Computer Vision EXAM

The exam can be taken in one of two ways:

1. by taking a **written** exam on the topics covered during the course. This method follows the scheduled exam dates.
2. by completing a **project** according to the methods outlined below.

In both cases registering for an exam session is necessary to record the grade.

Project modalities

- The project deadline is independent of the scheduled dates for the oral exams. If you decide to take the exam with a project, please contact the professor to agree on the project to be carried out.
- The project must be submitted within a year from the date of assignment (after this period, please contact the professor to update the proposal).
- The project can be carried out by 1 or 2 people. In the second case, a greater analysis and in-depth investigation is expected (e.g. comparing different techniques).
- Those who intend to pursue this path are required to produce:
  - functioning code with adequate experimentation and analysis.
  - A report that frames the context, formally describes the methodology pursued, and reports experimental results, and their analysis. The report can be written in either Italian or English of your choice.
  - Finally, the project must be presented (approximately one week after submitting the code and report) with a presentation of about 20 minutes (using slides if desired). Pay attention: during the presentation, questions may be asked to assess how much the concepts presented in class have been studied.
- Below are some project proposals. In case all these projects are assigned, the professor will propose other topics.

Project proposal

1. **Proposal: Evaluating the Generalization of Hierarchical Spatio-Temporal GCNs for Anomaly Detection in Hand Movements**

This project invokes the evaluation of the Hierarchical Spatio-Temporal Graph Convolutional Neural Network (HSTGCN), originally developed for human action anomaly detection, in the context of hand movement anomaly detection. The task focuses on fine-grained motion represented by a 42-node graph, where each node encodes the 3D position of a specific point on the hand over time. The objective is to assess whether the spatio-temporal modeling capabilities of HSTGCN are effective when transferred to this domain.

The anomaly detection strategy follows a predictive approach. The model is trained on sequences of normal hand movements, learning to forecast future hand positions from past ones. At test time, samples that produce high prediction errors are flagged as anomalies, based on the assumption that the model, trained only on normal data, is unable to accurately predict abnormal or unseen patterns.

The project entails adapting the HSTGCN architecture and data pipeline to accommodate hand-specific input. This includes formatting the data as 42-node spatio-temporal graphs, with each node capturing a 3D coordinate trajectory. The temporal continuity is maintained to fully exploit the model's dynamic feature extraction capabilities.

Performance evaluation will rely on quantitative and qualitative metrics, such as prediction error (mean squared error across all nodes) and ROC-AUC for anomaly classification. The evaluation will explore the model's sensitivity to different types of anomalies, such as irregular gestures or tremor-like motions.

Sources:

HTSGCN,

Dataset

2. Proposal: Evaluating the Transferability of Hierarchical Temporal Transformer (HTT) for 3D Hand Pose Estimation on EPIC-KITCHENS Dataset

This project investigates the performance and generalizability of the pre-trained Hierarchical Temporal Transformer (HTT), originally developed for 3D hand pose estimation and action recognition from egocentric RGB videos when applied to the EPIC-KITCHENS dataset. The HTT model is designed to extract 3D hand landmarks from video sequences, leveraging temporal attention mechanisms to capture fine-grained motion and spatial correlations. The project aims to assess whether the pre-trained HTT can effectively extract accurate 3D hand poses from a novel dataset without extensive retraining.

Quantitatively, the evaluation focuses on the successful extraction and visualization of 3D hand landmarks from video frames. The extracted landmarks are rendered over the video sequences to visually inspect temporal consistency and pose realism. Qualitatively, a comparative analysis is conducted between the 2D projections of the HTT-generated landmarks and those produced by MediaPipe's hand-tracking solution, which serves as a reference baseline. This comparison assesses alignment and anatomical plausibility in the image plane, providing insight into the fidelity of the 3D-to-2D mapping.

SOURCES:

HTT,

EPIC KITCHENS

3. Proposal: Evaluation of MedSAM Guided by Prompts Derived from Label Fusion

This project aims to evaluate the performance of MedSAM in replicating segmentation masks obtained through multiple annotators by using geometric prompts derived from fused annotations. The study is conducted without ground truth. It considers factors such as annotator disagreement, expertise level (ranging from 1 to 4), estimated image difficulty based on annotation variability, and the type of prompt used—either a centroid or a bounding box.

The dataset consists of 50 images, each annotated by approximately 30 individuals, without any available ground truth segmentations. The project pipeline begins with estimating the difficulty of each image using entropy, average Dice score, and area variation across annotations. Annotation fusion is then performed using three methods: Majority Voting, STAPLE, and Weighted Voting based on annotator expertise. Each method is applied separately to expert and non-expert groups, producing multiple fused masks. Prompts are extracted from these fused masks as either a centroid point or a bounding box, and MedSAM is run on each image using all possible combinations of prompts, fusion methods, and annotator groups.

