AI-Driven Immersive Learning:
The Future of Metaverse & Education

Dr. Yung-Hui Li

Director
AI Research Center
Hon Hai Research Institute

Outlines

1. GenAI & XR
2. Technical Bottlenecks
3. Overview of AI-Generated Images
4. PRO-U-GAT-IT
5. Conclusion

Rise of Generative AI

hashing for identifying duplicates, which is generally effective for most
However, for extremely large files or specific requirements, you might
the hashing method or consider efficiency improvements.

python

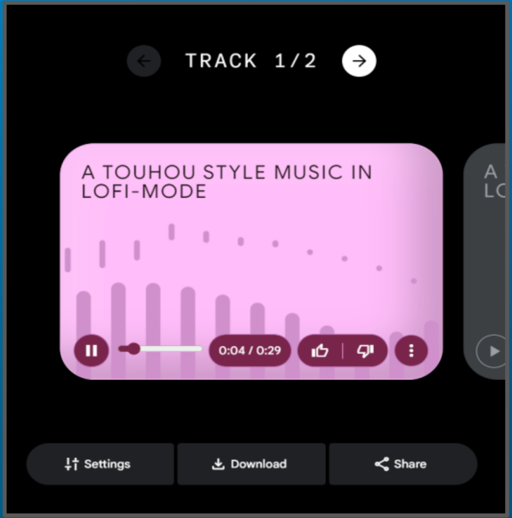
```python
import hashlib
import os

def file_hash(filepath):
    """Compute the SHA-256 hash of a file."""
    sha256_hash = hashlib.sha256()
    with open(filepath, "rb") as f:
        # Read and update hash in chunks of 4K
        for byte_block in iter(lambda: f.read(4096), b""):
            sha256_hash.update(byte_block)
    return sha256_hash.hexdigest()

def remove_duplicates(folder_path):
    """Remove duplicate files in the specified folder."""
    files_hash = {}
    for dirpath, _, filenames in os.walk(folder_path
```

Large Language Model

Based on Transformer technology, LLM is able to answer user's questions or generate articles.



Music LM

For music generation, creating melodies, harmonies, and even complete compositions.

Video Generation Model

Prompt: A cat "singing" opera with full orchestra, looking surprisingly profound.

Progress in AR/VR

Apple Vision Pro is a revolutionary spatial computing device that seamlessly integrates digital content into the physical world, allowing users to stay in the moment and connected to others.

Progress in AR/VR

FOXCONN 鴻海科技集團

Vision Pro users can turn any space into a personal cinema with a screen that feels 30 meters wide. The combination of virtual and real enhances the metaverse experience.

Future Growth of AR/VR

- **Shipments of AR/VR products are expected to increase significantly. This growth can be attributed to several key factors:**
  - **Hardware Improvement: With the advancement of AR/VR technology, hardware devices are becoming more lightweight and powerful.**
  - **Prices drop: As production scales up and R&D costs decrease, consumers can get these devices at a more affordable price.**
  - **Popularization of software: Metaverse, education, gaming, etc., these software are attracting more consumers and businesses to use AR/VR products.**

Applications of GenAI in the metaverse

FOXCONN 鴻海科技集團

Scene Generation

Diffusion Model

Customize the generated scene to meet the needs of users anytime, anywhere.

Real-Time conversation generation

Large Language Models

Design and generate dialog in real-time for NPC (Non-Player Character) in metaverse

AR / VR for Education

- **Virtual field trips**
  - With VR, students can visit far-flung places or historical scenes, such as ancient Egyptian pyramids, space stations, or ancient events, in a virtual environment. It stimulates students' curiosity in the study.
- **Interactive teaching**
  - AR and VR can be used to create interactive science experiments, explore complex scientific concepts and simulations, and enable students to experiment in a safer virtual environment.

Teaching Courses in VR

Source: https://vhil.stanford.edu/downloads/comm166

GenAI for Education (1/2)

- **Personalized Learning**
  - Generate customized learning materials and curriculum according to each student's unique needs.
- **Interactive Teaching**
  - Generate virtual TA avatars with GenAI technology, and use AR/VR and other devices to accompany students to read and learn, and improve learning efficiency and engagement.

GenAI for Education (2/2)

Analysis and Assessment for Student's Achievement

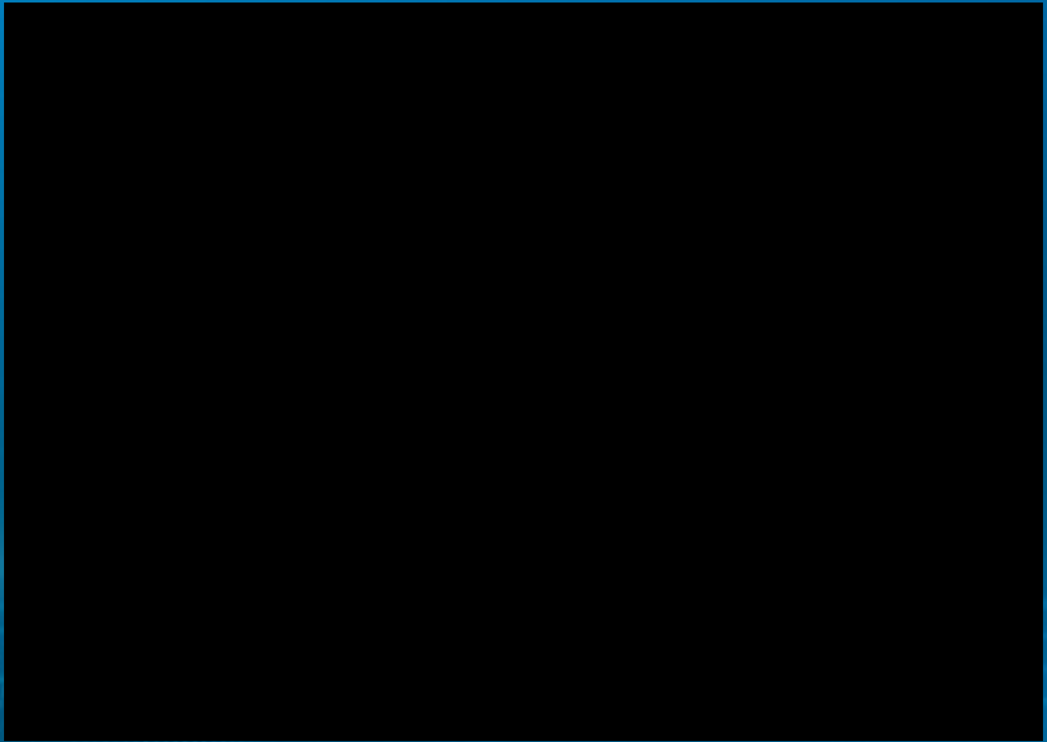
- Virtual teaching assistants can analyze student learning data, providing insights into learning progress, comprehension, and potential difficulties.

