Image and Video Generations

Lecture 0: Course Introduction

劉育綸

Yu-Lun (Alex) Liu

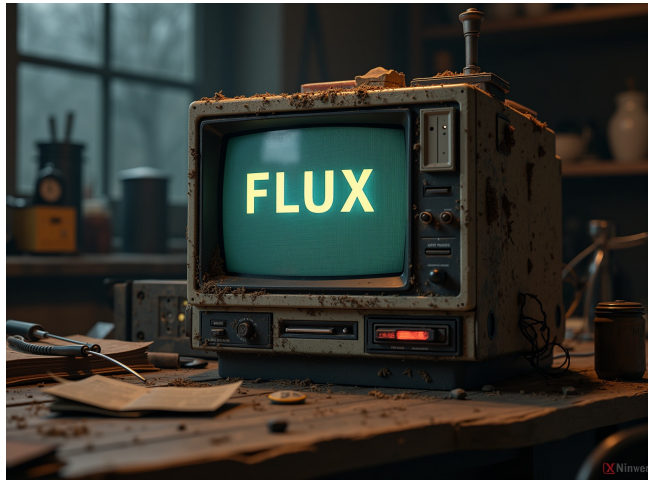
What do you know about image and video generation models?

air head · Made by shy kids with Sora

OpenAI Sora

Luma Dream Machine Ray2

Black Forest Labs Flux



Adobe Firefly

AI-Generated Art Won a Prize

Jason Allen won the digital-art competition at the Colorado State Fair last year for his piece "Theatre D'opera Spatial" that he created using the AI software Midjourney. Recently, the US Copyright Office refused to grant him a copyright for his piece, writing, "We have decided that we cannot register this copyright claim because the deposit does not contain any human authorship:' He plans to appeal.

AI ASMR – Glass Fruits Cutting

https://www.youtube.com/shorts/bJ2cmHFH9ew

Google DeepMind Veo 3

https://deepmind.google/models/veo/

Will Smith Eating Spaghetti test

2023

https://en.wikipedia.org/wiki/Will_Smith_Eating_Spaghetti_test

Will Smith Eating Spaghetti test

https://www.instagram.com/reel/C3i5vAZvRS3/

Will Smith Eating Spaghetti test

2023

2025

https://en.wikipedia.org/wiki/Will_Smith_Eating_Spaghetti_test

Google Nano Banana

Image Edit Arena

Rank (UB) ↑	Model ↑↓	Score ↑↓	95% CI (±) ↑↓	Votes ↑↓	Organization ↑↓	License ↑↓
1	gemini-2.5-flash-image-preview (nano-banana)	1362	±2	2,521,035	Google	Proprietary
2	flux-1-kontext-max	1191	±3	357,196	Black Fores...	Proprietary
3	flux-1-kontext-pro	1174	±2	2,015,530	Black Fores...	Proprietary
3	gpt-image-1	1170	±3	1,026,399	OpenAI	Proprietary
5	flux-1-kontext-dev	1152	±3	1,584,400	Black Fores...	Proprietary
6	qwen-image-edit	1145	±2	1,585,904	Alibaba	Apache 2.0
6	seededit-3.0	1142	±4	1,285,080	Bytedance	Proprietary
8	gemini-2.0-flash-preview-image-generation	1093	±3	1,700,785	Google	Proprietary
9	bagel	1044	±5	12,774	Bytedance	Apache 2.0
10	step1x-edit	1017	±4	138,399	StepFun	Apache 2.0

https://lmarena.ai/leaderboard/image-edit

Text-to-3D Generation

Vintage copper rotary telephone with intricate detailing.

Two-story brick house with red roof and fence.

Glowing orb on a stone pedestal.

Spherical robot with gold and silver design.

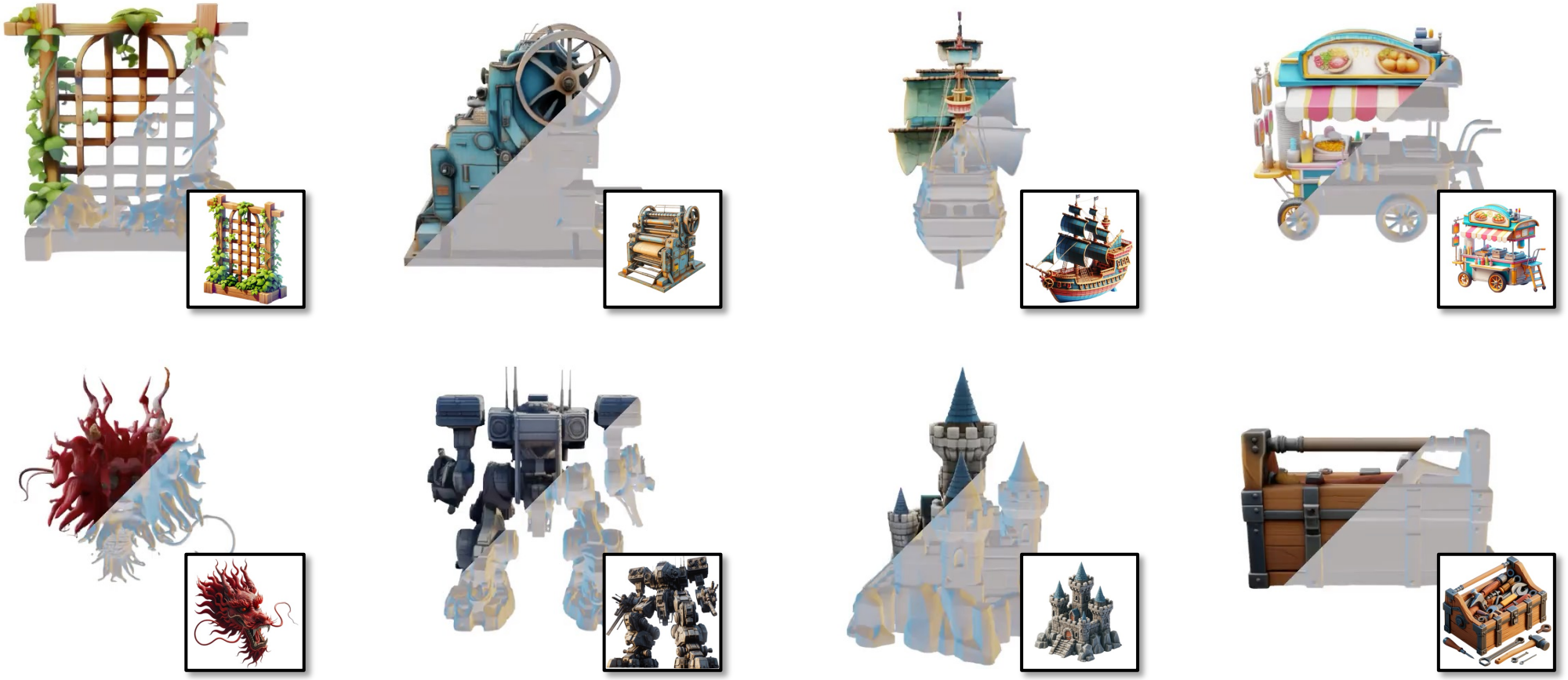
Bronze owl sculpture perched on a branch.

Futuristic robotic arm on a table.

A rustic log cabin with a stone chimney and a wooden porch.

Blocky, orange and teal robot with articulated limbs.

Image-to-3D Generation



Text-to-4D Generation

A storm trooper walking forward and vacuuming, best quality, 4K, HD.

Beer pouring into a glas.

A bee fluttering its wings fast.

A cat singing, best quality, 4K, HD.

Mage in purple robe running forward, full body, portrait, game, unreal, 4K, HD.

Santa Claus carrying a big bag is walking forward, portrait, game, unreal, 4K, HD.

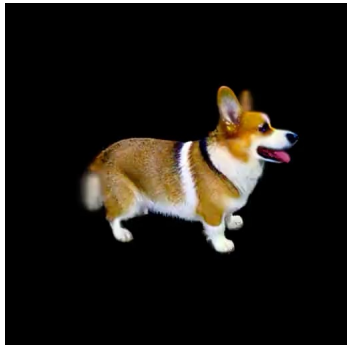
A dog wearing a Superhero outfit with red cape flying through the sky.

Flying dragon on fire.

An astronaut riding a horse, best quality, 4K, HD.

An ancient roman statue dancing, full body, portrait, game, unreal, 4K, HD.

