ODTÜ METU

TÜBİTAK

EURO 4SEE

Optimizing Deep Learning Systems for Hardware

Assoc. Prof. Erdem AKAGÜNDÜZ, METU

ncc@ulakbim.gov.tr

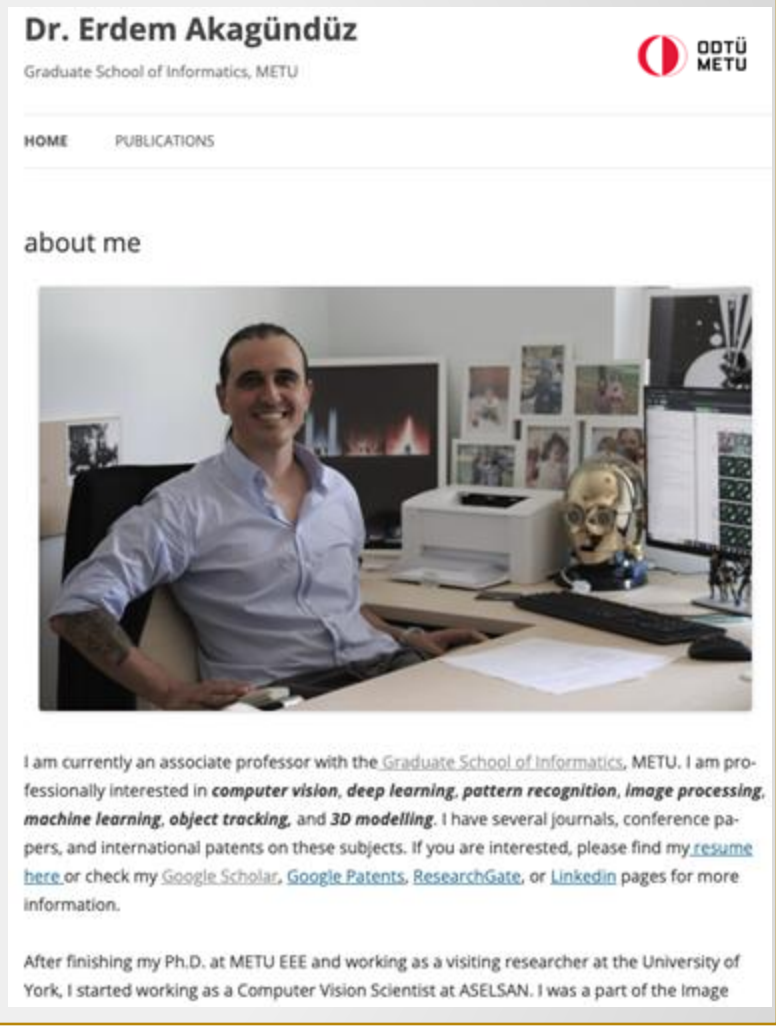
Welcome to the Course

4SEE

Meet the Instructor:

- Education
  - B.Sc. EEE METU, Turkey (1997-2001)
    B.Sc. Minor Program, Sociology METU (1998-2001)
  - M.Sc. EEE, METU, Turkey
    on "3D Modelling and Graphics" (2001-2004)
  - Ph.D. EEE, METU, Turkey
    on "3D Object/Surface Representation" (2004-2011)
- Professional
  - Research and Teaching Assistant, EEE METU, (2001 - 2008)
  - Visiting Phd Student, the Univ. of York (2008 - 2009)
  - Applied Scientist, ASELSAN, Inc, (2009 - 2016)
  - Research Associate, the Univ. of York, (2016)
  - Assistant Professor, EEE Çankaya Uni. (2018 - 2021)
  - Visiting Professor, the Univ. of York, (2022-2023)
  - Associate Professor, MMI II METU (2021 - …

https://blog.metu.edu.tr/akaerdem/



Dr. Erdem Akagündüz

Graduate School of Informatics, METU

ODTÜ METU

HOME PUBLICATIONS

about me

I am currently an associate professor with the Graduate School of Informatics, METU. I am professionally interested in ***computer vision***, ***deep learning***, ***pattern recognition***, ***image processing***, ***machine learning***, ***object tracking***, and ***3D modelling***. I have several journals, conference papers, and international patents on these subjects. If you are interested, please find my resume here or check my Google Scholar, Google Patents, ResearchGate, or Linkedin pages for more information.