Metaverse for Education

Foldit

- Participants play games to contribute to scientific research
- A new protein structure for the treatment of AIDS was discovered, a breakthrough achieved by 60,000 participants in 10 days

Demo（Foldit from Univ of Washington）



Source: https://www.youtube.com/watch?v=CtI7qpsoFqM&ab_channel=Foldit

Technical Bottlenecks

Technical Bottleneck of AR/VR

- **Require huge amount production for 2D/3D objects**
  - Creating high-quality AR/VR contents cost a lot of **time and money**.
  - Highly realistic 3D environments and objects need to be created by **professional** designers and developers, which is not only costly but also time-consuming.

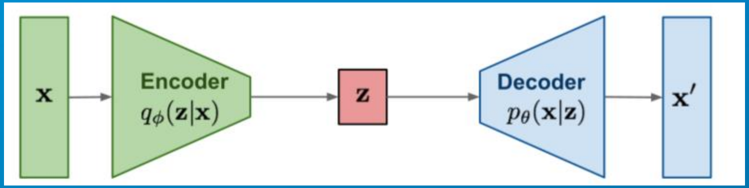
Technical Bottleneck of AR/VR

- **Limitations on computing resources**
  - Even working with AI, generating and processing high-resolution images in real-time requires significant computational resources.

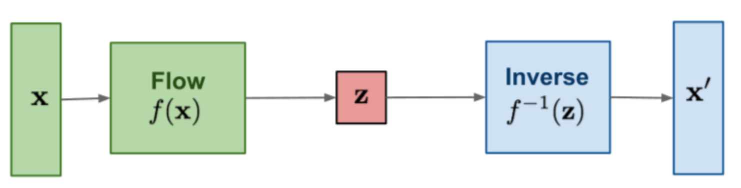
Overview of the Technology of AI-Generated Images

Deep Learning Models for AI-Generated Images

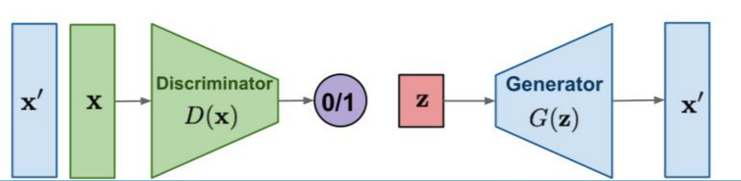
1. VAE



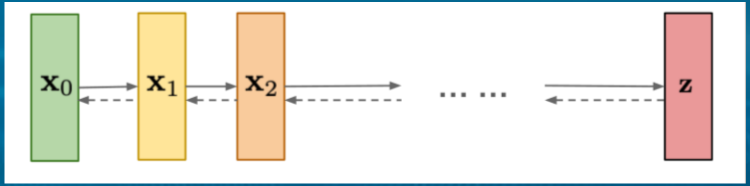
2. NF

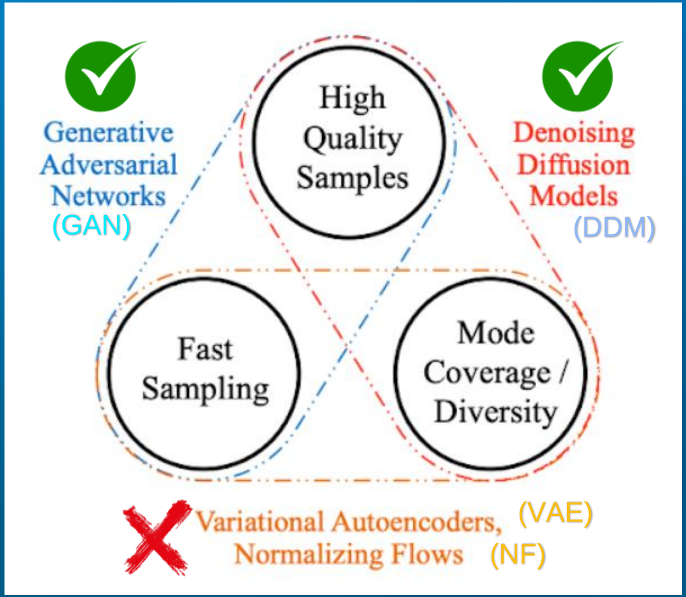


3. GAN



4. DDM



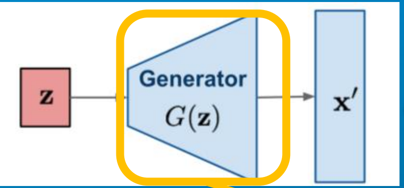


Paradigm	Quality	Diversity	Speed
GAN	✓	✗	✓
Diffusion	✓	✓	✗

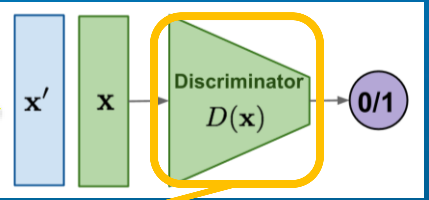
https://developer.nvidia.com/zh-cn/blog/improving-diffusion-models-as-an-alternative-to-gans-part-1/

Evolution of GAN (Generative Adversarial Networks)

1. **vanilla-GAN**
(unconditional)



Fight with each other!



$$L = \mathbb{E}_{z \sim p_z}\left[\log\left(1 - D\left(G\left(z, y\right), y\right)\right)\right] + \mathbb{E}_{x \sim p_{\text{data}}}\left[\log D\left(x, y\right)\right]$$

MNIST	LFW	Cifar10
Generated / GT	Generated / GT	Generated / GT

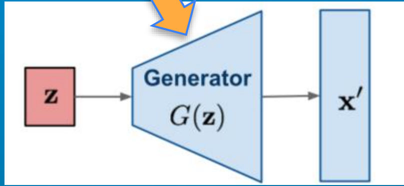
https://iopscience.iop.org/article/10.1088/2632-2153/ad1f77/pdf

Evolution of GAN (Generative Adversarial Networks)

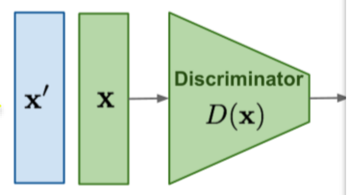
2. **Info-GAN**

(Conditional)

Class Label



Fight with each other!



