Deep Generative Models for Audio-Visual Data

Lecturer: Sasha Apartsin

1 Background

This course focuses on modern deep generative models trained on large-scale, unstructured, multimodal data, such as text and images. These models can produce high-quality outputs that closely resemble real data and can be controlled or conditioned on desired attributes. Many important real-world problems can be naturally framed as generative tasks, in which structured outputs are produced from given inputs, such as generating textual descriptions from images. Beyond their direct generative capabilities, a central focus of the course is their transformative role in **creating realistic, diverse, and labeled synthetic datasets**, enabling the training and evaluation of AI systems in domains where annotated data is limited, expensive, or unavailable.

2 Course Content

This course provides a comprehensive introduction to the architectures and foundation models that underpin modern generative AI, as well as the generative modelling approaches built on top of them. It begins with key challenges in computer vision and audio processing. The course then introduces core architectural foundations, including Transformers and Vision Transformers, and examines how these models support both discriminative and generative systems. Building on this foundation, students study major generative approaches, including Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and, with primary emphasis, diffusion models and the Stable Diffusion family. The course highlights modern techniques for controllable, fine-grained generation, as well as the use of generative models to create realistic synthetic training and evaluation data. Adopting a **code-first approach**, it presents core concepts through hands-on examples using contemporary libraries and tools.

3 Course Project

During the course, students will propose, implement, and present an **innovative project** based on the concepts studied in class. Each project will address a **novel vision- or audio-related problem** chosen by the students, with particular emphasis on using generative models to create **synthetic training and evaluation data**. The project will involve comparing different strategies for synthetic data generation and evaluating multiple deep learning models, including at least one model based on the Transformer architecture. Each team will deliver **three in-class presentations**: a project proposal, an interim progress update, and a final presentation. By the end of the course, teams will submit a **GitHub repository** containing all project materials, including presentation slides, source code, datasets, and a detailed README file.

4 PREREQUISITES

While all necessary background material, including machine/deep learning fundamentals and PyTorch library, will be introduced during the course, students are expected to have a solid understanding of foundational machine learning concepts and proficiency in Python. Prior experience with computer vision and deep learning is recommended and will be beneficial.

5 AI TOOLS POLICY

The use of AI tools for generating project code and presentation slides is **highly encouraged**, provided that two essential requirements are met: the project must demonstrate **novelty** by addressing a new and valuable task, and the team must maintain **full ownership** and responsibility for all submitted code and presentation materials.

6 EXPECTED OUTCOMES

Graduates of the course will acquire a broad, in-depth understanding of deep generative models and their applications. By the end of the course, students will have gained hands-on experience with state-of-the-art AI models and software libraries, developing practical skills through structured, guided projects. These projects will provide opportunities to design and implement AI solutions by combining, adapting, and extending existing components. The course emphasizes both technical proficiency and creative problem-solving, preparing students to innovate using modern AI technologies in real-world development scenarios. Students will develop a tangible AI project that can be showcased in a professional portfolio. At the same time, in-class presentations will enhance their ability to present and articulate complex technical work with confidence.

7 WEEKLY SCHEDULE

Below is an approximate weekly schedule outlining the subjects that will be covered. Please note that the actual order or content may vary depending on the class background, recent advancements in the field, or the specific focus of student projects.

Week	Theme	Selected subjects
1	Introduction	Typical vision and audio Tasks, project requirements, PyTorch/DL tutorials
2	Introduction to Language Foundation Models	HuggingFace libraries, LLM for text generation and representation, and an introduction to fine-tuning
3	Vision and Audio Foundation Models	Vision models for image representation and classification, audio quantization, and encoding
4	Synthetic Image Data Generation for Model Training	Introduction to stable diffusion pipelines, using generative models for creating synthetic data for image classification and object detection
5	Student presentations: Project Proposal	
6	Transformers Deep Dive	Tokenization and word embedding, attention, transformer blocks, pretraining, Parameter-Efficient Fine Tuning (PEFT)

7	Video/Audio Transformers	Vision Transformers, DeIT-DIN_SWIN models, DETR, Speech Transformers and Audio Encoder, Multimodal models: CLIP/BLIP, CLAP
8	Variational AutoEncoders	Autoencoder, AE variants, VAE, VAE variants
9	Student Presentations: Interim Report	
10	Stable Diffusion	Math background, Diffusion Models, Latent Diffusion Models, Guidance, Long Prompts, SDXL
11	Stable Diffusion Applications	Inpainting, DepthNet, Text2Video, Prompt2Prompt, Instruction Edit, Image Inversion, ControlNet, Dream Both, and Text Inversion
12	Generative Adversary Networks	GAN intro, DCGAN, WGAN, Pix2Pix/SRGAN, CycleGAN, StyleGAN, BigGAN
13	Student Presentation: Project Finals	

8 COURSE GRADE

Delivery	Grade
In-class project proposal presentation	No grade, for feedback and approval only
In-class project interim presentation	20%
In-class final project presentation	40%
Final project submission (GitHub Repo)	40%

9 REFERENCES

1. Foster, David. *Generative deep learning*, 2022.
2. Tomczak, Jakub M. *Deep generative modelling*, 2nd edition, 2024.
3. Sanseviero, Omar, *Hands-On Generative AI with Transformers and Diffusion Models*, 2024
4. Liu, Mark. *Learn Generative AI with PyTorch*, 2024.