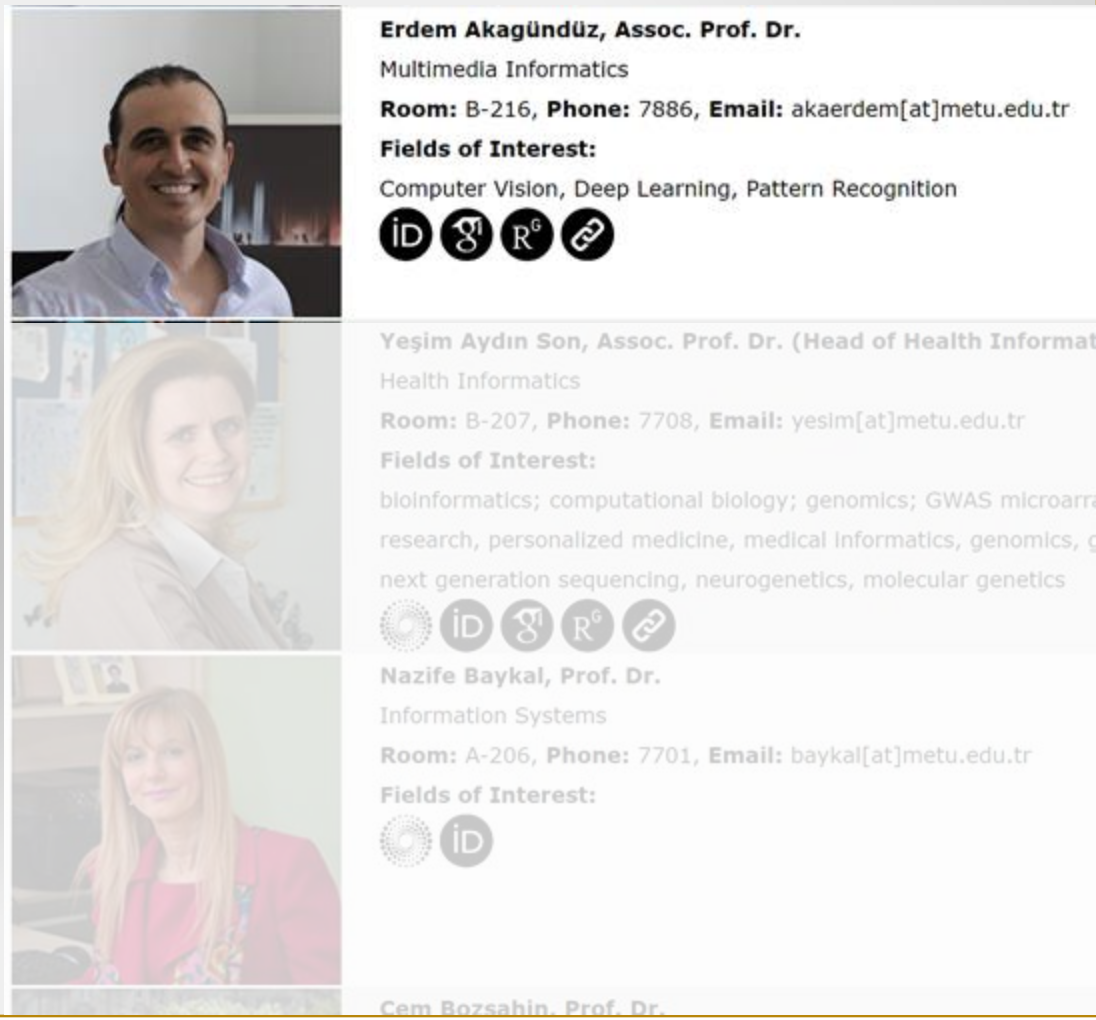
After finishing my Ph.D. at METU EEE and working as a visiting researcher at the University of York, I started working as a Computer Vision Scientist at ASELSAN. I was a part of the Image

Our Institute

METU Informatics Institute

- Associate Professor, MMI II METU
  - MSc, PhD Advisorships,
  - Research Project (TUBITAK) Administrations
  - Industrial Consultancies
  - Teaching
    - DI504 Foundations of Deep Learning
    - MMI711 Sequence Models in Multimedia
    - MMI714 Generative Models in Multimedia
    - IS566 Image Processing Algorithms
    - MMI704 Human Motion Capture, Analysis and Synthesis

https://ii.metu.edu.tr/full-time-faculty



Our Lab.

