

Part 1

→ Hubel & Wiesel (cats :))

→ orientation tuning

→ retinotopic organisation

fMRI ← where

MEG, EEG ← when

* Kobatake & Tanaka (1994)
 * Desimone et al. (1984)
 * Grill-Spector & Weiner (2014)

! Alexia part appears when you learn to read and it's always in the same place of the brain.

→ hierarchical theory of object recognition



Part 2

How human vision influenced computer vision?

→ Gabor filters

→ computer vision meets deep learning

→ solve invariant object recognition

→ ImageNet & AlexNet

Lecun et al, Nature (2015)

O'Reilly (2020) (?)

DiCarlo & Cox (2007)

* The Words I See
Fei-Fei Li

* Krizhevsky, Suts... (2012)

* Fukushima (1980)
Neocognitron

* JeNet, JeCun

→ poloclub GitHub CNN explainer

* Lindsay (2020)
computational neural networks

* Zeiler & Fergus (2014)

* Guckl & von Gennet (2015)

Part 3

- Do CNN networks predict brain activity? (features)
- HCNN layers (?)
- Grouping similarity
- Networks seem to mimic human brain
- Brain-Score
- Algo nauts
- Neuroconnectionist research cycle
- object recognition & scene processing

Lindsay (2020)

Jenkins et al. (2014)

Kriegeskorte (2014)

Kriegeskorte & Mur (2012)

King & Green et al (2018)

Doenig et al. (2022)

Green, Silson & Baker (2017)

Part 4

How computer vision & human vision (still) different?

- retinal image vs cortical map.
- transformers are better than neural networks
- Neuroscience & computer vision work together, they feed each other
- global statistics extraction, feedback/recurrent connections, peripheral sampling, ecological training, etc. ? ? ?

* Da Costa et. al. (2021)

* Cheung, Wells ...

* Müller, Scholte & Green (2020)

* Geirhos et. al. (2018)

* Geirhos et. al. (2019)

* Papello, Moques et. al
Simulating a Primary ...

→ An image worth 16x16 words

* Tuli, Dasgupta, ... Are convolutional NNs more like com-vis.?

* Connell et. al (2024)

Andrew French - Day 1 Lecture 2

→ "no learning"?

- * Classical tracking
- motion detection & tracking
- individual vs. multiple targets

→ pixel-level motion
→ optic flow
→ background models } detect movement
not tracking ::

- * if we can model the motion, we can predict objects' future locations.
- HCI, robotics, surveillance, medicine...

flow-field ↔ quiver plot

→ motion difference & background subtraction

two basic approaches
in motion detection

motion detection

- multiple object tracking -

- * Markov Chain Monte Carlo tracking
- * Metropolis-Hastings' algorithm
 - chain of predictions

→ Khan 2. et.al. (2004) An MCMC...

- * social extensions
 - behaviours
 - sharing motion information

- tracking -

* uncertainty

- * we want to be able to PREDICT where our target will be, and UPDATE our guess with a measurement.

* Kalman Filter

* Particle filters

→ real tracking is often multimodal

→ Isard & Blake (1998) Condensation

* appearance model(?)

* contour tracking

* mixed-state condensation

→ transition probability

* behavioural recognition ?

* the curse of dimensionality

still relevant in the
deep-learning era

Ego-centric Vision - Making Sense of the First-Person Perspective

* Steve Mann, WearCam, WearComp
applications in Personal Imaging

* Wearable devices

* Not limited to humans

↳ dogs, robots, shopping cards

* Blue Sky Outcomes

↳ robotics to understand surroundings

↳ personal assistants

- remind you what you forgot
- help you fix things
- remind you a recipe's next step etc.

↳ healthcare for assisting people

↳ ethic & privacy is an issue! !

↳ Hololens, Google Glass, etc.?

Ego4D dataset

GTEA, BEOLD, GTEA Gaze, UT Ego,
ADL, EGTEA Gaze +, EPIC-Kitchen-117
Chades-Ego, EPIC-Kitchen-100

↳ combining every dataset

Ego4D

Past — Present — Future

- Episodic memory
- Hand object interaction
- Forecasting
- Audio visual diarization
- Social interaction

HD-EPIC

Digital Twin of kitchen videos

- narration, fine grained annot
- audio annotation, object detection
- object - fixture assignment

- **Blind-Language**: Llona 1.2, Gemini Pro
- **Video-Language**: Video Llona 2, LangVA, Gemini Pro

Ego-Centric Tasks

- video understanding tasks
- ↳ action recognition, VQA

- captioning, retrieval, grounding
- ↳ audio localization, activity & object recognition

