN . . . . .	15
<b>7</b>	<b>Head Networks</b>	<b>17</b>
7.1	TOOD . . . . .	17
<b>8</b>	<b>Attention Mechanisms</b>	<b>19</b>
8.1	Attention with CNNs . . . . .	19
8.1.1	Deformable Convolutions . . . . .	19
8.2	Attention with ViTs . . . . .	19
<b>III</b>	<b>Core Tasks in Computer Vision</b>	<b>21</b>
<b>9</b>	<b>Image Classification</b>	<b>23</b>
<b>10</b>	<b>Object Detection</b>	<b>25</b>
10.1	Key Models . . . . .	25
10.1.1	YOLO . . . . .	25
10.1.2	R-CNN . . . . .	25
10.1.3	Faster R-CNN . . . . .	25
10.1.4	DETR . . . . .	25
10.1.5	RT-DETR . . . . .	25
10.1.6	RetinaNet . . . . .	25
10.2	The State of the Art Models . . . . .	25
10.2.1	YOLOv12 . . . . .	25
10.2.2	CO-DETR . . . . .	25
<b>11</b>	<b>Semantic Segmentation</b>	<b>27</b>
11.1	Key Models . . . . .	27
11.1.1	U-Net . . . . .	27

11.1.2	U2-Net . . . . .	27
11.1.3	SegFormer . . . . .	27
11.2	The State of the Art Models . . . . .	27
<b>12</b>	<b>Density Map Estimation</b>	<b>29</b>
12.1	Key Models . . . . .	29
12.1.1	CSRNet . . . . .	29
12.1.2	Cascaded CSRNet . . . . .	29
12.2	The State of the Art Models . . . . .	29
<b>IV</b>	<b>Production Deployment</b>	<b>31</b>
<b>13</b>	<b>Model Optimization</b>	<b>33</b>
13.1	Model Compression . . . . .	34
13.1.1	Quantization . . . . .	34
13.1.2	Pruning . . . . .	34
13.1.3	Knowledge Distillation . . . . .	34
13.1.4	Neural Architecture Search . . . . .	34
13.2	Model Acceleration . . . . .	34
13.2.1	TensorRT . . . . .	34
13.2.2	ONNX Runtime . . . . .	34
13.2.3	OpenVINO . . . . .	34
13.2.4	TensorFlow Lite . . . . .	34
13.3	Model Deployment . . . . .	34
13.3.1	TensorFlow Serving . . . . .	34
13.3.2	TorchServe . . . . .	34
13.3.3	NVIDIA Triton Inference Server . . . . .	34
13.3.4	MLflow . . . . .	34
13.3.5	Kubeflow . . . . .	34
13.4	Model Monitoring . . . . .	34
13.4.1	Prometheus . . . . .	34
13.4.2	Grafana . . . . .	34
13.4.3	Seldon . . . . .	34
13.4.4	Evidently . . . . .	34
13.5	Model Versioning . . . . .	34
13.5.1	DVC . . . . .	34
13.5.2	MLflow . . . . .	34
13.5.3	Weights & Biases . . . . .	34

## **V Explainability and Interpretability 35**

### **14 Explainability in Computer Vision 37**

14.1 Introduction to Explainability . . . . .	38
14.1.1 What is Explainability? . . . . .	38
14.1.2 Why is Explainability Important? . . . . .	38
14.1.3 Types of Explainability . . . . .	38
14.1.4 Challenges in Explainability . . . . .	38
14.2 Methods for Explainability . . . . .	38
14.2.1 Saliency Maps . . . . .	38
14.2.2 Grad-CAM . . . . .	38
14.2.3 Integrated Gradients . . . . .	38
14.2.4 LIME . . . . .	38
14.2.5 SHAP . . . . .	38
14.3 Interpretable Models . . . . .	38
14.3.1 Decision Trees . . . . .	38
14.3.2 Rule-Based Models . . . . .	38
14.3.3 Linear Models . . . . .	38
14.3.4 Prototype-Based Models . . . . .	38
14.4 Evaluating Explainability . . . . .	38
14.4.1 Quantitative Evaluation . . . . .	38
14.4.2 Qualitative Evaluation . . . . .	38
14.4.3 User Studies . . . . .	38
14.5 Applications of Explainability . . . . .	38
14.5.1 Medical Imaging . . . . .	38
14.5.2 Autonomous Vehicles . . . . .	38
14.5.3 Security and Privacy . . . . .	38
14.5.4 Fairness and Bias . . . . .	38



# Acknowledgements

A special thanks to all those who contributed to this work...





# Preface

This is the preface of the document...