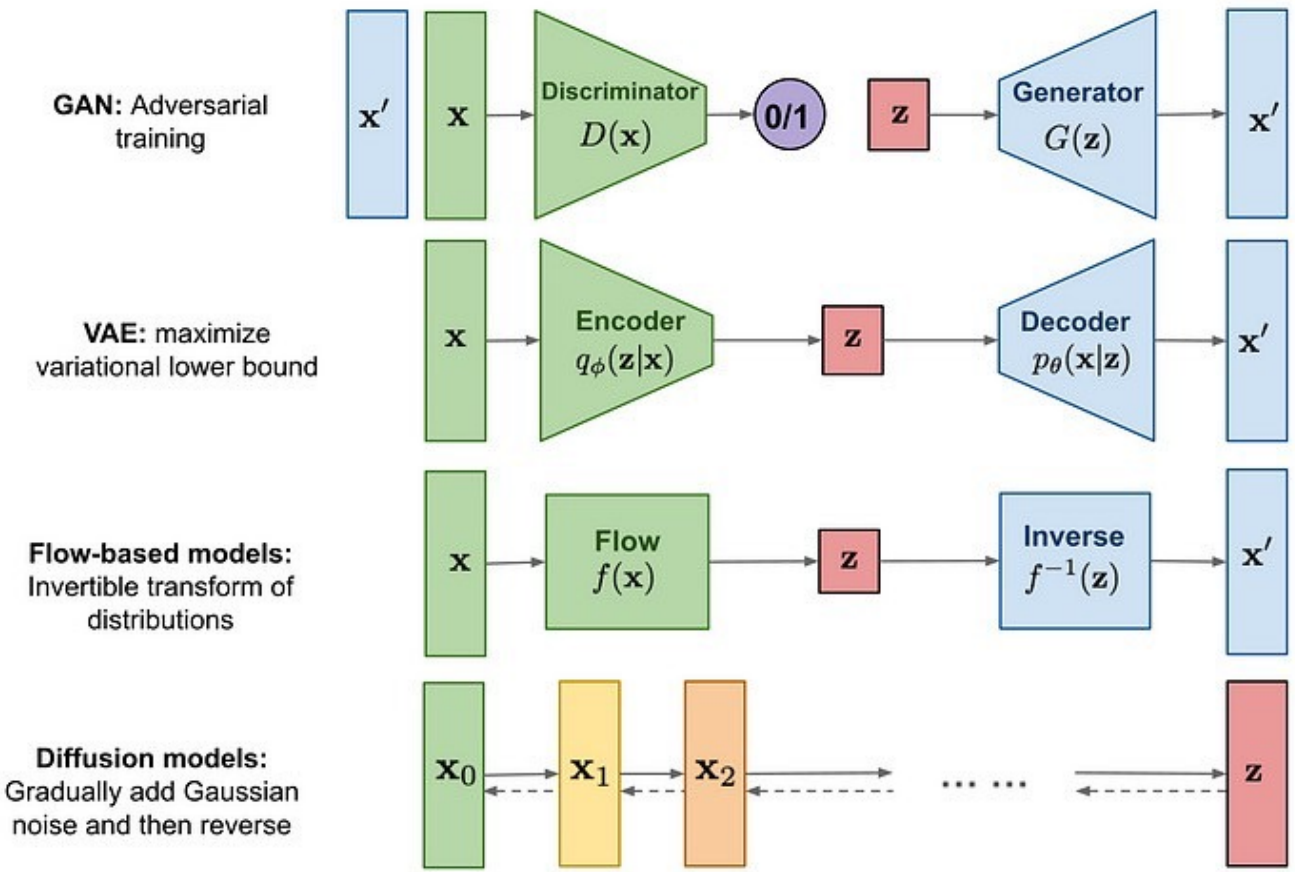
Dragon armor fluttering its wings fast.



A corgi is running fast.

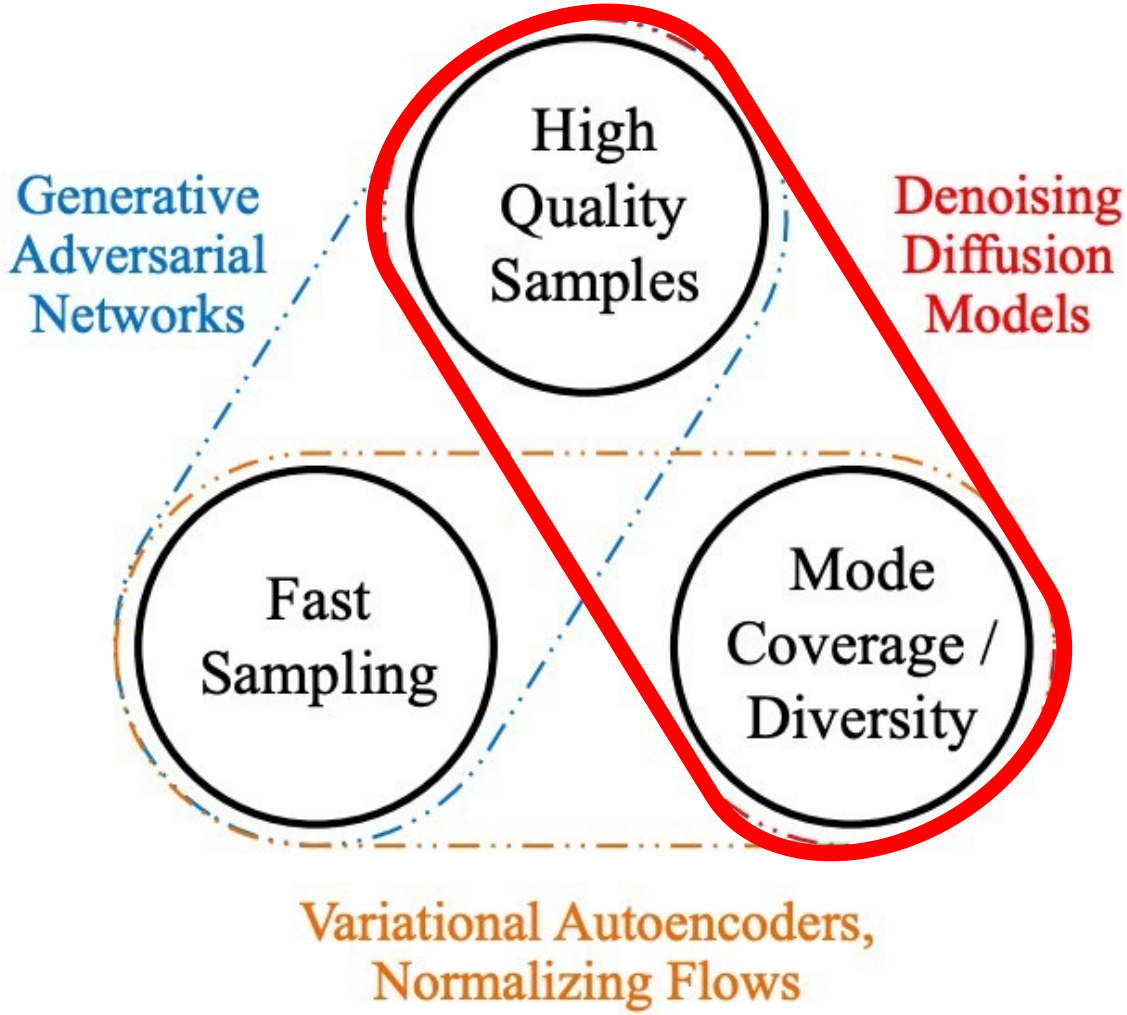
A Brief Overview of the Course

Generative Models



StyleGAN2

Generative Models – Comparison



Higher Quality & Diversity

Diffusion Models Beat GANs on Image Synthesis

Prafulla Dhariwal*
OpenAI
prafulla@openai.com

Alex Nichol*
OpenAI
alex@openai.com

Abstract

We show that diffusion models can achieve image sample quality superior to the current state-of-the-art generative models. We achieve this on unconditional image synthesis by finding a better architecture through a series of ablations. For conditional image synthesis, we further improve sample quality with classifier guidance: a simple, compute-efficient method for trading off diversity for fidelity using gradients from a classifier. We achieve an FID of 2.97 on ImageNet 128×128, 4.59 on ImageNet 256×256, and 7.72 on ImageNet 512×512, and we match BigGAN-deep even with as few as 25 forward passes per sample, all while maintaining better coverage of the distribution. Finally, we find that classifier guidance combines well with upsampling diffusion models, further improving FID to 3.94 on ImageNet 256×256 and 3.85 on ImageNet 512×512. We release our code at https://github.com/openai/guided-diffusion.