→ EgoVis workshop

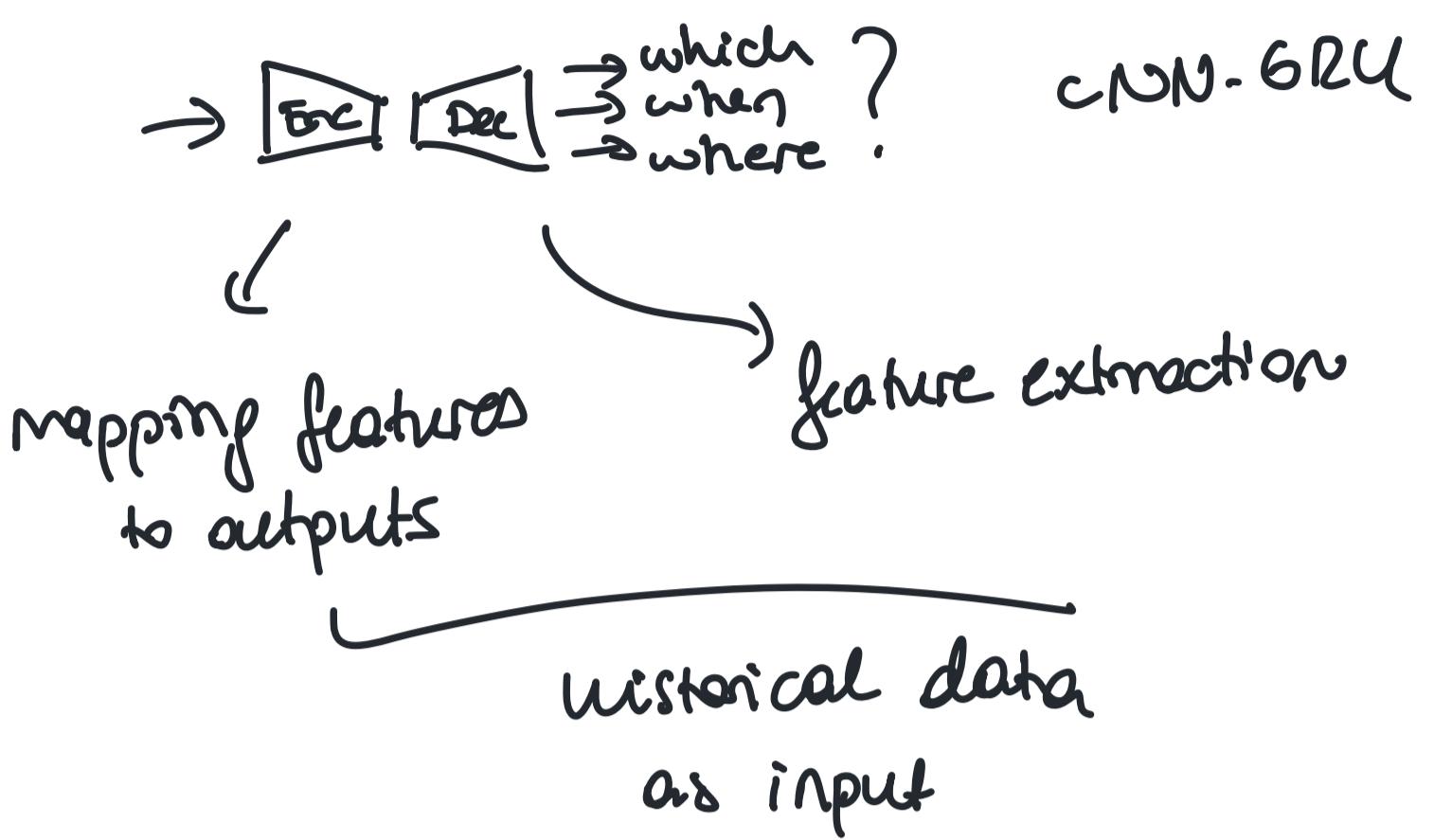
HierVL : Learning Hierarchical Video-Language Embeddings

- EgoVLP - previous study, not good (?)
- Assembly101 dataset (2022)
- ↳ assembly hands
- ↳ SVEgoNet : single view ego net

LEGO: Learning Egocentric Action Frame Generation via Visual Instruction Tuning

* CV + DL

which? → when? → where?
trajectory tensor



→ used tensors instead of vectors
to convey the uncertainty

* FC, CNN, GRU, TAN

* binary cross-entropy loss

Prof. Victor Sanchez - Warwick

Day 2 Lecture 2

CLIP: Contrastive Language-Image Pretraining
for Video Analytics

* Deepfakes, IAPR head of:

* SIPIlab,

• multimedia forensics & biometrics
↳ platform provenance

Project: Multicamera Trajectory Forecasting
tracking via person re-identification

Shyles, Guha, Sanchez (2021) Pattern Analysis...

Nootker, Sanchez (2025) WACV

TinyFaces project

MTCNN + IFS

Leyva, Sanchez, Li (2018)

Leyva, Shen, Bahadur ... (2025)
(Automatic Face & Gesture Recogn.)
IEEE

TinyFaces IFS detector

↳ Fisher Vector

* IFS to FVs → reduces the # of operations

Kulshreshtha & Guha (2018)
(ICIP)

M-EVA dataset

Multiview extended video with activities

Qi, Tan, Yao, ... (2023) - Yolo5Face

Yu, Huang ... (2024) - YoloFacev2

TinyFaces is still a challenge to be solved!

Recall: CLIP is proposed to be used in downstream tasks
→ WebImageText (WIT)

CLIP for segmentation

Lüddecke, Ecker (2022) → CUF

DACE: Distance-Aware Cross-Entropy Loss

HuggingFace: Spaces/Yiming-M / CLIP-EBC (?)

Crowd Counting

→ rely on annotated 2D coordinates of individuals' head centres in images.

Ma, Sanchez, Guha (2022) → CLIP

Ma, Sanchez, Guha (2025) → ICME (CLIP-EBC)

→ Reform crowd counting as a classification task



they introduced this idea to CLIP.

* zero-shot learning: can we facilitate?

