Advanced Deep Learning and Computer Vision

Code: CS32125CS	Course Type: Elective	Semester: I
Evolution Scheme (TA-MSE-ESE): 20-30-100	Total Marks: 150	L-T-P (Credits): 3-0-0 (3)

Pre-Requisites:

- Foundations of Machine Learning and Pattern Recognition
- Basics of Artificial Neural Networks and Backpropagation
- Multivariable Calculus, Linear Algebra, and Probability
- Proficiency in Python and Deep Learning Frameworks (PyTorch / TensorFlow).

Course Objectives:

1. To explore advanced spatial hierarchies and implement state-of-the-art dense prediction models for object detection and segmentation.
2. To formulate and train deep generative models, including Diffusion Models and GANs for high-fidelity image synthesis.
3. To analyze dynamic visual data by applying spatiotemporal neural architectures for video understanding and action recognition.
4. To investigate 3D computer vision and vision-language foundation models for zero-shot inference and embodied AI representations.

Course Outcomes: At the end of the course, the student will be able to

CO-1	Architect and evaluate advanced Convolutional Networks, Vision Transformers (ViTs), and dense object detectors/segmentation frameworks.
CO-2	Formulate the mathematics of latent generative models (GANs, DDPMs) and apply them to complex image-to-image translation synthesis.
CO-3	Design multi-stream and 3D spatiotemporal networks for complex video analytics, optical flow estimation, and action recognition.
CO-4	Integrate multi-view geometry with Implicit Neural Representations (NeRFs) and apply Vision-Language Models (CLIP) for multimodal tasks.

Course Articulation Matrix:

PO	PEO-1	PEO-2	PEO-3	PEO-4	PEO-5
CO-1	3	2	2	1	3
CO-2	3	2	2	1	3
CO-3	3	3	2	2	3
CO-4	3	3	2	2	3

1 - Slightly; 2 - Moderately; 3 – Substantially

Syllabus:

Unit-1: Advanced Spatial Hierarchies & Dense Prediction - Review of deep CNN hierarchies (ResNet, ConvNeXt). Self-Attention in Vision: Formulating the Vision Transformer (ViT) and Hierarchical Transformers (Swin). Self-Supervised Learning (SSL): Contrastive approaches (SimCLR) and Masked Image Modeling (MAE). Evolution of Object Detection: Single-Stage (YOLOv8/v9), Anchor-free paradigms (CenterNet), and Detection Transformers (DETR/Deformable DETR). Semantic, Instance, and Panoptic Segmentation architectures (UNet++, Mask2Former). Human Pose Estimation.

Unit-2: Deep Generative Models and Image Synthesis - Generative Adversarial Networks (GANs): Minimax formulation, Wasserstein GANs (WGAN-GP), and Spectral Normalization. High-fidelity synthesis via StyleGAN architectures. Conditional synthesis and Image-to-Image translation (Pix2Pix, ControlNet). Denoising Diffusion Probabilistic Models (DDPM): Forward/reverse processes, score-based generative modeling, and Langevin dynamics. Latent Diffusion Models (Stable Diffusion) and text-to-image conditioning mechanisms.

Unit-3: Spatiotemporal Analytics and Video Understanding - Challenges in temporal data modeling. Optical Flow estimation using deep learning (FlowNet, RAFT). Action Recognition paradigms: Two-stream convolutional networks. 3D Convolutional Neural Networks (C3D, I3D, SlowFastarchitectures). Spatiotemporal attention and Video Transformers (TimeSformer, Video Swin Transformer). Video Object Segmentation (VOS) and Multi- Object Tracking (MOT) using deep association metrics (DeepSORT).

Unit-4: 3D Computer Vision & Multimodal Foundation Models - Processing 3D Data: Point cloud geometry, PointNet, and PointNet++. 3D reconstruction from multi-view geometry. Implicit Neural Representations: Neural Radiance Fields (NeRFs) formulation, volumetric rendering, and mip-NeRF. Introduction to 3D Gaussian Splatting for realtime rendering. Vision-Language Foundation Models: Contrastive Language-Image Pretraining (CLIP), ALIGN. Zero-shot transfer capabilities, visual grounding, and open-vocabulary object detection.

Learning Resources:

Text Books:

1. Computer Vision: Algorithms and Applications Szeliski R. Springer 2022 (2nd Ed.).
2. Computer Vision: Models, Learning, and Inference Prince S.J.D. Cambridge Univ. Press 2012
3. Deep Learning Goodfellow I., Bengio Y.,Courville A. MIT Press 2016

Reference Books:

1. Generative Deep Learning: Teaching Machines to Paint, Write, Compose, and Play Foster D. O'Reilly Media 2023 (2nd Ed.)
2. Multiple View Geometry in Computer Vision Hartley R., Zisserman A. Cambridge Univ. Press 2004 (2nd Ed.)