Diffusion Models

- (+) High quality
- (+) Diversity
- (−) Slow

Diffusion Models

The generation process of a diffusion model is a denoising process.

Score-Based Model / DDPM / DDIM

Forward SDE (data → noise)

$\mathbf{x}(0)$ — $\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}$ → $\mathbf{x}(T)$

score function

$\mathbf{x}(0)$ ← $\mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - g^2(t)\nabla_{\mathbf{x}} \log p_t(\mathbf{x})\right] \mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}$ — $\mathbf{x}(T)$

Reverse SDE (noise → data)

Diffusion Models

- (+) High quality
- (+) Diversity
- (−) Slow

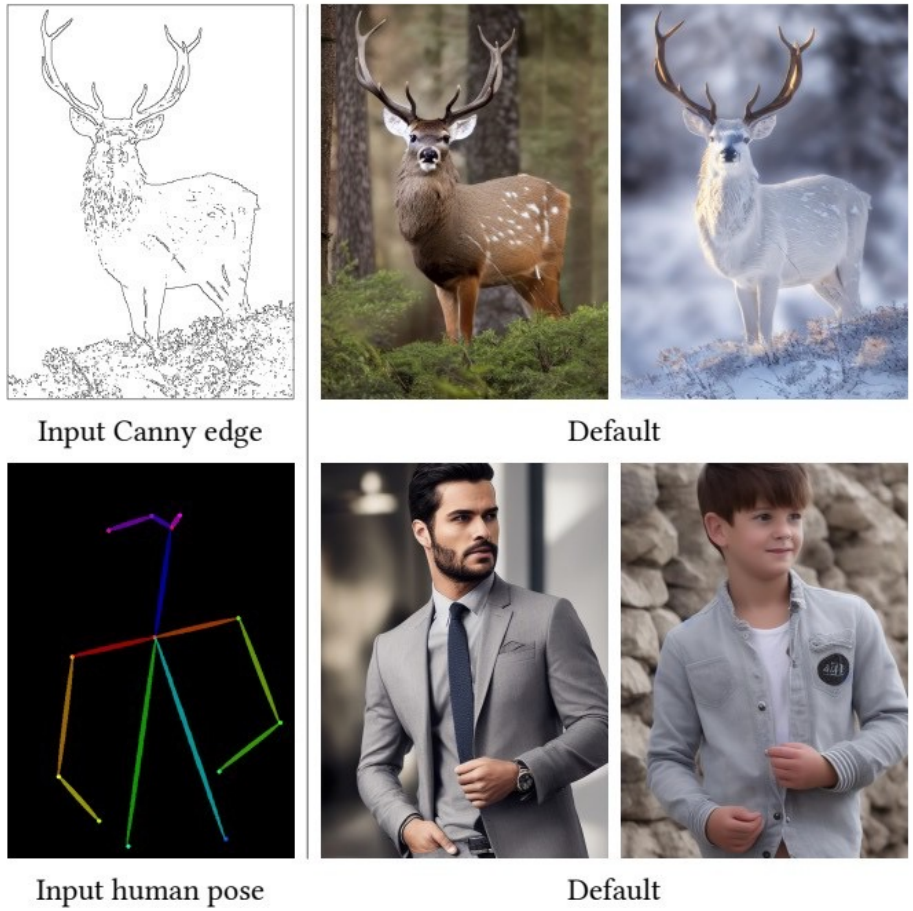
- (+) (Relatively) easy to implement and train
- (+) Easy to convert a conditional model
- (+) Easy to personalize
- (+) Easy to distill knowledge
- …