↳ natural language supervision

* contrastive learning

Radford, Kim, Hallacy ... learning transferable visual models from natural language supervision

* symmetric crossentropy loss

Attention is all you need

Fine-tune of CLIP → ResNet50: attention pooling
Vision Transformer

From Paragraphs to Pixels: LLMs in Video Understanding

@andrewjohngilbert.github.io

* Video understanding

- ↳ long range temporal dependencies
- ↳ ambiguous/overlapping events
- :
lots of challenges in this area.

YouTube: WhoDunnit? It is easy to miss
stuff you're not
looking for.

* Video Understanding with Large Language Models: A Survey, Yuanlong Tang et al.

* Unsupervised Learning of Visual Representations Using Videos, ICCV 2015

* Shuffle Learn ECCV (2016)

* Self-Supervised Video Rep CVPR 2017

* Arrow of Time, CVPR 2018 (?)

* ECCV 2018 - Tracking Inference by Colorizing Videos

* VideoMAE, NeurIPS 2022

* MOFO:, NeurIPS 2023 workshop

* CVPR 2023 - Learning Video ..

* Vid2Seq CVPR 2023

* VAST NeurIPS 2023, Sijian Chen

* FILS : Self-supervised Video Feature Prediction...

* Fine-Grained audible video description

* SWIN-BiT (?)

* AutoAD : CVF 2023

→ ? Donte AD ?

gap between pixels & n1 reasoning

Surveillance

Healthcare

Autonomous Vehicles

Entertainment

Impact &
ApplicationAI Storytelling

Art in Storytelling.

* Visual Agents in Software

- ↳ unreal engine
- or Game Engines mostly

} Visual Interface
Agent.

* MAGMA : Multimodal Agentic Foundation

Tang et al. 2025

* UI-TADS, 2025

Audio-Visual Fusion (?)

* Romanovna... "OWL..." CVF 2023

* Egocentric audio-visual object localization, CVF 2023

* Self-supervised moving vehicle ...

* Epic-fusion : Audio visual temporal binding for egocentric action recognition.

→ Dataset: Youtube8MTHUMOS

Local (?)
Learnable Getting
Cross Attention

* Multi-Resolution Audio-Visual Feature Fusion...

* DEL: Dense Event Localisation...? 2025

→ Action Quality Assessment

→ Scale to this.

ECCV 2022

with temporal
padding transformer

* learning to score Olympic Events

Day 2 Lecture 6

Representation learning

- * bad vs good features
- Wei Koll et.al. Concept Bottleneck Models, ICML 2020

* success depends on data representation

↳ what is good representation?

- compact (minimal)
- explanatory (sufficient)
- disentangled (independent factors)
- interpretable
- make subsequent problem easier

Eli Cole, Caltech (slide credits)

* Transfer learning (ImageNet)

- Bengio et al. Representation Learning: A Review and Perspectives

credit: Justin Johnson eecs498

Self-Supervised learning

"Most of human and animal learning ... " ~Yann LeCun
(On true AI)

* how to learn features from the unlabelled data?

* Papers with Code: Self Supervised Image Classification

* Effective methods for learning representations.

- Gidaris et.al. Unsupervised Repr. Learning... ICLR 201

- Weng & Kim, Self-Supervised learning, NeurIPS 2021

Self-prediction

SimCLR (2018)

- Chen et.al. A Simple Framework for Contrastive learning ...

• Github: google-research/simclr

@sthalles.github.io/a-few-words-on-representation-learning

* Once trained, the representation can be transferred to other tasks.

Multi-modal Contrastive learning

- Radford et.al. Learning Transferable Visual Models from NL Supervision ICML 2021

* zero-shot classification

* there are limitations

↳ can require large batch size

↳ relies on "good" negative selection

↳ space of plausible augmentations can be dataset specific and must be defined in advance

* masked language models

↳ BERT masked language models

* vision transformers (ViT)

- Dosovitskiy et.al. An Image is worth 16x16 Words. ICLR

- He et.al. Masked Autoencoders Are Scalable Vision Learners CVPR 2022

- Zhang et.al. 2016, 2017

- Doersch et.al 2015

- Noroozi & Favaro, 2016 ?

- Noroozi et.al 2017

Recent advances

- * non-contrastive Siamese networks
- * BYOL, SimSiam
- * DINO
- Emerging Properties in Self-supervised Vision Transformers, ICCV 2021
- * DINOv2 builds on DINO & iBOT
 - DINOv2 : learning Robust Visual Features...

- limits of SSL
- * most SSL pretrained with ImageNet
- Cole et al. When Does Contrastive Visual Representation Learning Work? CVPR'22
- * dataset size matters
 - ↳ amount of unlabelled data for pretraining
 - ↳ amount of labelled data for supervised learning
- * iNAT
 - Von Henn et al Benchmarking... (?)
 - Zong, Andher... Self Supervised ... Vision @ UoE

- * Monty Hall problem.
- * Bayes' rule
 - How can we make it better if we don't know where it fails...
 - Understanding Deep Learning - Simon Pierce
 - * Bias-variance trade-off

Neill DF Campbell - UCL

Day 3 lecture 1

09.09

Uncertainty & Evaluation in Computer Vision

• Poggi ... On the uncertainty of self-supervised...