Class Label

A task-oriented classifier is added to supervise the generation of corresponding class-specific data

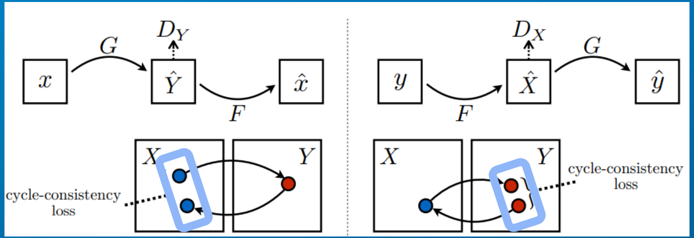
(a) Azimuth (pose)

(b) Presence or absence of glasses

Evolution of GAN (Generative Adversarial Networks)

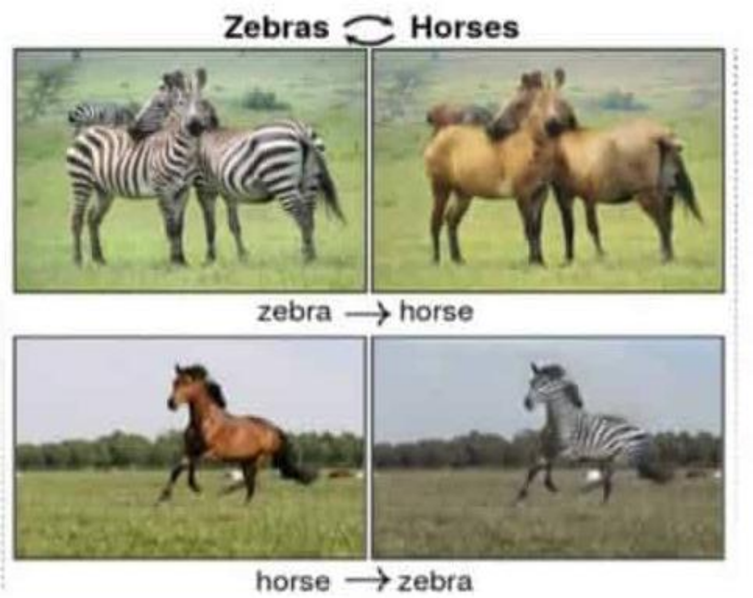
3. Cycle-GAN

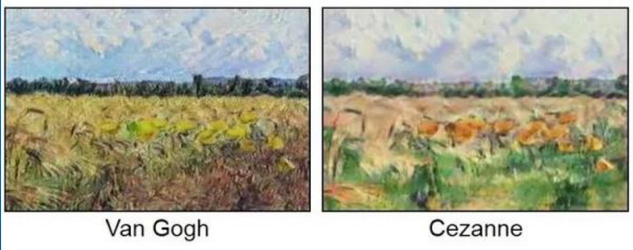
$$\mathcal{L}_{\text{cycle}}(G_{XY}, G_{YX}) = \mathbb{E}_{x\sim p_{\text{data}}}\left[\|G_{YX}(G_{XY}(x)) - x\|_1\right] + \mathbb{E}_{y\sim p_{\text{data}}}\left[\|G_{XY}(G_{YX}(y)) - y\|_1\right]$$



Cross Domain Generation

Style Transfer





https://iopscience.iop.org/article/10.1088/2632-2153/ad1f77/pdf

Diffusion Model

Image ← Noise

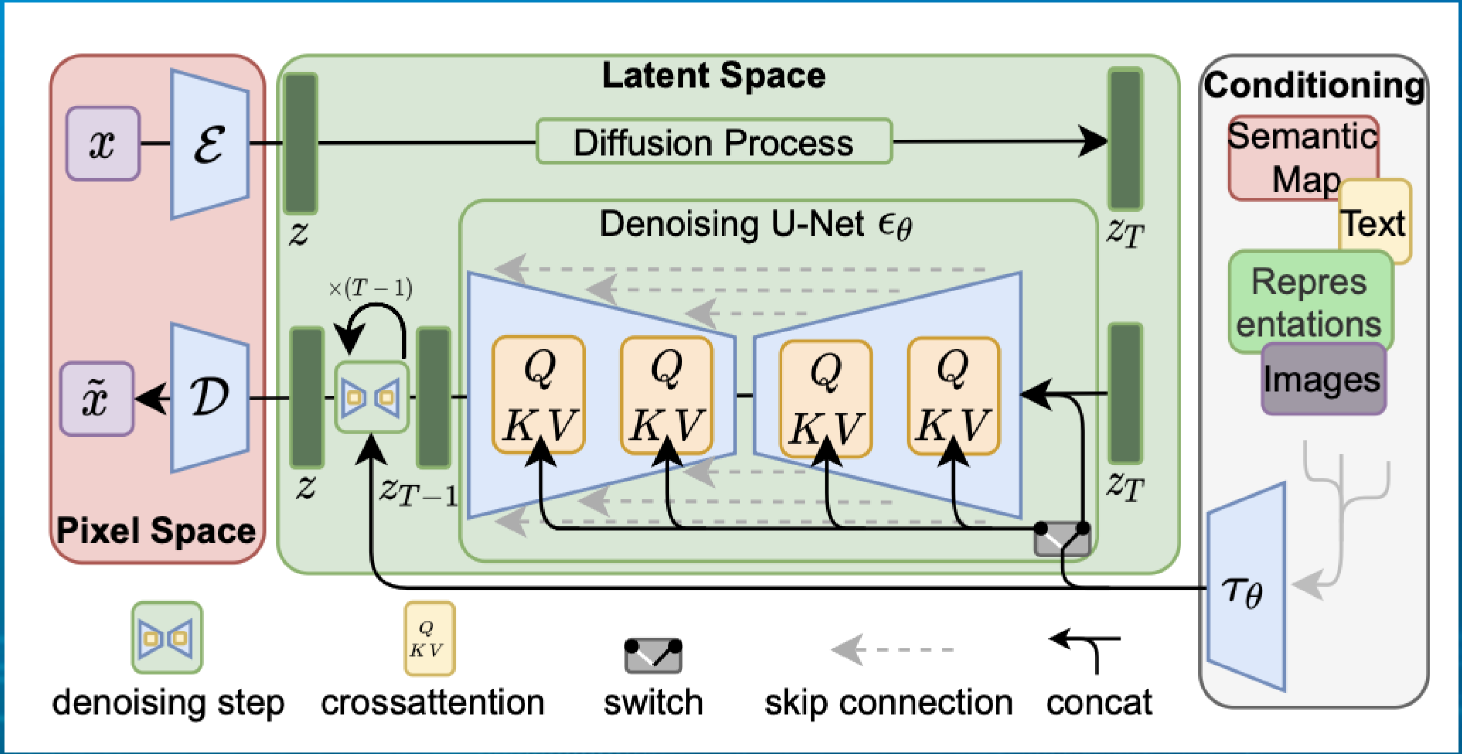
$x(0)$ $p_\theta(x_{t-1}|x_t)$ $x(T)$

$x(0)$ $x(T)$

$q(x_t|x_{t-1})$

Image → Noise

High-Resolution Image Synthesis with Latent Diffusion Models



[2112.10752] High-Resolution Image Synthesis with Latent Diffusion Models (arxiv.org)