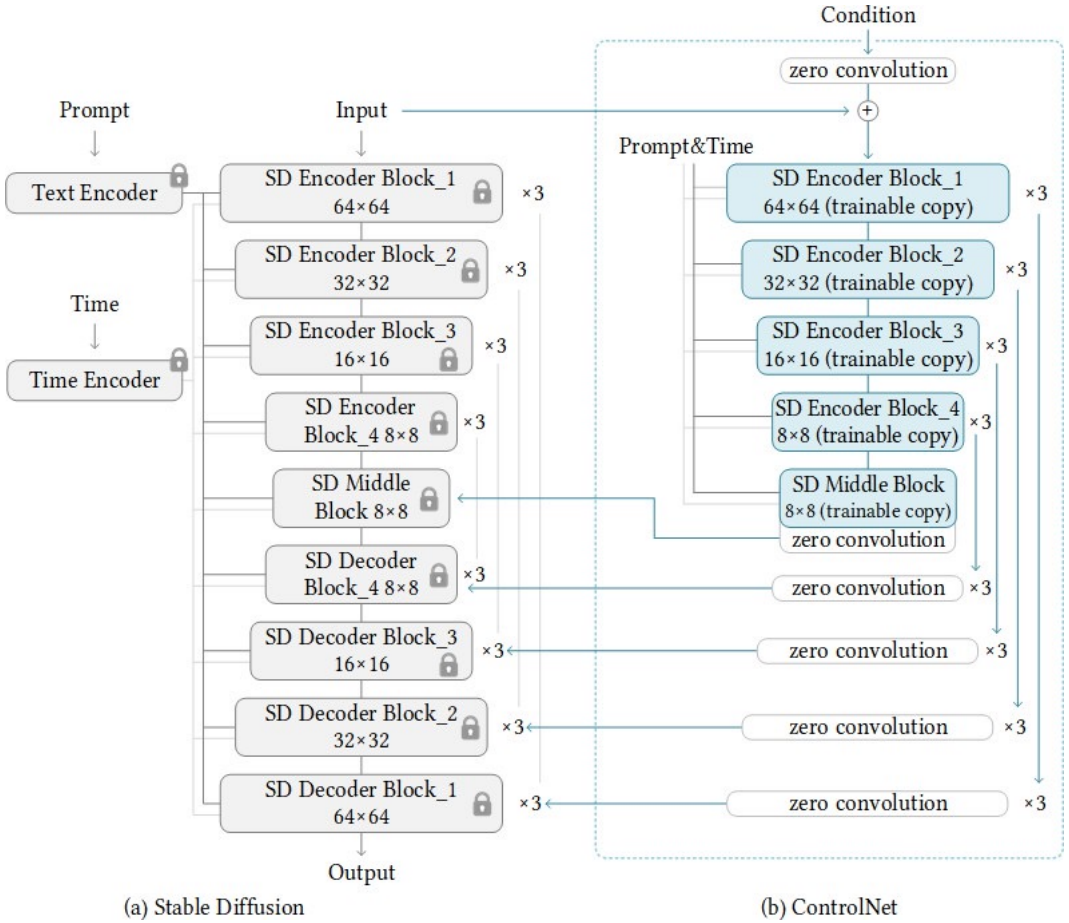
Leveraging Class Labels or Prompts

Classifier Guidance / Classifier-Free Guidance

Conditional Generation

ControlNet





Stylization

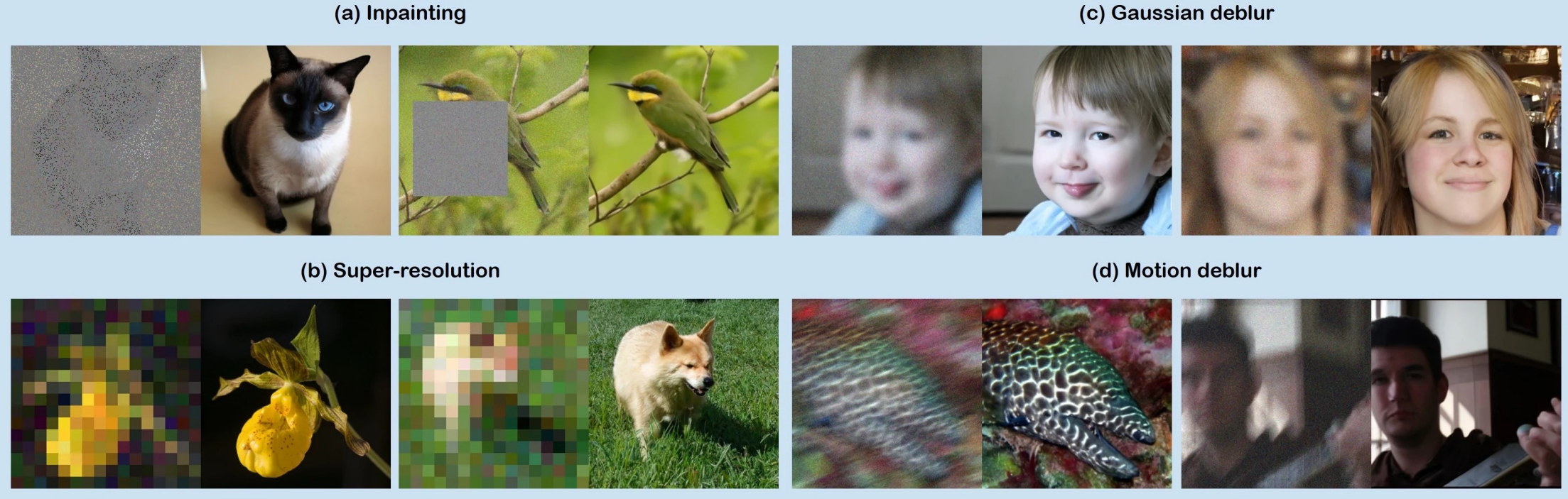
LoRA

Personalization

DreamBooth

Inverse Problems

Diffusion Posterior Sampling

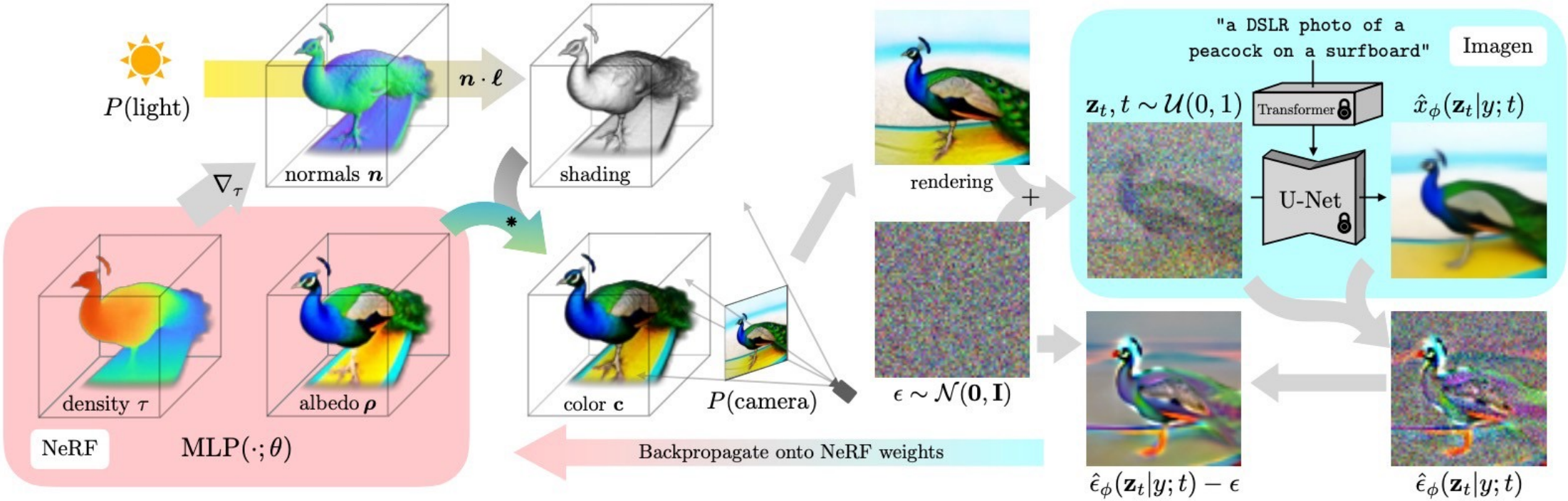


Video Editing

Chen, Ting-Hsuan, et al. "NaRCan: Natural Refined Canonical Image with Integration of Diffusion Prior for Video Editing." *NeurIPS*. 2024.