* why do we care about uncertainty? -

↳ ambiguity in task, in our models

↳ downstream decision making

· safe · robust · transparent

↳ improved performance

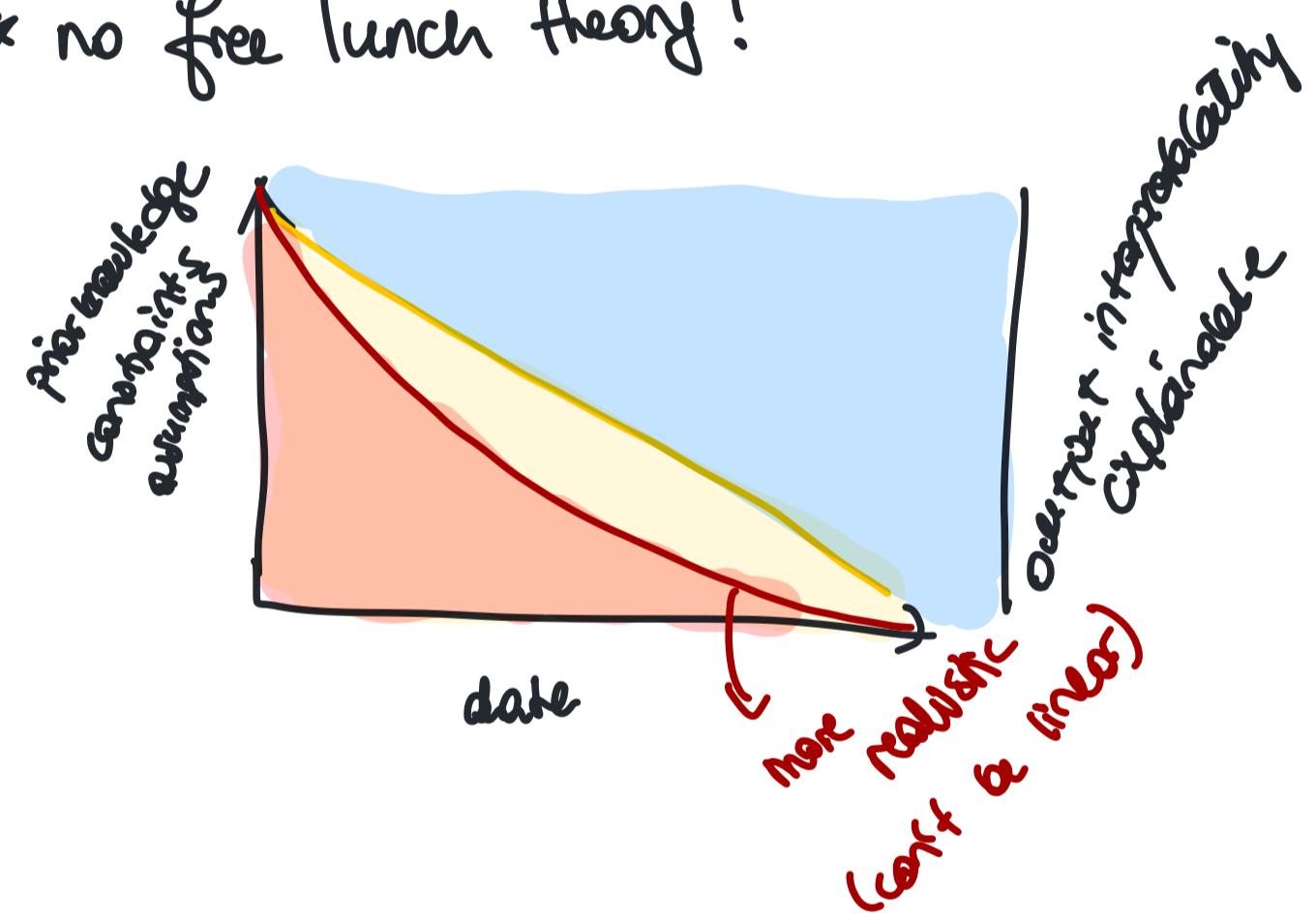
· data efficiency

· self-supervision

↳ evaluation & model selection!

* there are danger if we avoid probabilistic approach

* no free lunch theory!



Laplace: "the theory probability is the common sense reduced to calculus." (?)

* statistical significance

→ Youtube: Crimes against data

* Human36M dataset

* ablation studies

- Simple and Scalable predictive uncertainty estimation using deep ensembles
- on uncertainty of self-supervised monocular depth estimation
- learning structured gaussians to app. deep ensembles.

- * GAN, VAE, DDM

- * no semantic manipulation

↳ so structured generative models

- * generative model: probability dist, we can sample from

→ give latent variables Semantic meanings, by forcing the decoder

- AIR (ECCV NIPS'16) - SPACE (CVPR'20)

- SPAIR (Crawford AAAI'?)