Image to Image Translation: PRO-U-GAT-IT

Lee, H.-Y.; Li, Y.-H.;
Lee, T.-H.; Aslam, M.S. "Progressively Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization for Image-to-Image Translation," *Sensors* **2023**,*23*,6858.
https://doi.org/10.3390/s23156858

sensors MDPI

Article

Progressively Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization for Image-to-Image Translation

Hong-Yu Lee [1], Yung-Hui Li [2], Ting-Hsuan Lee [1] and Muhammad Saqlain Aslam [2,*]

1 Department of Computer Science and Information Engineering, National Central University, Taoyuan 32001, Taiwan; sam3u7858@gmail.com (H.-Y.L.); s109522056@g.ncu.edu.tw (T.-H.L.)
2 AI Research Center, Hon Hai Research Institute, Taipei 114699, Taiwan; yunghui.li@foxconn.com
* Correspondence: saqlain.msa@foxconn.com

Abstract: Unsupervised image-to-image translation has received considerable attention due to the recent remarkable advancements in generative adversarial networks (GANs). In image-to-image translation, state-of-the-art methods use unpaired image data to learn mappings between the source and target domains. However, despite their promising results, existing approaches often fail in challenging conditions, particularly when images have various target instances and a translation task involves significant transitions in shape and visual artifacts when translating low-level information rather than high-level semantics. To tackle the problem, we propose a novel framework called Progressive Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization (PRO-U-GAT-IT) for the unsupervised image-to-image translation task. In contrast to existing attention-based models that fail to handle geometric transitions between the source and target domains, our model can translate images requiring extensive and holistic changes in shape. Experimental results show the superiority of the proposed approach compared to the existing state-of-the-art models on different datasets.

Keywords: anime; cartoon styles; generative adversarial networks; image-to-image translation; style transfer

check for updates

Citation: Lee, H.-Y.; Li, Y.-H.; Lee, T.-H.; Aslam, M.S. Progressively Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization for Image-to-Image Translation. *Sensors* **2023**, *23*, 6858. https://doi.org/10.3390/s23156858

Academic Editor: Wataru Sato

Received: 13 June 2023
Revised: 16 July 2023
Accepted: 30 July 2023
Published: 1 August 2023

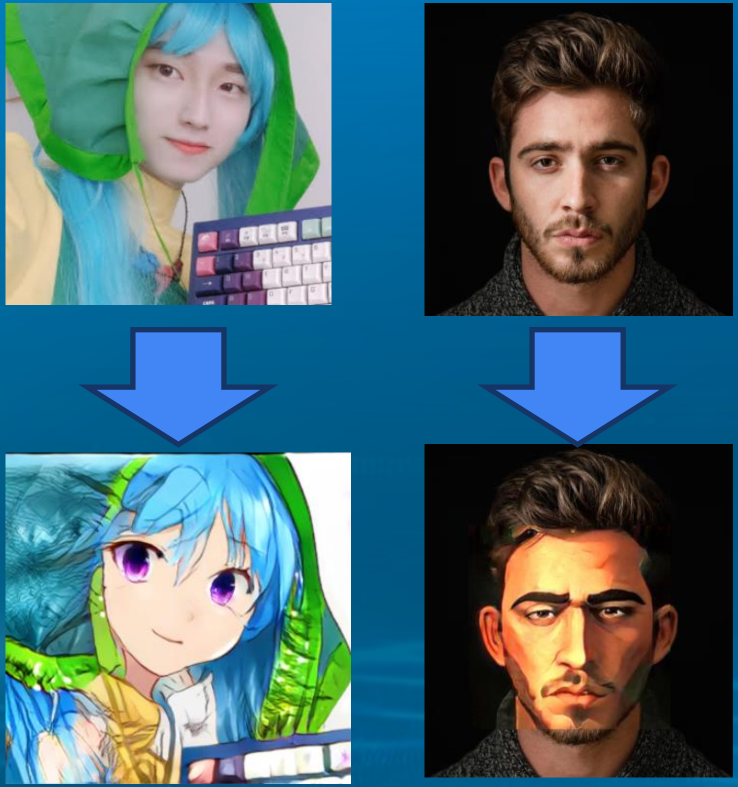
1. Introduction

In recent years, generative adversarial networks (GANs) have made significant progress in image-to-image translation. Researchers in machine learning and computer vision have given this topic considerable attention because of the wide range of practical applications available [1,2]. These include image inpainting [3,4], colorization [5,6], super- https://doi.org/10.3390/s23156858]. Image-to-image translation refers to a category of vision and graphics problems in which the goal is to learn the mapping between an input image (source domain) and an output image (target domain) from a set of aligned image pairs [11]. In the case of portrait stylization, various methods have been explored, such as self-to-anime [1] and cartoon [12]. There are, however, many tasks that will not offer paired training data. When paired data are provided, the mapping model can be trained using a conditional generative model [13–15] or a simple regression model [5,16,17] in a supervised manner.

Various works [18–25] have successfully translated images in unsupervised settings without available paired data by assuming shared latent space [22] and cycle consistency assumptions [11,21]. Nevertheless, supervised approaches require paired datasets for training, which can be laborious and expensive, if possible, to prepare manually. In contrast, unsupervised methods need a large volume of unpaired data and frequently need help to reach stable training convergence and generate high-resolution results [26].

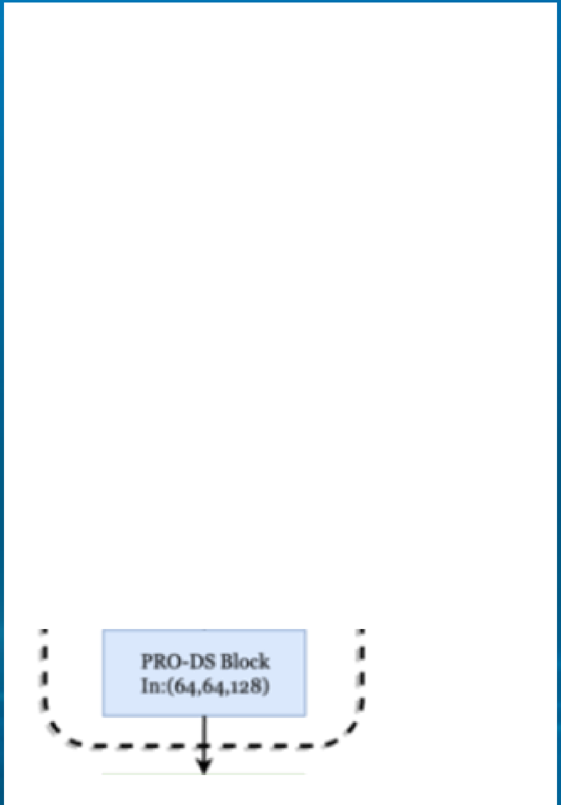
Introduction

- **A GAN variant for Image Style Transfer**
  - PRO-U-GAT-IT is a new kind of GAN, good for the task of **I2I** (Image to Image Translation).
  - The model is able to transfer the input image into another one with different style, while preserving the details of the content.