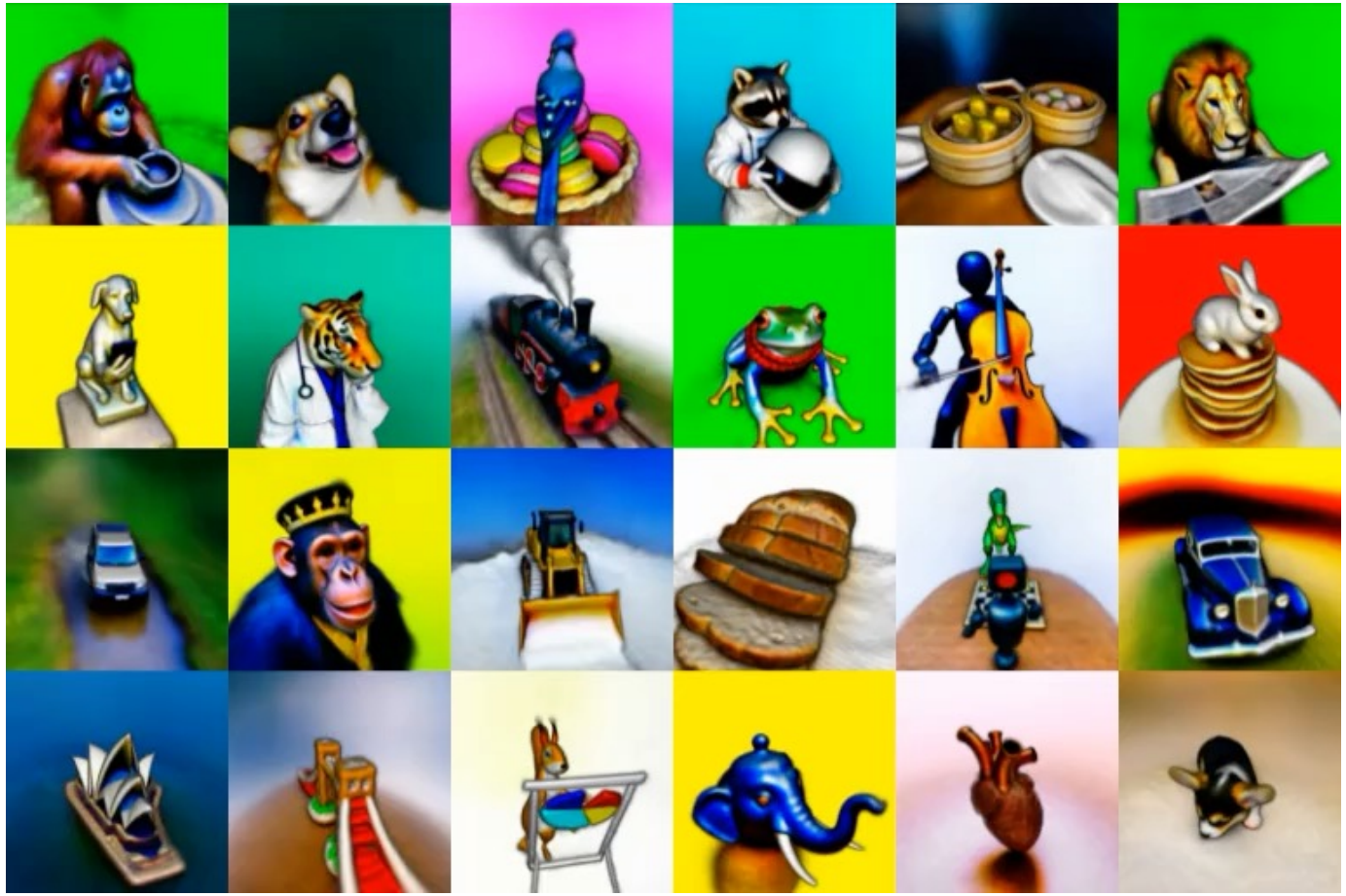
Knowledge Distillation

Score Distillation Sampling



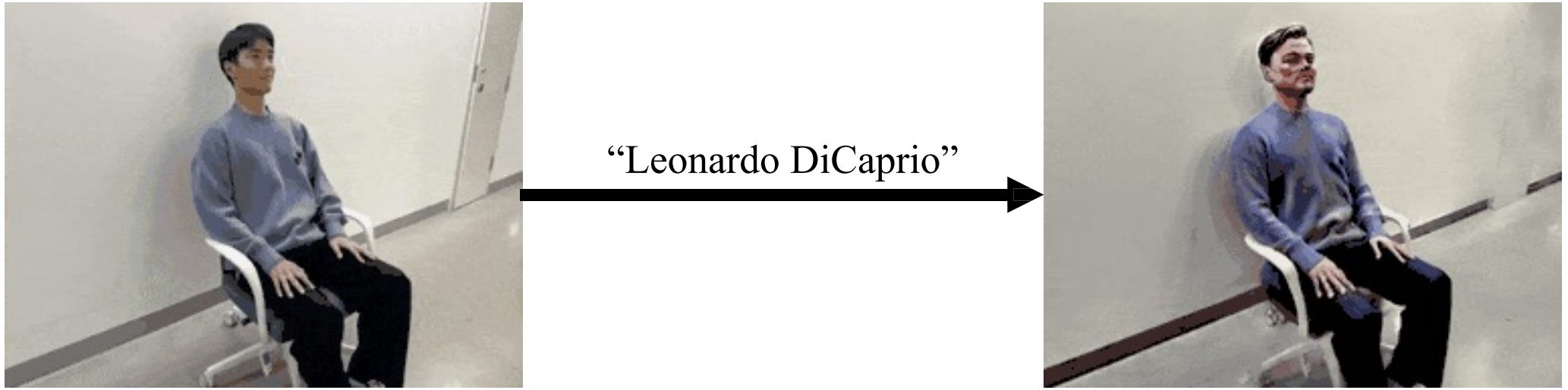
Knowledge Distillation

Score Distillation Sampling

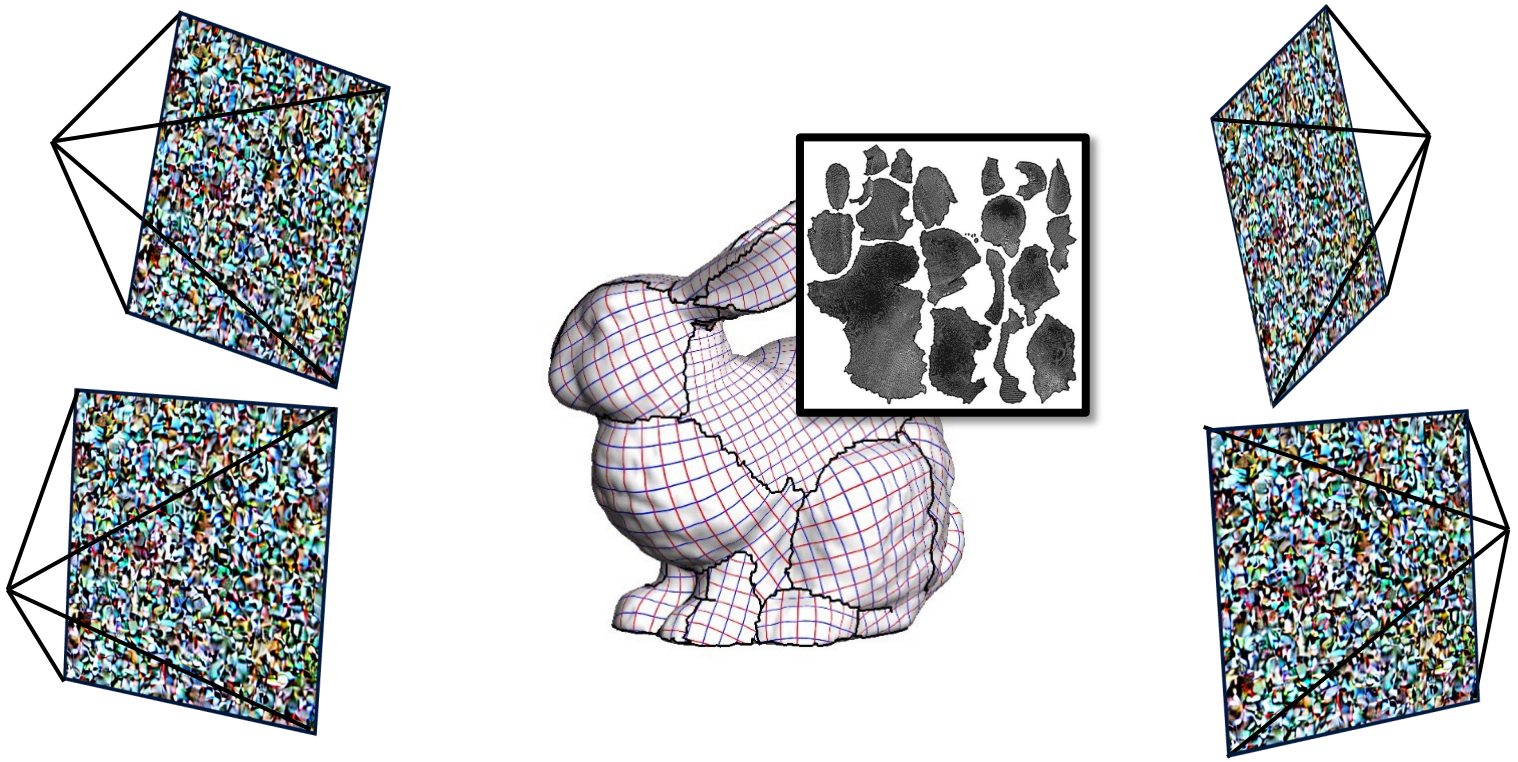


Knowledge Distillation

Posterior Distillation Sampling

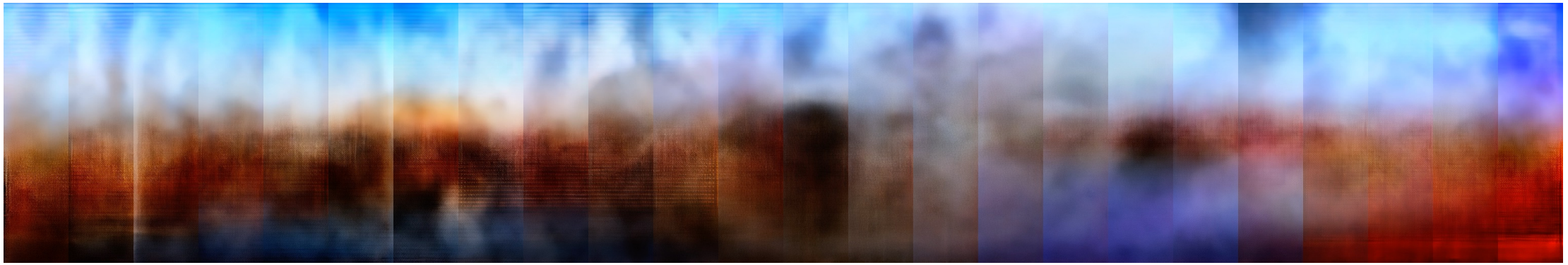


Joint Denoising



Joint Denoising

SyncDiffusion



Joint Denoising

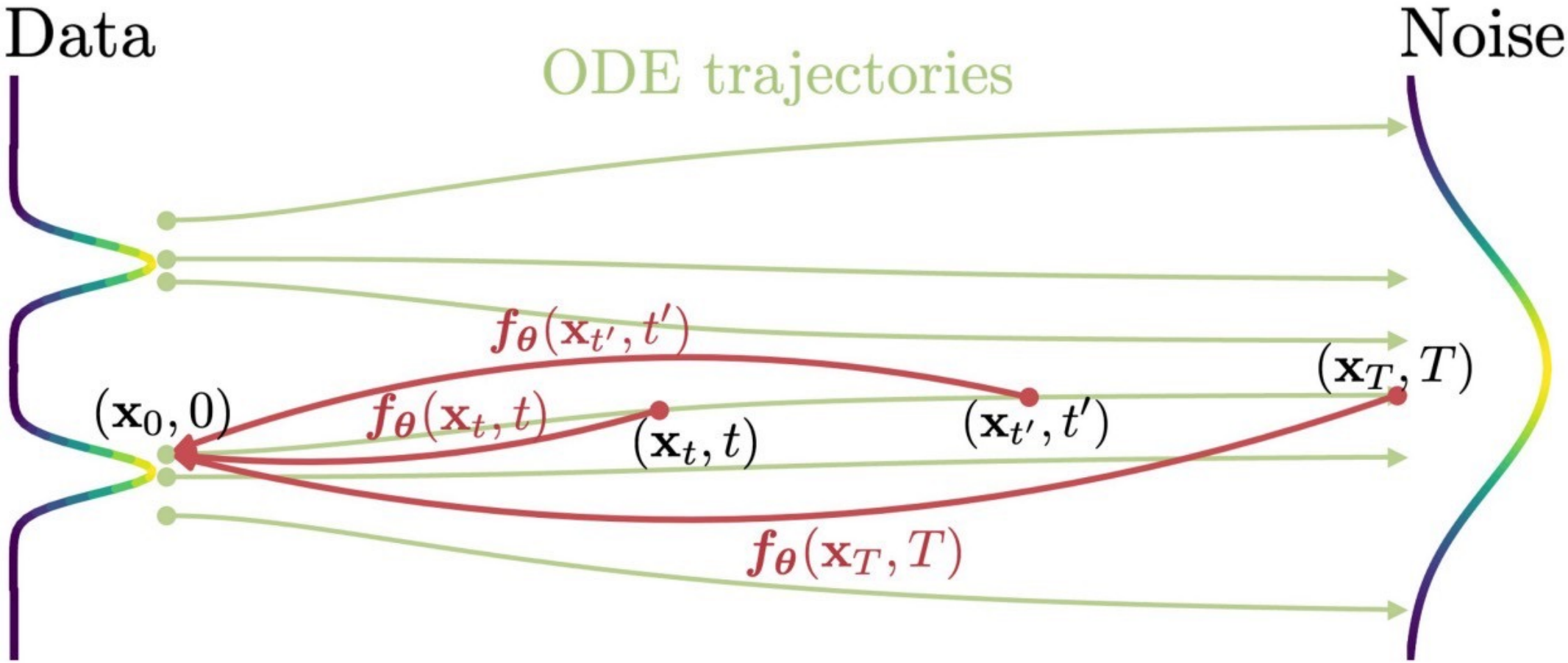
SyncTweedies

"A Chinese style lantern"