- Anchukiewicz : ICMl 2020 ocl workshop

- * NeRF

- SoftRas (Liu 2019)

- NMR (Kato 2018)

- DIRT (Henderson 2018)

meshes

} differentiable rendering

- Mildenhall et. al., ECCV 2020

- GRAF, Schwartz et. al. NeurIPS 2020

- Henderson, Tsimenaki, Lampert - CVPR (2020)

- RenderDiffusion, CVPR 2023

- Denoising Diffusion via Image-Based Rendering, ICLR 2024

- PixelNeRF

- Sampling 3D Gaussian Scenes in Seconds with latent

Diffusion models · Henderson et al. 2024

- arxiv.org/abs/1612.03928

- iee: 9446880

- arxiv.org/abs/1805.04730

- arxiv.org/abs/2203.05469

- AlexNet (2012)

- VGGNet (2014)

- ResNets (2015)

- arxiv.org/abs/1512.03385

- Inception v3 (2016)

- NASNet (2018)

- arxiv.org/abs/1808.05377

- Zoph & Le (2017)

- arxiv.org/pdf/1707.07012

- * CIFAR-10

- arxiv.org/pdf/1802.01548.pdf

- arxiv.org/pdf/1807.11626.pdf

- arxiv.org/pdf/1808-05377.pdf

- * one-shot NAS

- * zero-shot NAS

Efficient models for Computer Vision

- arxiv.org/abs/1711.02613

- arxiv.org/abs/2211.10438v7

- * methods → structured

↳ pruning ↳ unstructured

↳ quantisation

↳ knowledge distillation

↳ neural architecture search

- * channel pruning → heuristically

- arxiv.org/abs/1802.03494

- * FP32 to INT8

- * symmetric quantisation

- * asymmetric quantisation

- * quantising weights
activations (post-training)

- * logits, softmax, log loss

- * distillation

↳ teacher & student

- arxiv.org/abs/1312.6184

#DARTS

- arxiv.org/pdf/1806.09055.pdf

#NASWOT

- arxiv.org/pdf/2006.04647.pdf

- arxiv.org/pdf/2301.11300.pdf

- arxiv.org/pdf/2010.11929v2.pdf

- * GPViT (ICLR 2023)

* specialist architectures

Mike Morgan - Uni. of Sheffield

* SLAM Primer

* SLAM/AI

* robotics industry.

Localisation is solved ... ?

- ↳ insufficiently accurate
- ↳ subject to interference
- ↳ requires infrastructure
- ↳ not a general solution

* Simultaneous Localisation & Mapping (SLAM)

• Bohg, Trin & Hug "Simultaneous localization and mapping" Part 1 & 2
IEEE robotics & automation (2006)

* Using vision to create world models (maps)

• NeRF-SLAM: Real-time Dense Monocular SLAM with Neural Radiance Fields

* indirect SLAM

• ORB SLAM3 → totally open source

* Where SLAM deployed today?

↳ AR/VR

• YouTube: Opteon 1 minute introduction

reverse-engineering biological brains, creating general purpose neuromorphic software

* SoTA autonomy ↗ create human algorithms
↳ how humans solve autonomy
* Natural autonomy ↗ extract nature algorithms
↳ how nature solve auto.
→ the struggle is to build adaptable machines

* Natural Intelligence

- ↳ identify valuable natural behaviours in nature
- ↳ separate algorithms from neural implementation
- ↳ map algorithms to standard compute h/w

* neuromorphic software for machines.

* intelligence isn't artificial, it's natural.