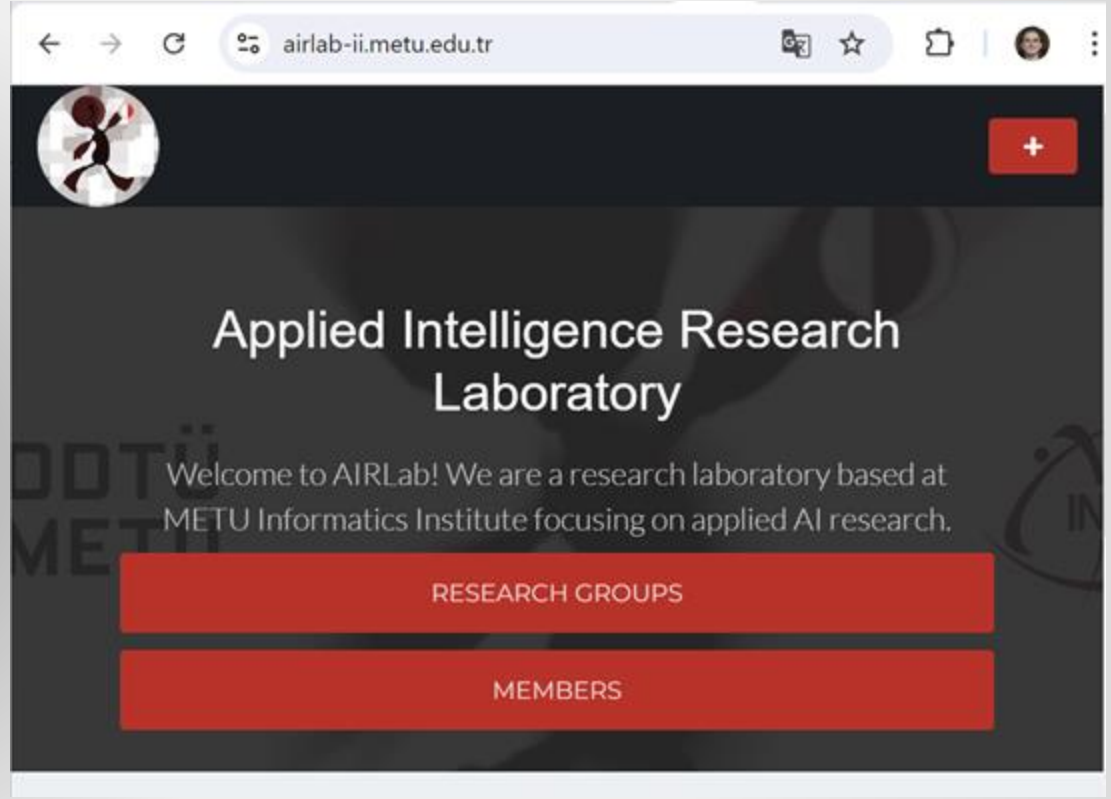
METU II AIRLab

4SEE

https://airlab-ii.metu.edu.tr/

- **AIRLab** is home to different research groups focusing on applying AI problems across diverse domains including, but not limited to, computer vision, geoinformatics, earthquake engineering, and decision systems.

- For comprehensive insights into specific projects, publications, and team members, please navigate to the individual pages of each research group.



Pre-knowledge/P-rerequisite(s)

(Must)

- Experience with Training Deep Neural Networks
  - *i.e. taken at least an intro level DL course*
- Familiarity with “Computer Architecture” Concepts

(Optional/Recommended)

- Familiarity with Deep Learning Model Optimization

Course Overview

- Understand how hardware and system design impact deep learning performance and efficiency.
- Learn strategies to optimize models and systems, from memory usage to computation and precision.
- Explore scaling concepts and frameworks for deploying deep learning across edge, cloud, and HPC environments.

What this course is

- This course explores the relation between hardware and deep learning model
- It presents optimization strategies, separately
  - at the model
  - and system levels.
- It briefly introduces deployment through the example of scaling models in HPC using DeepSpeed.

What this course isn’t

- It isn’t a hands-on course with setup, configuration, or installation steps.
- It isn’t a complete guide to scaling deep models across diverse platforms.
  - (*Instead, it is an introduction aimed at learners who are willing to take the first steps toward scaling and deployment.*)

Introduction and Set Up/Configure/Install …..

- No installation, setup, or configuration is required.
- The course is entirely theoretical, including the introduction to the HPC framework at the end.

Outline

- **Part I : Fundamentals**
- Part II : Hardware Types & Memory Hierarchy
- Part III : Model-Level Optimizations
- Part IV : System-Level Optimizations
- Part V : Introduction to Scaling Deep Learning in HPC

EURO 4SEE

Thanks!

Co-funded by
the European Union

EuroHPC
Joint Undertaking

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101191697. The JU receives support from the Digital Europe Programme and Germany, Türkiye, Republic of North Macedonia, Montenegro, Serbia, Bosnia and Herzegovina.