"A car with graffiti"

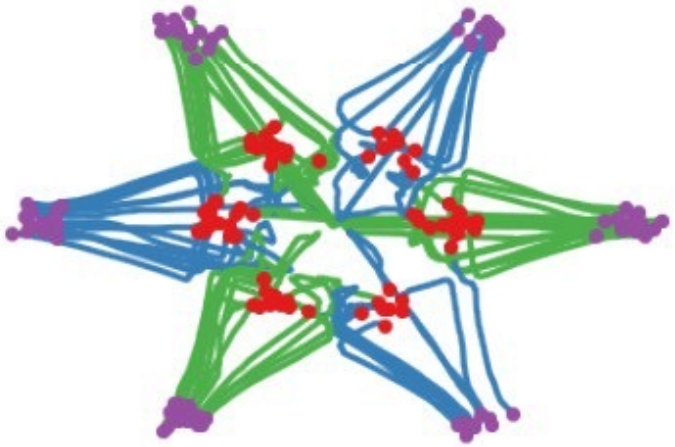
Few-Step Generation

Consistency Models



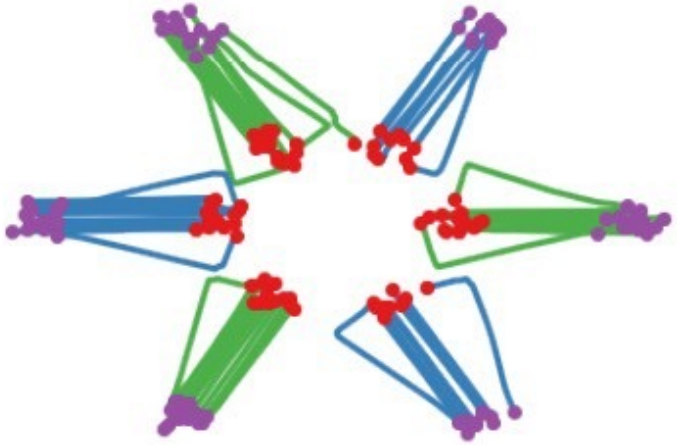
Rectified Flow

ReFlow



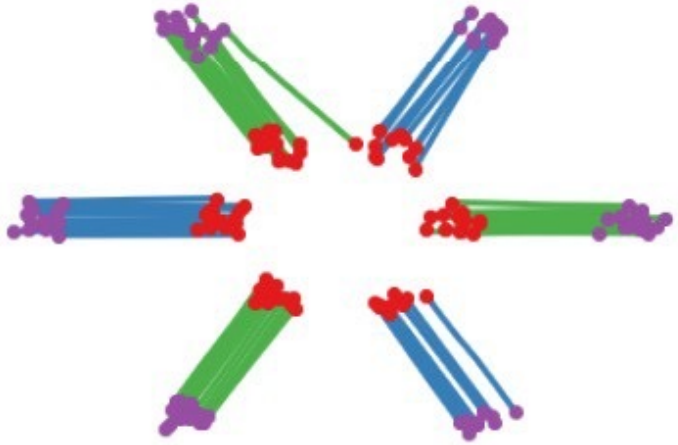
(a)The 1st rectified flow $\boldsymbol{Z}^1$

$\boldsymbol{Z}^1 = \texttt{RectFlow}((X_0, X_1))$



(b) Reflow $\boldsymbol{Z}^2$

$\boldsymbol{Z}^2 = \texttt{RectFlow}((Z_0^1, Z_1^1))$



(c) Reflow $\boldsymbol{Z}^3$

$\boldsymbol{Z}^3 = \texttt{RectFlow}((Z_0^2, Z_1^2))$

Topics

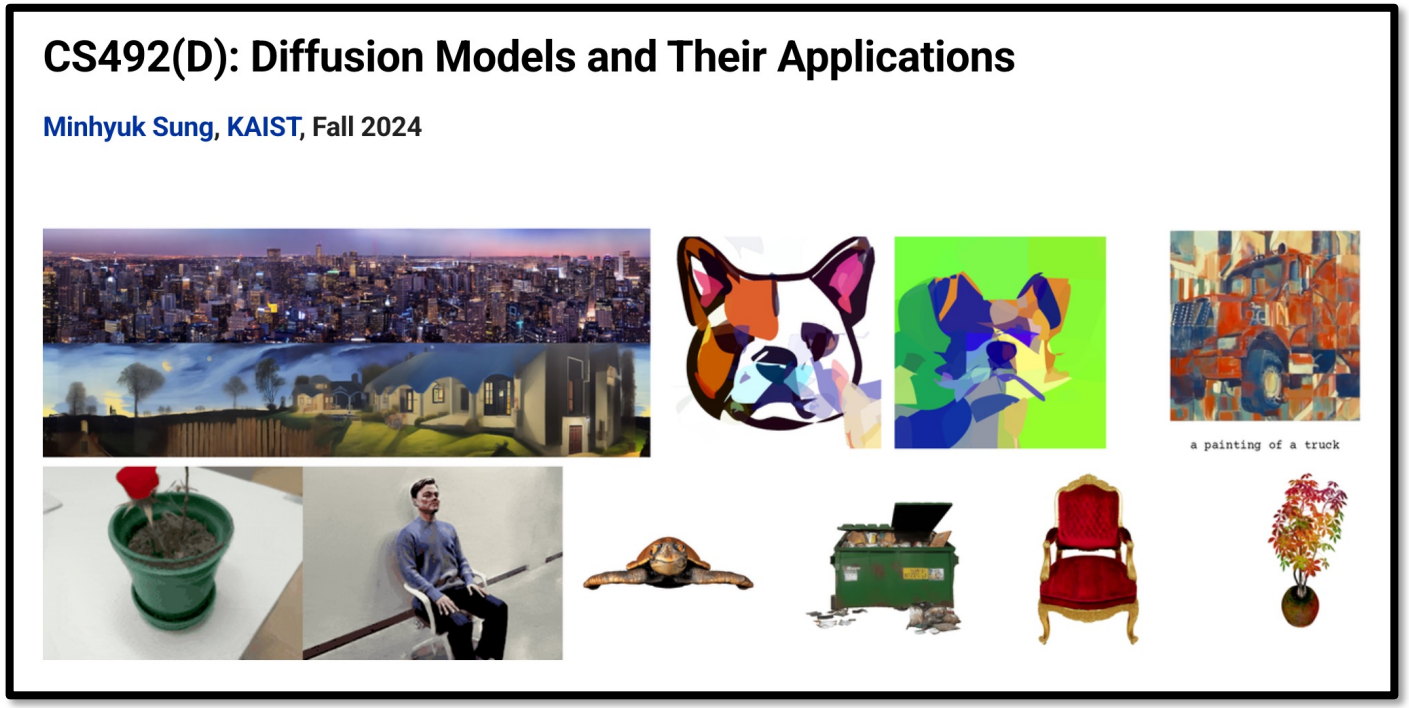
- Background of Generative Models
- DDPM / DDIM / Score-Based Models
- CFG / Latent Diffusion
- Conditional Generation
- Stylization / Personalization
- Inverse Problem
- Knowledge Distillation
- Diffusion Synchronization
- SDE/ODE Solvers
- Consistency Models / Flow-Based Models
- DiT / Applications / Future of Generative Models

Resources

KAIST, Fall 2024, CS492(D):