Armin Mustafa - Uni. of Surrey

4D Machine Perception of Complex Scenes

* Single & multi-view reconstructions

↳ * generative AI

* video understanding

* 4D Perception

Computer vision

Generative AI

NLP

Machine learning

* Personalised media

↳ AI for me project

* 4D Vision

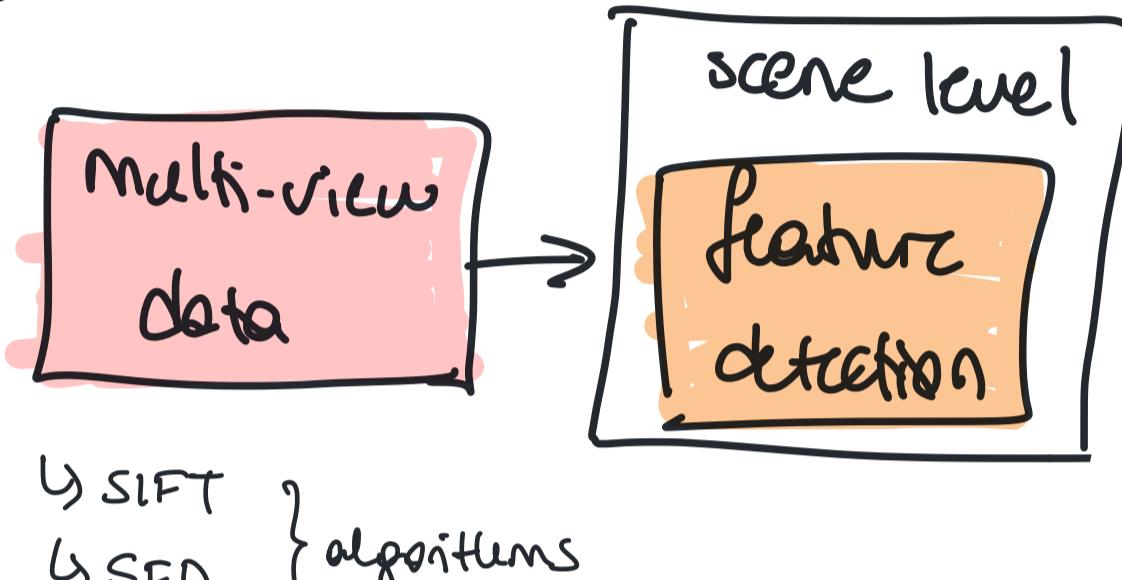
3D from monocular videos

↳ parametric reconstruction

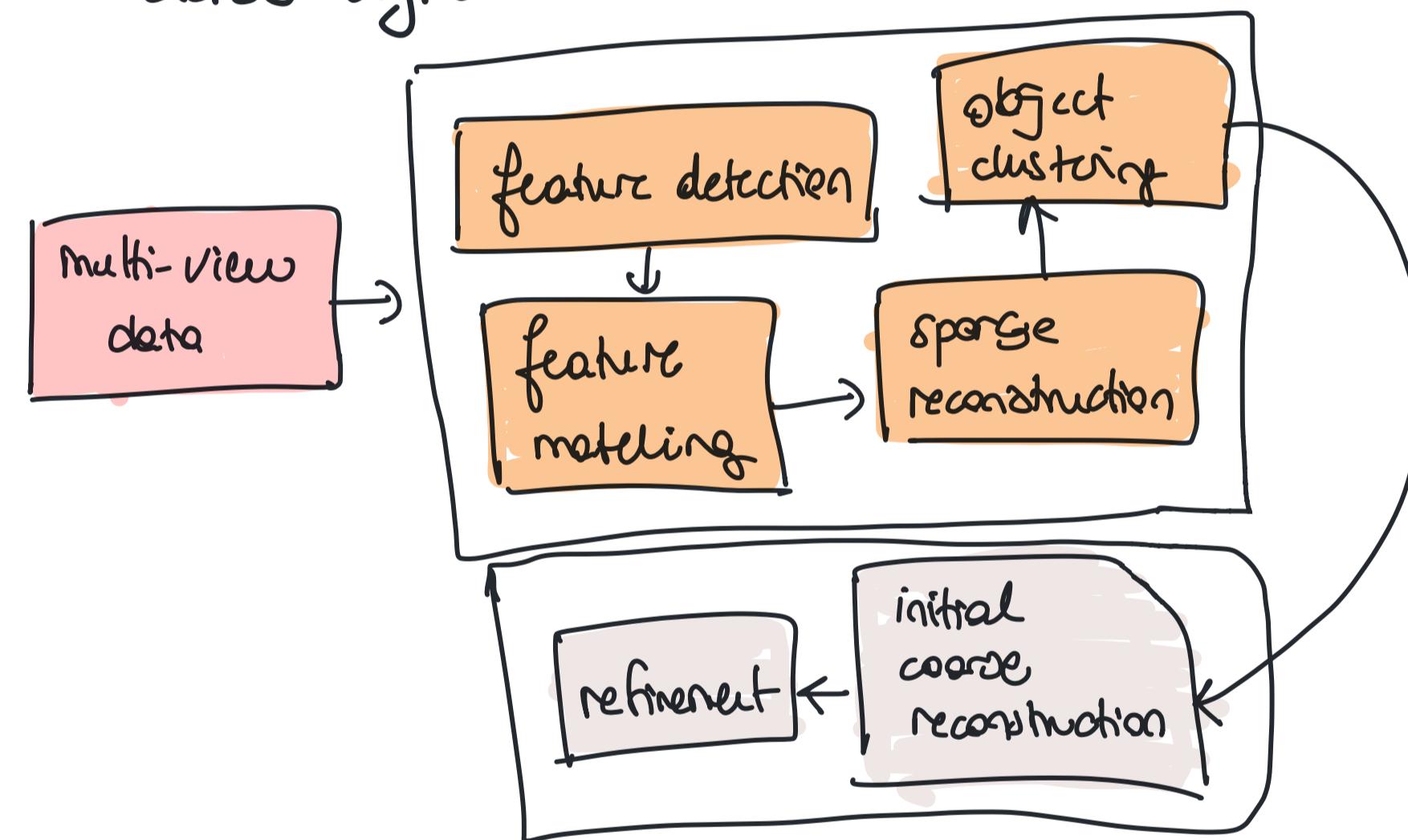
- depend on parametric 3D models
- estimate model parameters
- ? ?

(photos)

* general scene reconstruction



* dense dynamic 3D reconstruction



* machine learning based 3D reconstruction

* dense dynamic 3D reconstruction

↳ training data S2P2 dataset (?)