# I Foundations of Modern Computer Vision



# Chapter 1

## What is Computer Vision?

### 1.1 Basic Concepts

Computer vision is a field of artificial intelligence that focuses on enabling machines to interpret and understand visual information from the world, similar to how humans do. It involves the development of algorithms and models that can analyze images and videos, extract meaningful features, and make decisions based on visual data. Computer vision has applications in various domains, including autonomous vehicles, medical imaging, surveillance, robotics, and augmented reality.

### 1.2 The Role of Deep Learning in Computer Vision



## Chapter 2

# Convolutional Neural Networks (CNNs)

### 2.1 The Basics of Convolution

### 2.2 Pooling Layers

### 2.3 Wide Convolutions and Point-wise Convolutions

### 2.4 Activation Functions





# Chapter 3

## Optimization in Deep Learning

3.1 Gradient Descent and Backpropagation

3.2 Optimizers

3.3 Challenges in Optimization



# Chapter 4

## Building Blocks of Modern CNNs

### 4.1 Batch Normalization

### 4.2 UpSampling and Transposed Convolutions

### 4.3 Gradient Flow Through Layers



## **II   Core Architectures and Building Blocks**



# Chapter 5

## Backbone Networks

### 5.1 CNN-based Backbones

#### 5.1.1 ResNet

#### 5.1.2 ResNeXt

#### 5.1.3 EfficientNet

#### 5.1.4 MobileNet

### 5.2 Introduction to Transformers in Vision

#### 5.2.1 Vision Transformer

### 5.3 Advanced Transformer Backbones

#### 5.3.1 Swin Transformer





# Chapter 6

## Neck Networks

### 6.1 FPNs

### 6.2 PANs

### 6.3 BiFPN



# Chapter 7

## Head Networks

### 7.1 TOOD



# Chapter 8

## Attention Mechanisms

### 8.1 Attention with CNNs

#### 8.1.1 Deformable Convolutions

### 8.2 Attention with ViTs



### **III    Core Tasks in Computer Vision**





## Chapter 9

# Image Classification



# Chapter 10

## Object Detection

### 10.1 Key Models

#### 10.1.1 YOLO

#### 10.1.2 R-CNN

#### 10.1.3 Faster R-CNN

#### 10.1.4 DETR

#### 10.1.5 RT-DETR

#### 10.1.6 RetinaNet

### 10.2 The State of the Art Models

#### 10.2.1 YOLOv12

#### 10.2.2 CO-DETR



# Chapter 11

## Semantic Segmentation

### 11.1 Key Models

#### 11.1.1 U-Net

#### 11.1.2 U2-Net

#### 11.1.3 SegFormer

### 11.2 The State of the Art Models



# Chapter 12

## Density Map Estimation

### 12.1 Key Models

#### 12.1.1 CSRNet

#### 12.1.2 Cascaded CSRNet

### 12.2 The State of the Art Models





## IV Production Deployment





# Chapter 13

## Model Optimization

### 13.1 Model Compression

#### 13.1.1 Quantization

#### 13.1.2 Pruning

#### 13.1.3 Knowledge Distillation

#### 13.1.4 Neural Architecture Search

### 13.2 Model Acceleration

#### 13.2.1 TensorRT

#### 13.2.2 ONNX Runtime

#### 13.2.3 OpenVINO

#### 13.2.4 TensorFlow Lite

### 13.3 Model Deployment

#### 13.3.1 TensorFlow Serving

#### 13.3.2 TorchServe

#### 13.3.3 NVIDIA Triton Inference Server

#### 13.3.4 MLflow

#### 13.3.5 KubeFlow

### 13.4 Model Monitoring

#### 13.4.1 Prometheus

#### 13.4.2 Grafana

## **V Explainability and Interpretability**





# Chapter 14

## Explainability in Computer Vision

### 14.1 Introduction to Explainability

#### 14.1.1 What is Explainability?

#### 14.1.2 Why is Explainability Important?

#### 14.1.3 Types of Explainability

#### 14.1.4 Challenges in Explainability

### 14.2 Methods for Explainability

#### 14.2.1 Saliency Maps

#### 14.2.2 Grad-CAM

#### 14.2.3 Integrated Gradients

#### 14.2.4 LIME

#### 14.2.5 SHAP

### 14.3 Interpretable Models

#### 14.3.1 Decision Trees

#### 14.3.2 Rule-Based Models

#### 14.3.3 Linear Models

#### 14.3.4 Prototype-Based Models

### 14.4 Evaluating Explainability

#### 14.4.1 Quantitative Evaluation