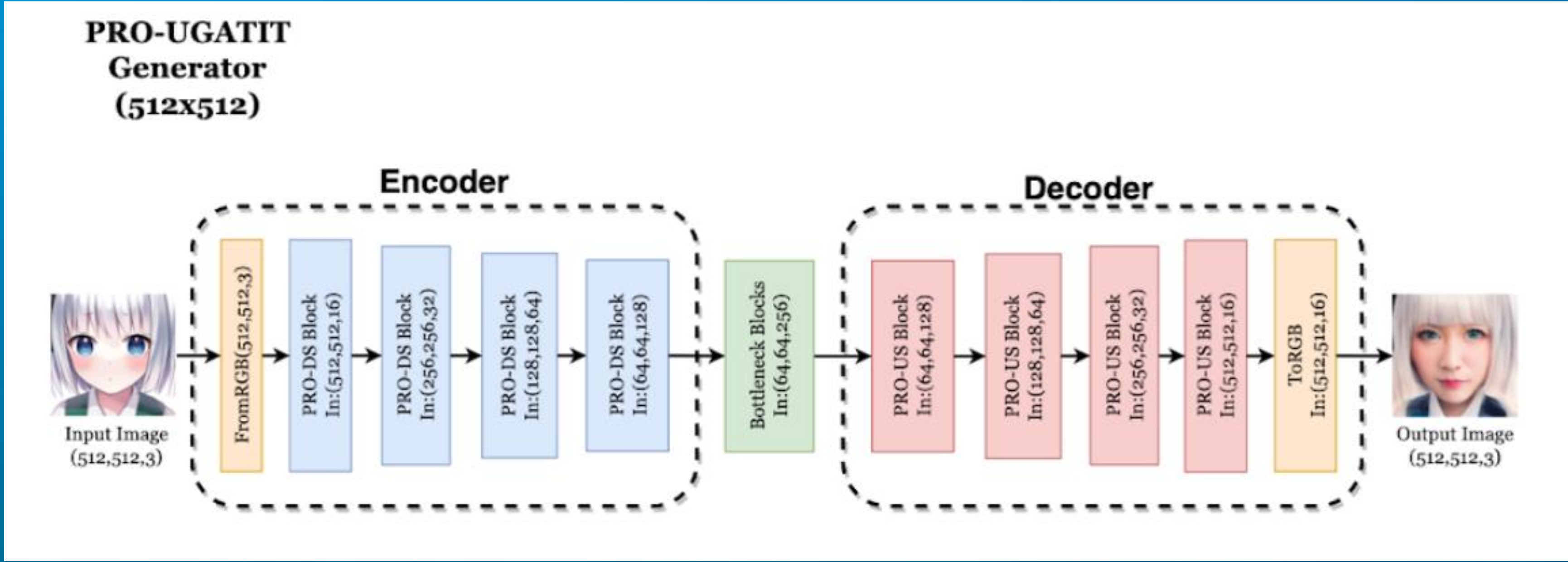


Introduction

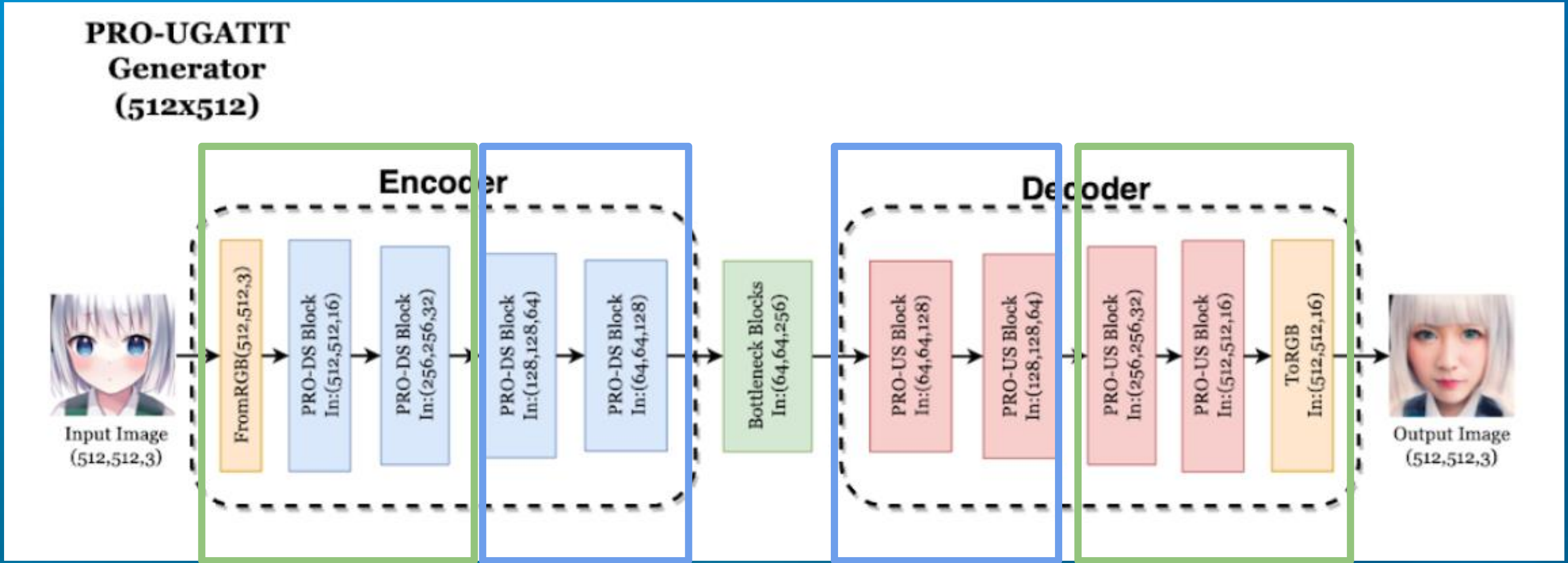
- **A New Design based on Modularization**
  - During the training stage, two modules (PRO-DS Block & PRO-US Block) are inserted into the structure <u>**incrementally and dynamically**</u> based on the computational requirement.