* 3D Video to 4D Models

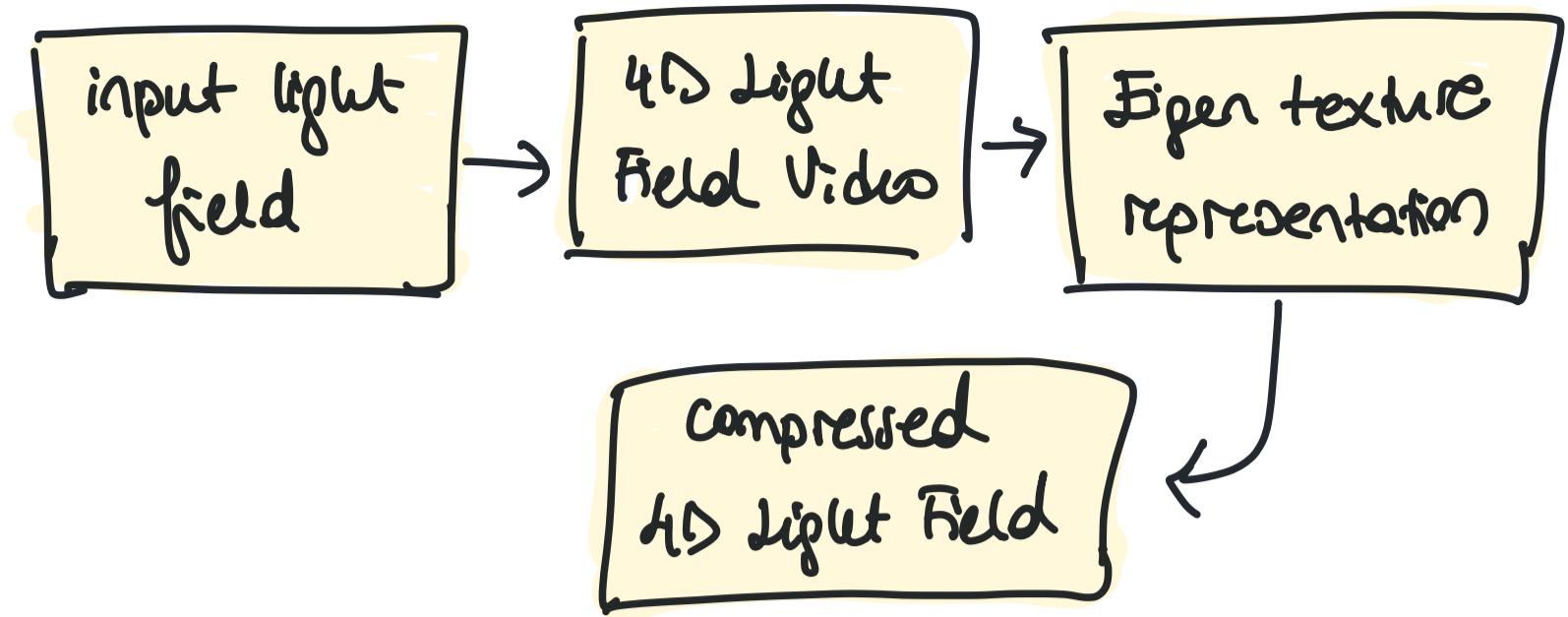
(photo)

• video: 4D light-field

* Kino & the Double World: movie :)

* epipolar plane information (EPI)

* Light-field Videos



why 4D Scene understanding?

→ estimate semantics, reconstruction and motion simultaneously

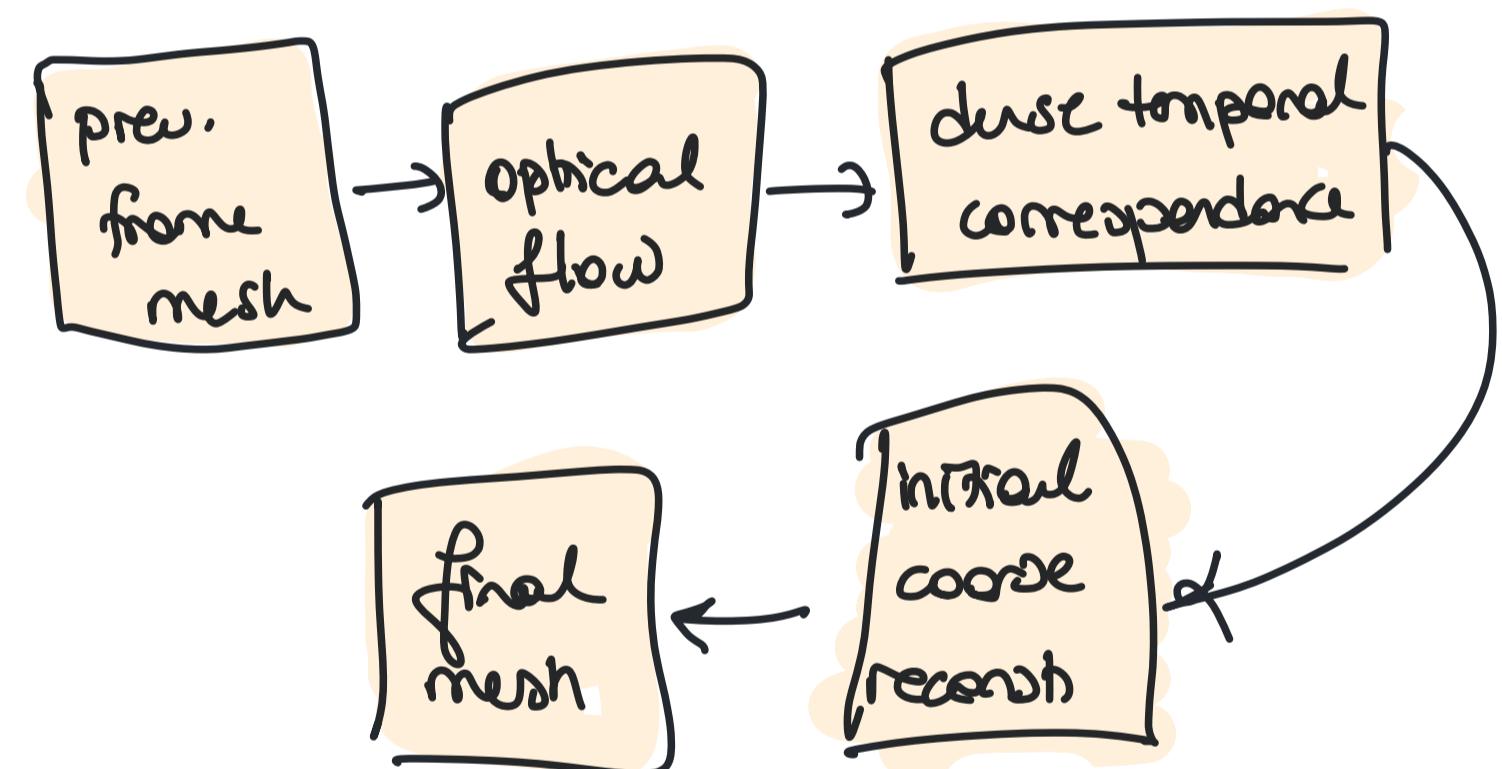
* sparse temporal tracks

↳ introduce temporal coherence

↳ appearance, motion & semantic inf.