Diffusion Models and Their Applications

https://mhsung.github.io/kaist-cs492d-fall-2024/

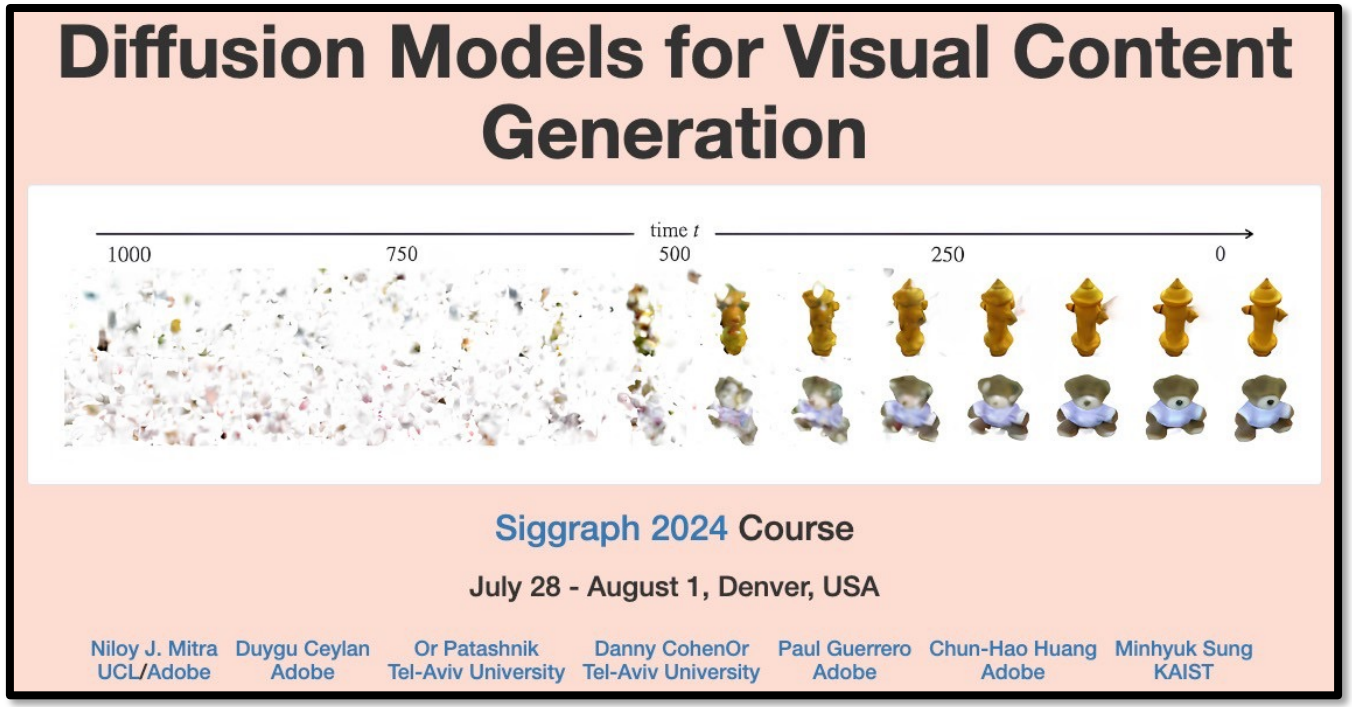


Resources

SIGGRAPH 2024 Course:

Diffusion Models for Visual Content Generation

https://geometry.cs.ucl.ac.uk/courses/diffusion4ContentCreation_sigg24/



Resources

Tutorials

- CVPR 2023: Denoising Diffusion Models: A Generative Learning Big Bang.
- CVPR 2024: 3D/4D Generation and Modeling with Generative Priors.
- CVPR 2024: Diffusion-based Video Generative Models.
- CVPR 2025: From Video Generation to World Model.

Blogs

- "Generative Modeling by Estimating Gradients of the Data Distribution", Yang Song.
- "What are Diffusion Models?", Lilian Weng.
- "Understanding Diffusion Models: A Unified Perspective". Calvin Luo.
- "Tutorial on Diffusion Models for Imaging and Vision". Stanley H. Chan.
- "Step-by-Step Diffusion: An Elementary Tutorial". Preetum Nakkiran, Arwen Bradley, Hattie Zhou, and Madhu Advani.

In this course…

We will discuss diffusion models, covering both their theoretical foundations and practical applications.

Prerequisite

- Background in machine learning/deep learning. We'll specifically focus on diffusion models (while briefly discussing the background of generative models).
- Experience with neural network implementation. There will be programming assignments and a final project.
- Recommended prior courses:
  - 線性代數
  - 機率
  - 微分方程
  - 深度學習+深度學習實驗

Week	Date	Topic	Assignments
1	2025-09-04	GAN / VAE	
2	2025-09-11	DDPM	#1 – DDPM
3	2025-09-18	DDIM	
4	2025-09-25	CFG / Latent Diffusion / ControlNet / LoRA / Zero-Shot Applications	#2 – DDIM & LoRA
5	2025-10-02	DDIM Inversion / Score Distillation **(attending CoRL 2025)**	
6	2025-10-09	Diffusion Synchronization / Inverse Problems	#3 – Distillation
7	2025-10-16	Probability Flow ODE / DPM-Solver	
8	2025-10-23	Flow Matching **(attending ICCV 2025)**	#4 – Flow Matching
9	2025-10-30		
10	2025-11-06	Final Project Proposal	
11	2025-11-13		
12	2025-11-20		
13	2025-11-27	Paper Presentation	
14	2025-12-04	Guest Lecture – Ta-Ying (Tim) Cheng	
15	2025-12-11	Guest Lecture – Chieh (Hubert) Lin	
16	2025-12-18	Guest Lecture – Chih-Hao Lin	
17	2025-12-25	Final Project Presentation	

Lectures

- 100 mins IN-PERSON lecture + 50 mins ONLINE recorded video.

- Recordings will be available on E3.

Guest Lectures

Ta-Ying Cheng
Research scientist
@ Netflix
Topic: Material editing

Chieh Hubert Lin
Research scientist
@ Architectural design startup
Topic: 3D inpainting

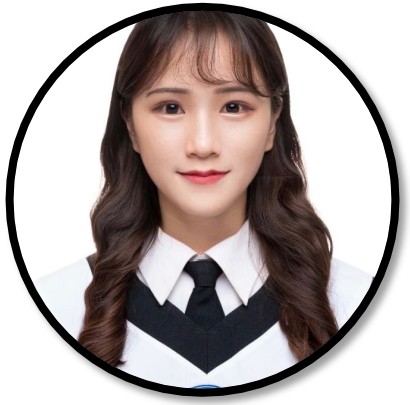
Chih-Hao Lin
3rd year PhD student
@ UIUC
Topic: Weather Synthesis

Course Logistics

Yu-Lun Liu (Instructor)

- Assistant Professor in the Department of Computer Science.
- Homepage: https://yulunalexliu.github.io/
- Office: EC713.

Teaching Assistants



黃怡川（資科工所博三）
yichuanh.cs12@nycu.edu.tw

李杰穎（資科工所博一）
jayinnn.cs14@nycu.edu.tw

陳映寰（資科工所碩二）
yinghuanchen.cs13@nycu.edu.tw

張欀齡（資科工所碩二）
siang1105.cs13@nycu.edu.tw

鄭淮薰（資科工所碩二）
huaish.cs13@nycu.edu.tw

柯柏旭（資科工所碩二）
hentci.cs13@nycu.edu.tw

Evaluation

- Programming Assignment: 40% (10% each)
- Paper Presentation: 15%
  - 10%: Hacker's Deliverables (code, demo, presentation)
  - 5%: 問問題
- Final Project: 40%
  - 10% Proposal Presentation
  - 15% Oral & Demo Presentation
  - 10% Final Report (4 pages in CVPR LaTeX format)
  - 5%: 問問題
- Participation in Guest Lectures: 5%

Programming Assignments

- Assignment1 (DDPM)
- Assignment2 (DDIM & LoRA)
- Assignment3 (Distillation)
- Assignment4 (Flow Matching)

Prerequisite

- You'll need basic programming skills in Python and PyTorch to complete the programming assignments.

- **Start the programming assignments as early as possible!**

Submission

- Each programming assignment is due **two weeks after** the assignment session.
- Submit your solutions on E3.
- **No late submissions allowed.** I.e., you will get zero credit even with only 1 minute late submitting deliverables on E3.