Generator Architecture



Generator Architecture



Learning
local features

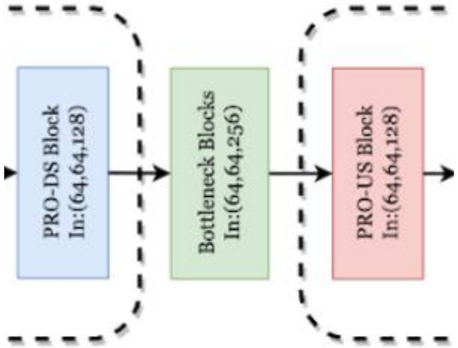
Learning
global feature

Generating
global pattern

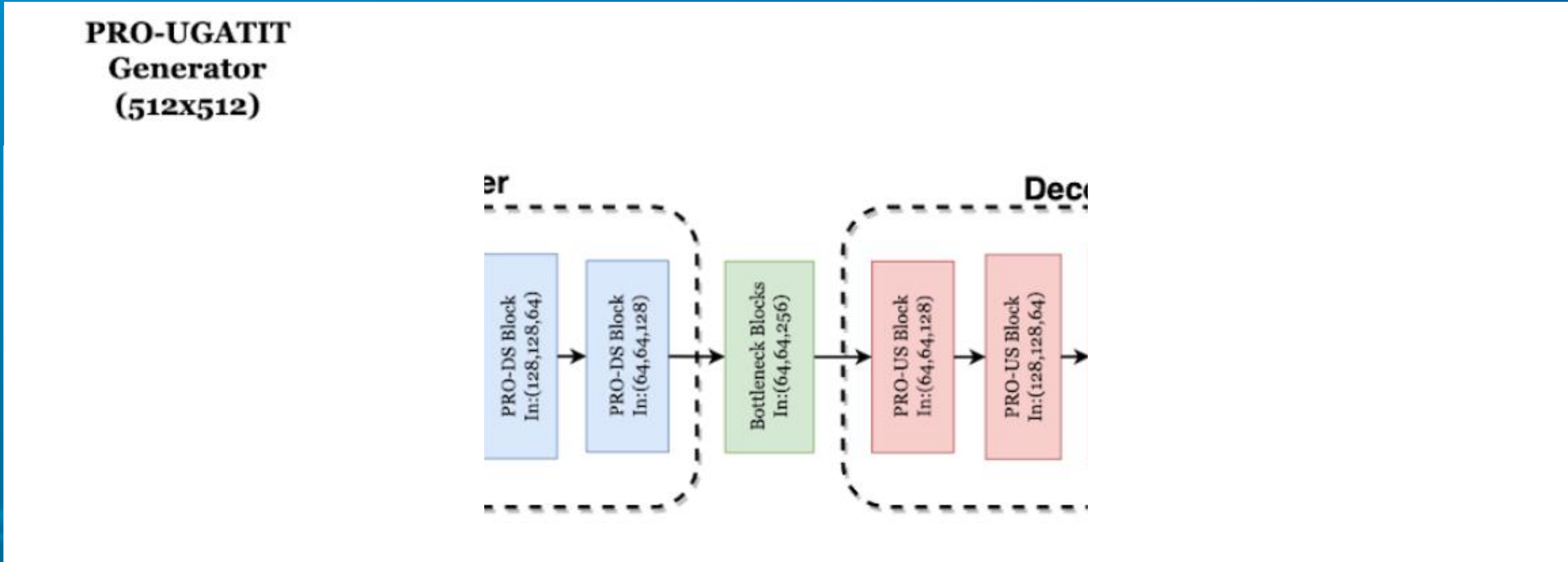
Generating
local pattern

Progressive Training

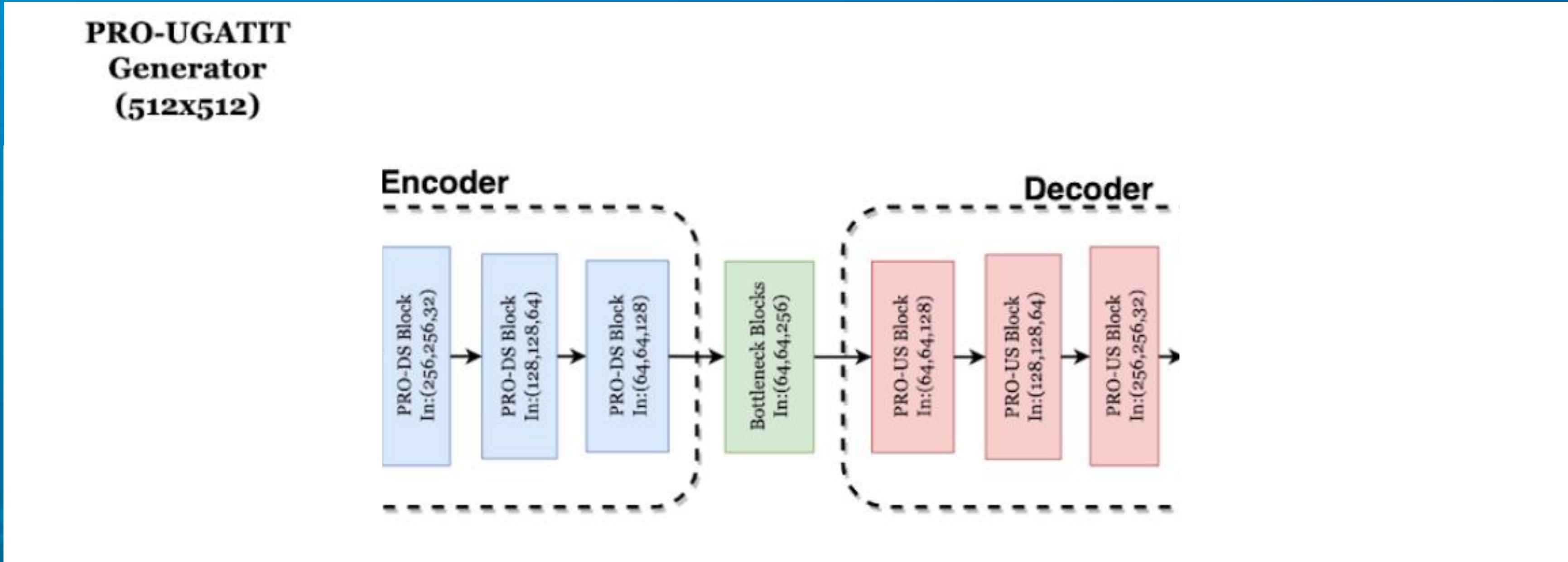
PRO-UGATIT
Generator
(512x512)