↳ 3D human pose - part of sparse temporal tracks.

* temporal coherence



* geodesic star convexity (GSC) in optimisation

⇒ semantics + segmentation + depth inform

! could be applied to scene reconstruction methods!

VR, video creation, ...

⇒ single-image 3D human reconstruction

⇒ 3D Virtual Human Dataset

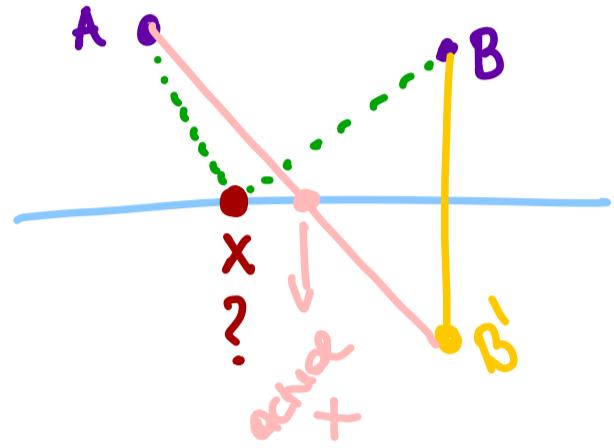
Tolga Birdal - ICL Day 4 lecture 3

Shape Shifters: Geometry, Topology, and Learning for and Beyond Computer Vision

* Representation has been known to be notorious problem in 3D vision

* 14th century Persian manuscript of Euclid's Elements

* Heron's solution



$$Z = \{x_i, y_i\}$$

inputs. labels

\hookrightarrow data

$$\hat{R}(f) = \frac{1}{n} \sum_{i=0}^n \ell(f(x_i), y_i)$$

$$R(f) = \mathbb{E} [\ell(f(x_i), y_i)]$$

loss function

integral over life

$|\hat{R}(f) - R(f)| \rightarrow$ generalisation error

* invariance & equivariance

@ maurice-weller.gitlab.io

* Fourier transforms

* Laplacian \rightarrow Pierre Laplace

\rightarrow what is not representable with graphs?

@ robertpepperell.com

"induce an indeterminate mental state"

\rightarrow microgenesis, Gestalt

! the whole \neq the sum of its parts

* como B3 · Dante

\rightarrow cognition is synergistic

Binod Bhattacharai - Uni. of Aberdeen

Day 4 lecture 4

Towards Trustworthy AI in medical Imaging

* ~80% of all healthcare data comes from medical images

* Modality 1: X-Ray (Radiography)

* Modality 2: Computed Tomography (CT)

* Modality 3: Magnetic Resonance Imaging (MRI)

* Modality 4: Ultrasound Imaging.
(proven to be very safe??)

\Rightarrow endoscopic imaging.

The Human Factor and Motivation of AI

\rightarrow most abnormalities are missed by human

\rightarrow 1% increase in detection in anomalies

3% decrease in cancer risk

* AI as an Assistive Tool in Medical Imaging

• 3D U-Net (Cicik et al. 2016)

• SegFormer3D, Perera et al. 2024

• Explainable Vision Transformers and Radiomics for COVID-19 Detection in Chest X-Rays

* Vision Language Models

CLIP, BLIP, Flamingo, LLaVA, GiT, MedCLIP, BioViL

* Medical Report generation

• Zhou et al. ICCV 2021

* Question answering

- Lee et. al. ICLR 2024

Challenges and Future Direction for Trustworthy AI

* European High Commission Expert Group on AI (2019), trustworthy ai:

- ↳ reliability / robustness
- ↳ privacy
- ↳ fairness
- ↳ explainability

* Federated learning for privacy protection

- Fed Average (McMahan et. al. AISTATS 2017)
- FedProx (Sahu et. al. MLSys 2020)
- Federated learning for Optical Coherence Tomography (Angain et. al. 2024)
- Li et. al. MLSys 2020
- Karimireddy et. al. PMLR 2020
- Poudel 2024 (?)
- Poudel 2025 (?)
- Li 2025 (?)

* multimodal FL with missing modalities

- multimodal FL by Augmentation by Retrieval in clients (Poudel et. al., MICCAI 2024)
- Feature Imputation Network (FIN) (Poudel et. al. MILA 2025)

* What makes medical AI unreliable?

- heterogeneous data
- long-tail problem
- ???

- Feature-based OOD detection using NCDD score in GI (Pokhrel et. al. MICCAI 2025)

? nearest center ?
? deficit ??

- NERO : Explainable Out-of-Distribution Detection with Neuron-level Relevance (Chhetri et. al., MICCAI 2025)

- Layerwise relevance propagation: an overview (2018)

- From Attribution Maps to Human-understandable explanations through Concept Relevance Propagation, 2023 Nature machine Intelligence

- Benchmark for Hallucination Detection in VLM for GI Imaging. (Pokhrel et. al. MICCAI 2025)