The output masks are compared against the fused masks that generated the prompts, using evaluation metrics such as Dice coefficient, Intersection-over-Union, Hausdorff Distance (HD95), precision, recall, and volume difference.

The final analysis investigates which prompt types and fusion methods yield better results, whether MedSAM replicates expert or non-expert fused masks more accurately, how performance varies across image difficulty levels, and if non-experts can produce reliable fused masks. Twelve combinations are tested in total to fully characterize MedSAM's behavior across these conditions.

SOURCES:

https://github.com/bowang-lab/MedSAM
https://www.nature.com/articles/s41467-024-44824-z

4. Proposal: Study and application of Causal Recurrent Variational Autoencoder (CR-VAE) for EEG Time Series Generation

This project aims to explore and apply the Causal Recurrent Variational Autoencoder (CR-VAE) — a state-of-the-art generative model that integrates temporal and causal structure into its learning and generation process — to real EEG (electroencephalography) data. Unlike traditional recurrent VAEs, CR-VAE introduces a multi-head decoder architecture, where each head generates one specific dimension of the multivariate time series. Through a sparsity-inducing regularization, the model learns a Granger-causal adjacency matrix that reveals directional dependencies among the variables, enabling a causally consistent generation of synthetic data.

The project will involve training CR-VAE on EEG datasets and assessing its performance in generating realistic, temporally coherent, and causally plausible EEG signals. Evaluation will be based on:

- Quantitative metrics of generation quality (e.g., reconstruction loss, predictive power)
- Causal interpretability via the inferred Granger causal graph,
- Utility in downstream tasks, such as data augmentation for classification or anomaly detection in clinical settings.

Ultimately, the project will evaluate whether CR-VAE can serve as a reliable tool for modeling complex brain signals, producing high-quality synthetic EEG data while simultaneously uncovering meaningful causal structures within the data.

SOURCES:

GitHub link

Paper link

5. Proposal: Representation Learning for EEG using VAEEG: A Variational Autoencoder Approach

This project focuses on the study and application of VAEEG, a deep learning architecture based on Variational Autoencoders (VAEs), specifically designed to extract meaningful representations from EEG (electroencephalography) data. VAEEG aims to learn a low-dimensional latent space that captures the intrinsic structure of raw EEG signals in an unsupervised manner, facilitating a compact and informative encoding of neural activity.

The core objective of this project is twofold:

1. To analyze and implement the VAEEG architecture, evaluating its ability to encode EEG signals into robust latent representations.
2. To leverage these learned representations for downstream tasks, such as classification (e.g., disease diagnosis, cognitive state detection) or regression (e.g., age or severity score prediction).

GitHub Link

Paper Link

6. Proposal: Latent alignment in deep learning models for EEG decoding

The issue of *subject shift*—that is, variability between individuals that hinders the generalization of machine learning models—is a critical challenge in many biomedical contexts, especially in EEG signal analysis. Due to deep inter-subject differences—morphological, functional, or linked to experimental conditions—models trained on one set of subjects often perform poorly on unseen individuals.

The recent work by Bakas et al. (2025), **"Latent Alignment in Deep Learning Models for EEG Decoding"**, introduces a novel and effective method for tackling domain shift through **unsupervised alignment in the latent feature space** of deep learning models. Instead of aligning EEG signals at the input level, their **Latent Alignment** approach performs standardization in the **intermediate layers (latent space)** of the model, using **subject-specific statistics** derived from trial sets. This alignment is applied consistently during both **training and inference**, making the model more robust to subject variability without requiring labeled data from the target subject.

The goal of this proposal is to **study and apply the Latent Alignment framework to biomedical scenarios affected by strong subject shift** like EEG signals. In particular, we aim to investigate:

- The effectiveness of latent-space alignment in improving generalization to unseen subjects;

- The impact of aligning features at multiple model stages on classification performance and latent space structure;

This study will validate the approach on public datasets, comparing its performance to standard baselines and alternative domain adaptation techniques, with a particular focus on robustness to class imbalance.