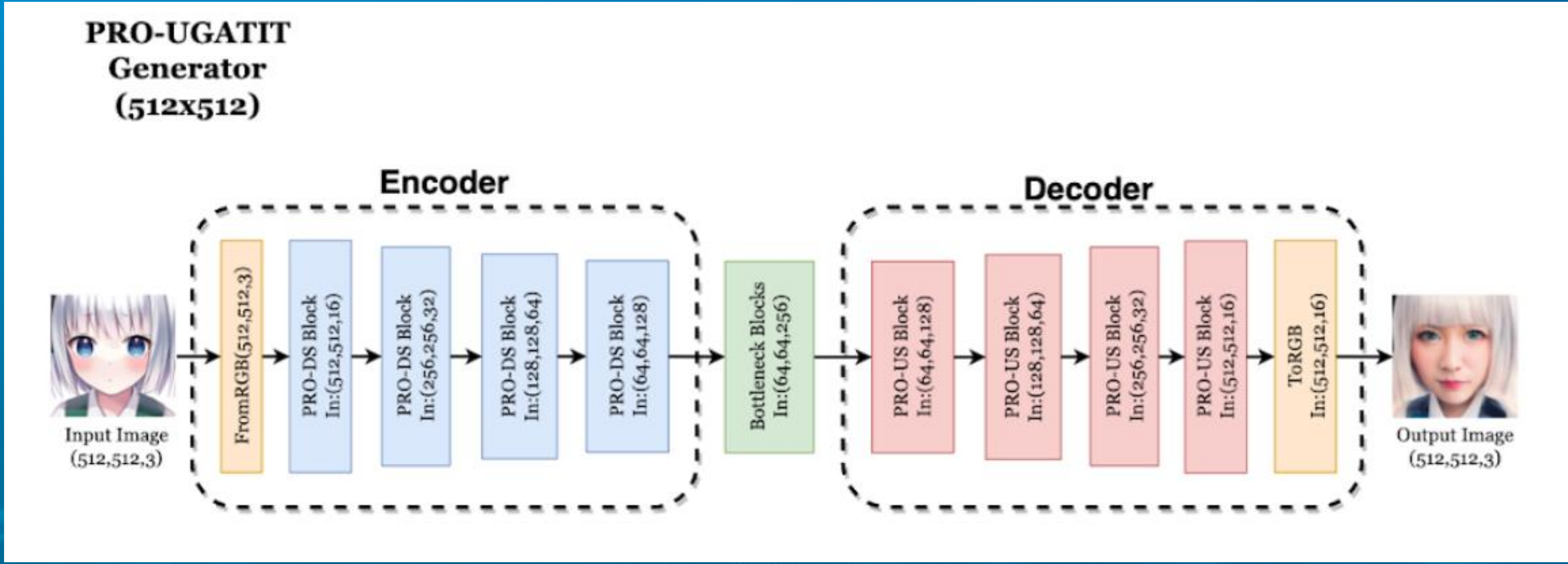
Progressive Training



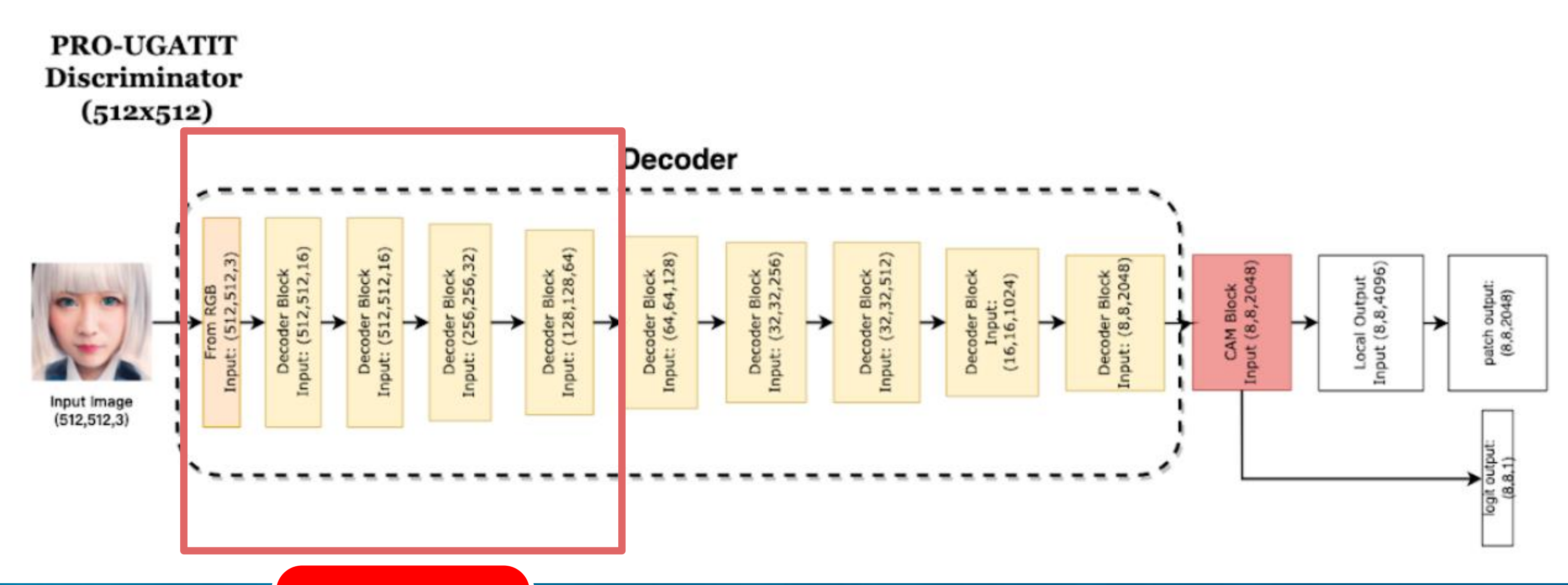
Progressive Training



Progressive Training



Discriminator Architecture



Progressively added into model

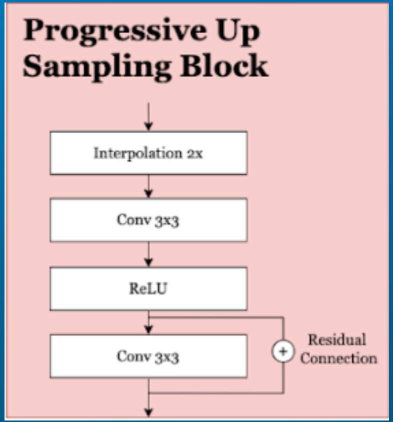
Contributions (1/2)

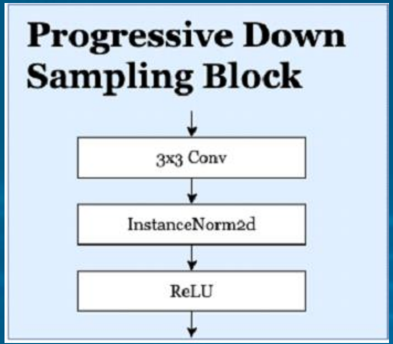
■ Progressive Training

- With progressive training, our model can start learning from lower-resolution images and gradually transition to higher-resolution images after it grows more capable.
- This training strategy helps the model better capture features from rough to detail, and can be fine-tuned according to needs at any stage to achieve better performance.