Paper Presentation

- 每組**四**位學生（請現在就開始找組員！）

- 報告的組別：報告近三年發表於 **CVPR/ICCV/ECCV/NeurIPS/ICLR/ICML/SIGGRAPH/SIGGRAPH Asia main conference** 的論文
  - 根據你報告的好壞我會給予評分（總成績 5%）
  - **Hacker's Deliverables**（總成績 5%）

- 其他所有組別：問問題
  - 在 presentation 中間或結束後皆可發問
  - 每個 presentation 我會選出最 insightful 的問題加分（每次總成績 1%）

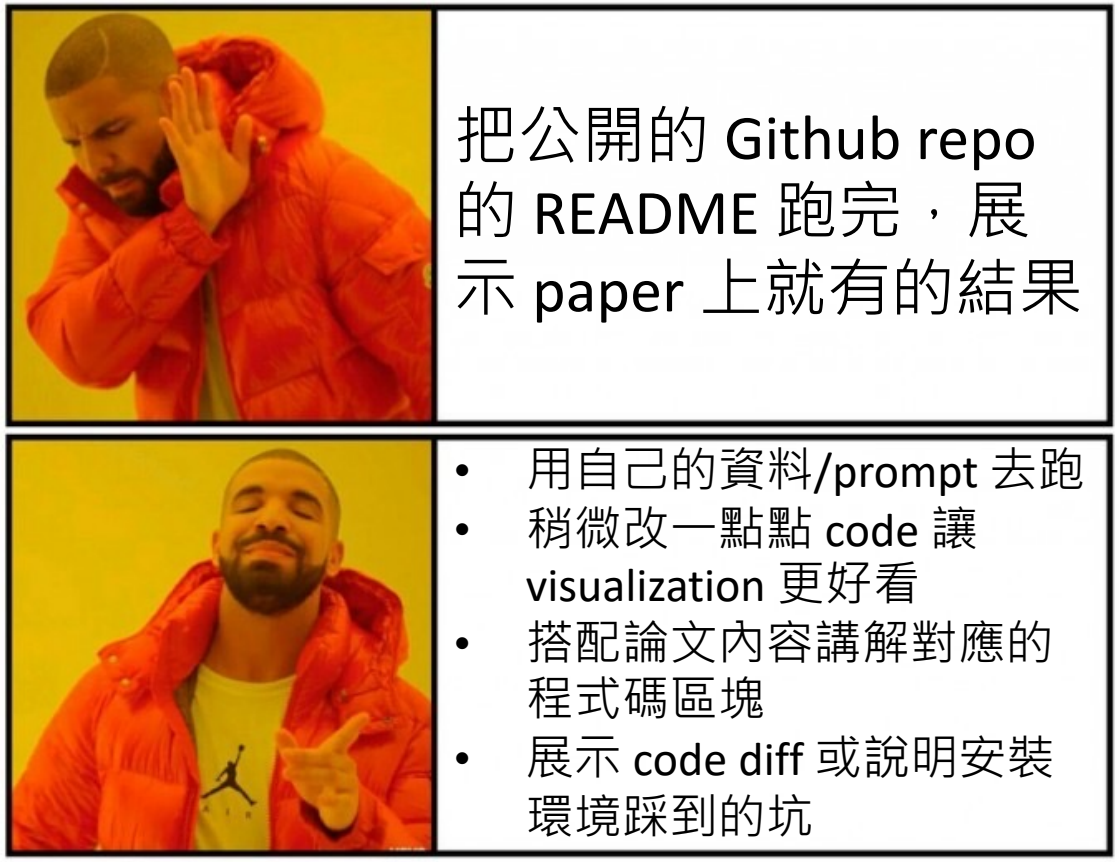
Paper Presentation – Hacker

報告的組別必須扮演一個需要盡快展示這篇論文的**駭客**

- 在你自己的資料/任務上展示這篇論文的結果
- 準備與班上同學分享演算法的核心程式碼並展示你的實作
- 不要只是下載並執行現有的實作

https://courses.grainger.illinois.edu/cs598shw/fa2024/slides/00_Intro.pdf

Paper Presentation – **Hacker**



Paper Presentation – **其他組別問問題**

不著邊際非常飄渺的問題：“這篇 paper 的東西能用在 video 上嗎？”

- “為什麼不用 X，要這麼複雜？好處是 Y 嗎？”
- “用 Z 演算法以及 A 改動，就可以 training-free 了嗎？”
- “這樣做會有 B 的問題，也許可以使用 C 方法解決，因為…”

Final Project

- 每組四位學生（與 Paper Presentation 組別相同）
- 三種選項：
  - [量變] 改善現有論文的演算法（生成品質、速度等等）
  - [質變] 創意開發新應用或演算法（影像 -> 視訊, training -> zero-shot）
  - [復現] Re-implement 近五年任何一篇沒有公開程式碼的頂尖會議論文

Final Project **Oral Presentation**

- 報告的組別：
  - 根據創意/實用度/完整度/報告的好壞等等我會給予評分（總成績 15%）
- 其他所有組別：扮演 reviewer 問問題
  - 在 presentation 中間或結束後皆可發問
  - 每個 presentation 我會選出最 insightful 的問題加分（每次總成績 1%）

https://www.computer.org/publications/tech-news/events/ieee-cvpr-conference-on-computer-vision-and-pattern-recognition-2019-paper-presentations

Final Project **Final Report**

四頁的 paper in CVPR LaTeX format

- Title
- Abstract
- Introduction
- Related Work
- Method
- Experiments
- Conclusion
- References

Robust Dynamic Radiance Fields

Yu-Lun Liu[2*] Chen Gao[1] Andreas Meuleman[3*] Hung-Yu Tseng[1] Ayush Saraf[1]
Changil Kim[1] Yung-Yu Chuang[2] Johannes Kopf[1] Jia-Bin Huang[1,4]
[1]Meta [2]National Taiwan University [3]KAIST [4]University of Maryland, College Park
https://robust-dynrf.github.io/

Figure 1. **Robust space-time synthesis from dynamic monocular videos.** Our method takes a casually captured video as input and reconstructs the camera trajectory and dynamic radiance fields. Conventional SfM system such as COLMAP fails to recover camera poses even when using ground truth motion masks. As a result, existing dynamic radiance field methods that require accurate pose estimation do not work on these challenging dynamic scenes. Our work tackles this *robustness* problem and showcases high-fidelity dynamic view synthesis results on a wide variety of videos.

Abstract

Dynamic radiance field reconstruction methods aim to model the time-varying structure and appearance of a dynamic scene. Existing methods, however, assume that accurate camera poses can be reliably estimated by Structure from Motion (SfM) algorithms. These methods, thus, are unreliable as SfM algorithms often fail or produce erroneous poses on challenging videos with highly dynamic objects, poorly textured surfaces, and rotating camera motion. We address this robustness *issue by jointly estimating the static and dynamic radiance fields along with the camera parameters (poses and focal length). We demonstrate the robustness of our approach via extensive quantitative and qualitative experiments. Our results show favorable performance over the state-of-the-art dynamic view synthesis methods.*

1. Introduction

Videos capture and preserve memorable moments of our lives. However, when watching regular videos, viewers observe the scene from fixed viewpoints and cannot interactively navigate the scene afterward. Dynamic view synthesis techniques aim to create photorealistic novel views of dynamic scenes from arbitrary camera angles and points of view. These systems are essential for innovative applications such as video stabilization [33, 42], virtual reality [7, 15], and view interpolation [13, 85], which enable free-viewpoint videos and let users interact with the video sequence. It facilitates downstream applications like virtual reality, virtual 3D teleportation, and 3D replays of live professional sports events.

Dynamic view synthesis systems typically rely on expensive and laborious setups, such as fixed multi-camera capture rigs [7, 10, 15, 50, 85], which require simultaneous capture from multiple cameras. However, recent advancements have enabled the generation of dynamic novel views from a single stereo or RGB camera, previously limited to human performance capture [16, 28] or small animals [65]. While some methods can handle unstructured video input [1, 3], they typically require precise camera poses estimated via SfM systems. Nonetheless, there have been

*This work was done while Yu-Lun and Andreas were interns at Meta.

AI Coding Assistant Tool Policy

- 你可以在 programming assignments 和 final project 中使用 AI coding tools，例如 ChatGPT、Copilot、Codex 和 Code Intelligence

- 但直接抄網路上或別人的程式碼仍然是嚴格禁止的
  - 學期成績零分，並向學校報告

Plagiarism

- 我們將使用 Turnitin 與 Moss 來比對原創性檢測與遏止抄襲

- Turnitin: https://www.turnitin.tw/

- Moss: https://theory.stanford.edu/~aiken/moss/

Computing Resource

本課程**不提供**任何運算資源，請使用個人/實驗室/付費雲端資源完成 programming assignments 與 final project

你會需要至少一張有 12 GB VRAM 的 NVIDIA GPU