GitHub link

Paper Link

7. Proposal: Camera-based Respiratory Measurement from Monocular 3D Estimation

This project investigates the estimation of respiratory rate from upper-body videos acquired via a monocular RGB camera. Prior literature has shown that applying deep learning directly on RGB frames often yields suboptimal results in remote respiration measurement, underscoring the need for appropriate inductive biases. To this end, we propose incorporating a geometric pre-processing stage that transforms each video frame into a 3D point cloud or consecutive frames into a scene flow representation. This transformation encourages the model to focus on task-relevant motions, particularly the anterior-posterior, mediolateral, and vertical displacements of the thorax.

The project objectives are summarized as follows:

1. Preprocessing: Apply state-of-the-art monocular estimation methods to generate either 3D point clouds (RGB2Point) or scene flow (SelfMonoSF, SFExpansion, MonoPLFlowNet). This step is applied to videos in synthetic (SCAMPS) and real (COHFACE, BP4D+) datasets (possibly, with optional foreground extraction).

2. Preliminary Validation: Visualize a subset of generated point clouds/scene flow fields to qualitatively assess body reconstruction. Evaluate the utility of the geometric representations for respiration estimation through the following steps:

   2.1. Convert each 3D video into three 1D signals (e.g., median X, Y, Z coordinates per frame); optionally apply PCA to extract the principal mode of variation.
   2.2. Bandpass filter the signals and ground truth (GT) in the respiratory range [0.1, 0.5] Hz.
   2.3. Estimate respiratory frequency via the dominant peak in the Power Spectral Density (PSD), computed using Welch's method.
   2.4. Assess prediction accuracy, between predicted and GT frequencies using RMSE, MAE, PCC, and CCC.

3. Supervised Deep Learning: Train a neural model from scratch on the 3D point cloud or scene flow pre-processed videos from SCAMPS and COHFACE, using either a frequency-domain loss or negative Pearson correlation loss (losses are listed in the available RPPG-Toolbox).

4. Generalization: Evaluate the trained model on BP4D+ by estimating the respiratory frequency and applying steps 2.2, 2.3 and 2.4.

The entire pipeline, from pre-processing to model training and evaluation, will be implemented within the **resPyre** framework (https://github.com/phuselab/resPyre), which will serve as the execution environment for this study. The **RPPG-Toolbox** framework (https://github.com/ubicomplab/rPPG-Toolbox) can be used as an additional reference. Care must be taken to align RGB frames with ground-truth signals during training, account for different sampling rates across datasets, and manage videos of variable lengths.

POINTCLOUD FROM RGB:

A) [WACV 2025] RGB2Point:3D Point Cloud Generation from Single RGB Images [CODE: https://github.com/JaeLee18/RGB2point]

SCENE FLOW FROM RGB:

A) [CVPR 2020 (Oral)] Self-Supervised Monocular Scene Flow Estimation [CODE: https://github.com/visinf/self-mono-sf]

B) [CVPR 2020 (Oral)] Upgrading Optical Flow to 3D Scene Flow through Optical Expansion [CODE: https://github.com/gengshan-y/expansion]

C) [ECCV 2022] MonoPLFlowNet: Permutohedral Lattice FlowNet for Real-Scale 3D Scene Flow Estimation with Monocular Images [CODE: https://github.com/BlarkLee/MonoPLFlowNet]

D) [CVPR 2025 (Oral, Best Paper Candidate)] Zero-Shot Monocular Scene Flow Estimation in the Wild [CODE: coming soon]

8. Proposal: *Dental Enumeration and Diagnosis on Panoramic X-rays (2D Images)*

Panoramic X-rays are widely used in dental practice to provide a comprehensive view of the oral cavity and aid in treatment planning for various dental conditions. However, interpreting these images can be a time-consuming process that can distract clinicians from essential clinical activities. Moreover, misdiagnosis is a significant concern, as general practitioners may lack specialized training in radiology, and communication errors can occur due to work exhaustion.

In recent years, advancements in artificial intelligence (AI) have paved the way for automated dental radiology analysis. However, developing automated algorithms for panoramic X-ray analysis is challenging due to variations in anatomy and the lack of publicly available annotated data. Despite these challenges, the potential benefits of utilizing AI in dental radiology analysis cannot be ignored, as it can significantly improve treatment outcomes and patient satisfaction. Therefore, there is a growing need for research to explore and develop effective AI algorithms for dental radiology.

Core objectives of this project:

- address this challenge by developing a self-supervised model in order to cope with the lack of annotated data in this field.
- studying the gap in performances between self-supervised and supervised models.

Data: The DENTEX dataset comprises panoramic dental X-rays obtained from three different institutions using standard clinical conditions but varying equipment and imaging protocols, resulting in diverse image quality reflecting heterogeneous clinical practice. The dataset includes X-rays from patients aged 12 and above, randomly selected from the hospital's database to ensure patient privacy and confidentiality.

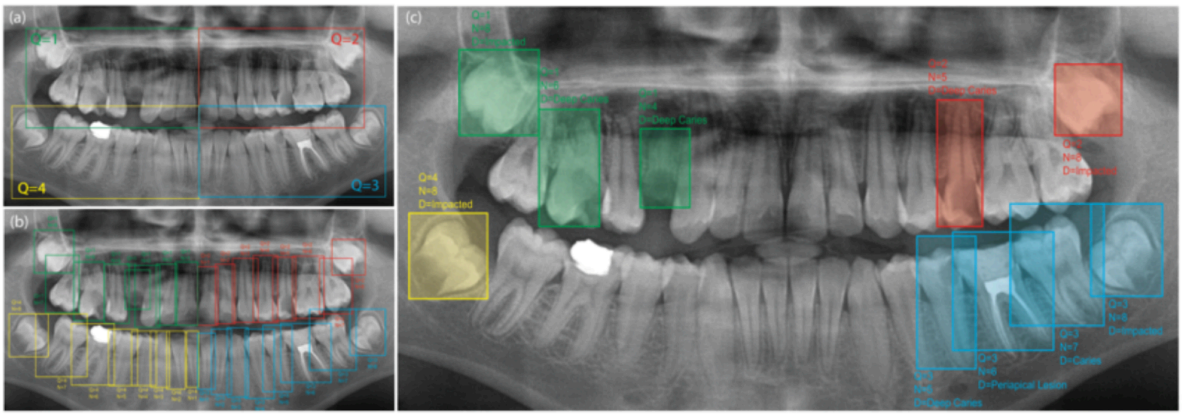
To enable effective use of the FDI system, the dataset is hierarchically organized into three types of data;

(a) 693 X-rays labeled for quadrant detection and quadrant classes only,

(b) 634 X-rays labeled for tooth detection with quadrant and tooth enumeration classes,

(c) 1005 X-rays fully labeled for abnormal tooth detection with quadrant, tooth enumeration, and diagnosis classes.

The diagnosis class includes four specific categories: caries, deep caries, periapical lesions, and impacted teeth. An additional 1571 unlabeled X-rays are provided for pre-training.


(a)
Q=1
Q=2
Q=4
Q=3
(b)
(c)
Q=1
N=8
D=Impacted
Q=4
N=8
D=Impacted
Q=2
N=5
D=Caries, Cavities
Q=2
N=6
D=Impacted
Q=3
N=6
D=Impacted
Q=3
N=7
D=Caries
Q=3
N=6
D=Periapical Lesion

9. PROPOSAL: 3D Teeth Scan Segmentation and Labeling (3D data)

Computer-aided design (CAD) tools have become increasingly popular in modern dentistry for highly accurate treatment planning. In particular, in orthodontic CAD systems, advanced intraoral scanners (IOSs) are now widely used as they provide precise digital surface models of the dentition. Such models can dramatically help dentists simulate teeth extraction, move, deletion, and rearrangement and ease therefore the prediction of treatment outcomes. Hence, digital teeth models have the potential to release dentists from otherwise tedious and time consuming tasks.

Although IOSs are becoming widespread in clinical dental practice, there are only few contributions on teeth segmentation/labeling available in the literature and no publicly available database. A fundamental issue that appears with IOS data is the ability to reliably segment and identify teeth in scanned observations. Teeth segmentation and labeling is difficult as a result of the inherent similarities between teeth shapes as well as their ambiguous positions on jaws.

In addition, it faces several challenges:

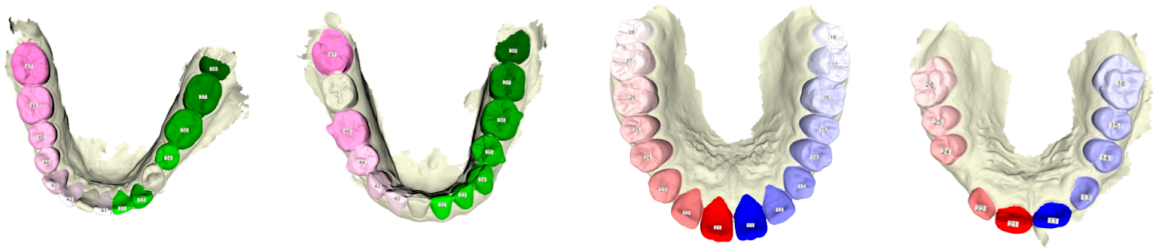
1. The teeth position and shape variation across subjects.
2. The presence of abnormalities in dentition. For example, teeth crowding which results in teeth misalignment and thus non-explicit boundaries between neighboring teeth. Moreover, lacking teeth and holes are commonly seen among people.
3. Damaged teeth.
4. The presence of braces, and other dental equipment

Core objectives of this project:

- address this challenge by developing a self-supervised model in order to cope with the lack of annotated data in this field.
- studying the gap in performances between self-supervised and supervised models.

Data:

A total of 1800 3D intra-oral scans have been collected for 900 patients covering their upper and lower jaws separately.



10. PROPOSAL: Multi-Structure segmentation in CBCT Volumes (Volumetric Data)

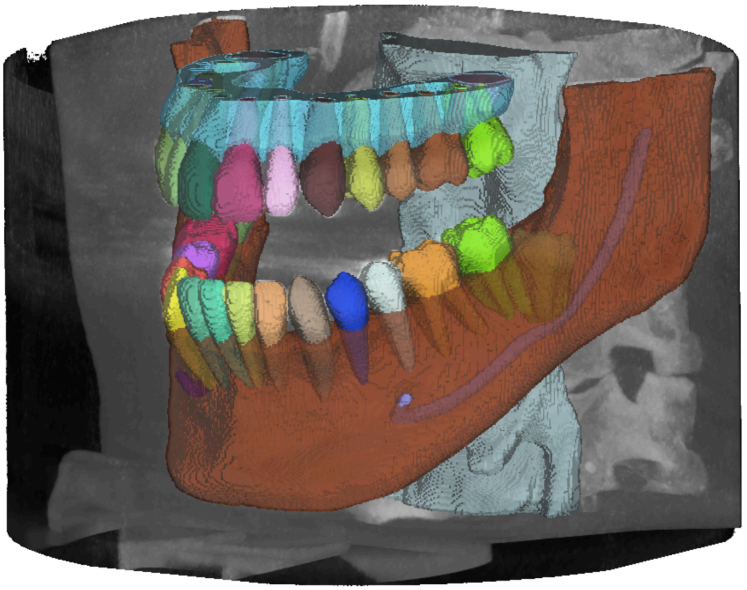
The use of Cone Beam Computed Tomography (CBCT) is growing not only in dentistry but across the broader field of head and neck surgery. CBCT offers key advantages, including short acquisition times and low radiation doses, while still providing excellent visualization of anatomical structures—particularly hard tissues. Consequently, a wide range of anatomical regions must be considered during segmentation, such as the mandible, teeth, maxillary bone, and pharynx. These structures are relevant across multiple surgical specialties, as well as in routine clinical and anesthesiological practice. In this context, deep learning models can assist medical professionals in surgical planning by enabling automated, voxel-level segmentation.

Core objectives of this project:

- address this challenge by developing a self-supervised model in order to cope with the lack of annotated data in this field.
- studying the gap in performances between self-supervised and supervised models.

Data:

The ToothFairy2 dataset follows the nnU-Net dataset format. Datasets consist of three components: raw images, corresponding segmentation maps and a dataset.json file specifying some metadata. The class IDs are an extension of the FDI World Dental Federation notation and include: Background, Jaws, Inferior Alveolar Canals, Maxillary Sinus, Pharynx, Bridges, Crowns, Implants, and Upper and Lower Teeth (Wisdom Teeth included), for a total of 42 classes.



11. PROPOSAL: Indoor Scene Segmentation (3D Point Clouds)

This project aims to conduct a comparative analysis of state-of-the-art deep learning models for indoor scene segmentation using 3D point cloud data. Accurate scene segmentation is a critical task in various applications, including robotics, augmented reality, and indoor navigation systems. By evaluating the performance of multiple models on a standardized dataset, we aim to gain insights into their strengths, limitations, and suitability for real-world deployment.

Models for Comparison:
The project will focus on the following models:

- PointNet++ – An extension of the original PointNet architecture, PointNet++ captures local features at multiple scales using hierarchical sampling and grouping techniques.

- DGCNN (Dynamic Graph CNN) – A model that constructs local graphs dynamically and uses edge convolution operations to capture geometric relationships between points.

- PAConv (Position Adaptive Convolution) – A novel model that enhances convolution operations in point clouds by learning dynamic kernel weights based on relative positions, improving local feature representation.

Dataset:
The Stanford Large-Scale 3D Indoor Spaces Dataset (S3DIS) will serve as the primary dataset for training and evaluation. It consists of richly annotated indoor spaces scanned from six large-scale areas, making it ideal for benchmarking segmentation models in diverse environments.

The core objective is to compare and contrast the different models