Contributions (2/2)

- **Flexible Architecture with Modular Design**
  - Flexibility and scalability: The modular design allows the model to be adapted and scaled according to different application requirements and image characteristics.
  - Such design strategy allows researchers or developers to add or remove modules based on specific tasks, improving the suitability and efficiency of the model

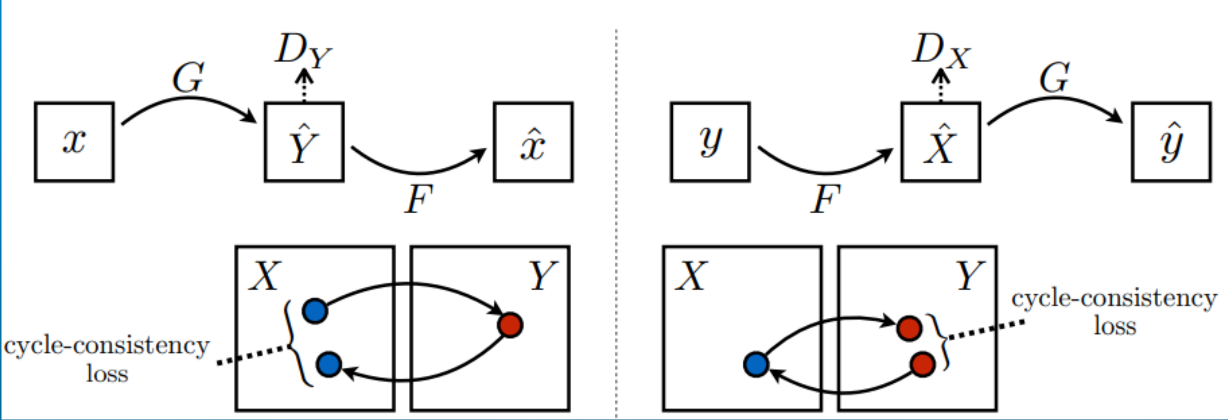




Loss Function (1)

Cycle Loss

- An image that goes through twice transformation (from source to target ; and from target to source) should be the same



$$L_{Cycle}^{s\to t} = E_{x\sim Xs}[\,|x - G_{t\to s}(\,G_{s\to t}(\,x)\,)\,|_1]$$

$$L_{Cycle} = L_{Cycle}^{s\to t} + L_{Cycle}^{t\to s}$$

Loss Function (2)

- **Identity Loss**
  - It helps preserve the consistency of input and output color composition by enforcing identity mapping when real samples of the **target domain** are given as the input to the generator.

$$L_{identity}^{s\rightarrow t} = E_{x\sim X^t}[\,|x - G_{s\rightarrow t}(x)|_1]$$

$$L_{identity} = L_{identity}^{s\rightarrow t} + L_{identity}^{t\rightarrow s}$$

Loss Function (3)

LsGAN Loss

- Use **least-squared-loss** instead of cross entropy loss to give a more accurate estimation about the reconstruction quality
- Advantage:
  - More stable training process
  - Faster convergence speed
  - Improved quality of generated samples
  - Avoid gradient saturation problems

$$L_{lsgan}^{s\to t} = E_{x\sim Xt}\left[\left(D_t(x)\right)^2\right] + E_{x\sim Xs}\left[\left(1 - D_t(G_{s\to t}(x))\right)^2\right]$$

$$L_{lsgan} = L_{lsgan}^{s\to t} + L_{lsgan}^{t\to s}$$

Loss Function (4)

- **CAM Loss (Class Activation Mapping Loss):**
  - Conditioned on the consistency of the object between image level and feature level
  - Minimize the difference in CAM (Class Activation Map)

$$L_{cam}^{s\rightarrow t} = E_{x\sim X s}\Big[\log\big(\eta_s(x)\big)\Big] + E_{x\sim X t}\Big[\log\big(1-\eta_s(x)\big)\Big]$$

$$L_{cam}^{D_t} = E_{x\sim X t}\Big[\big(\eta_{D t}(x)\big)^2\Big] + E_{x\sim X s}\Big[\Big(\eta_{D t}\big(G_{s\rightarrow t}(x)^2\big)\Big)\Big]$$

$$L_{cam} = L_{cam}^{s\rightarrow t} + L_{cam}^{D_t} + L_{cam}^{t\rightarrow s} + L_{cam}^{D_s}$$

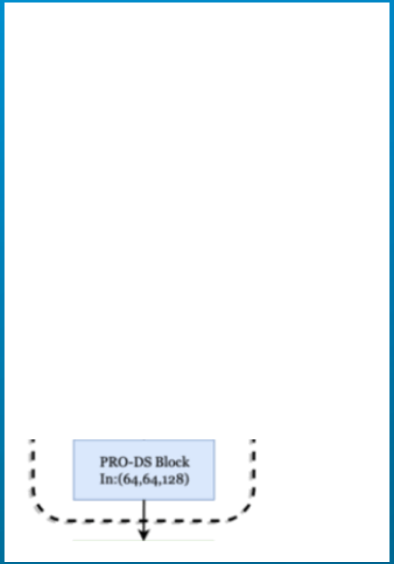
Total Loss

- **The final Loss Function is the result of the weighted combination of the above four Losses**

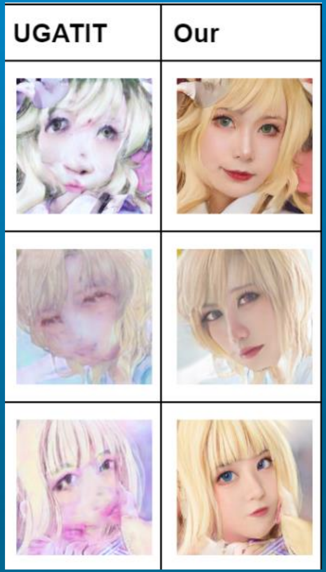
$$L = \lambda_1 L_{Cycle} + \lambda_2 L_{identity} + \lambda_3 L_{lsgan} + \lambda_4 L_{cam}.$$

where $\lambda_1 = 1, \lambda_2 = 10, \lambda_3 = 10, \lambda_4 = 1000.$

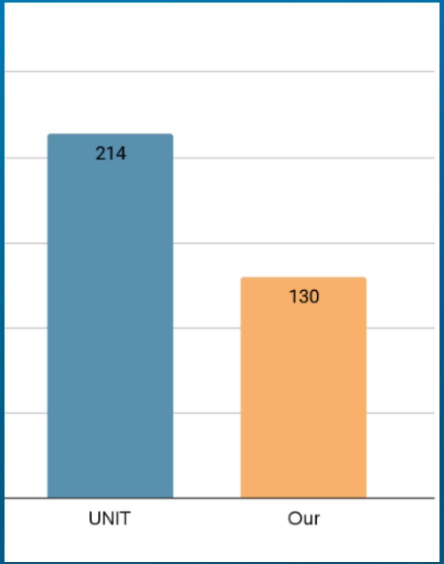
Three Advantages



High Scalability

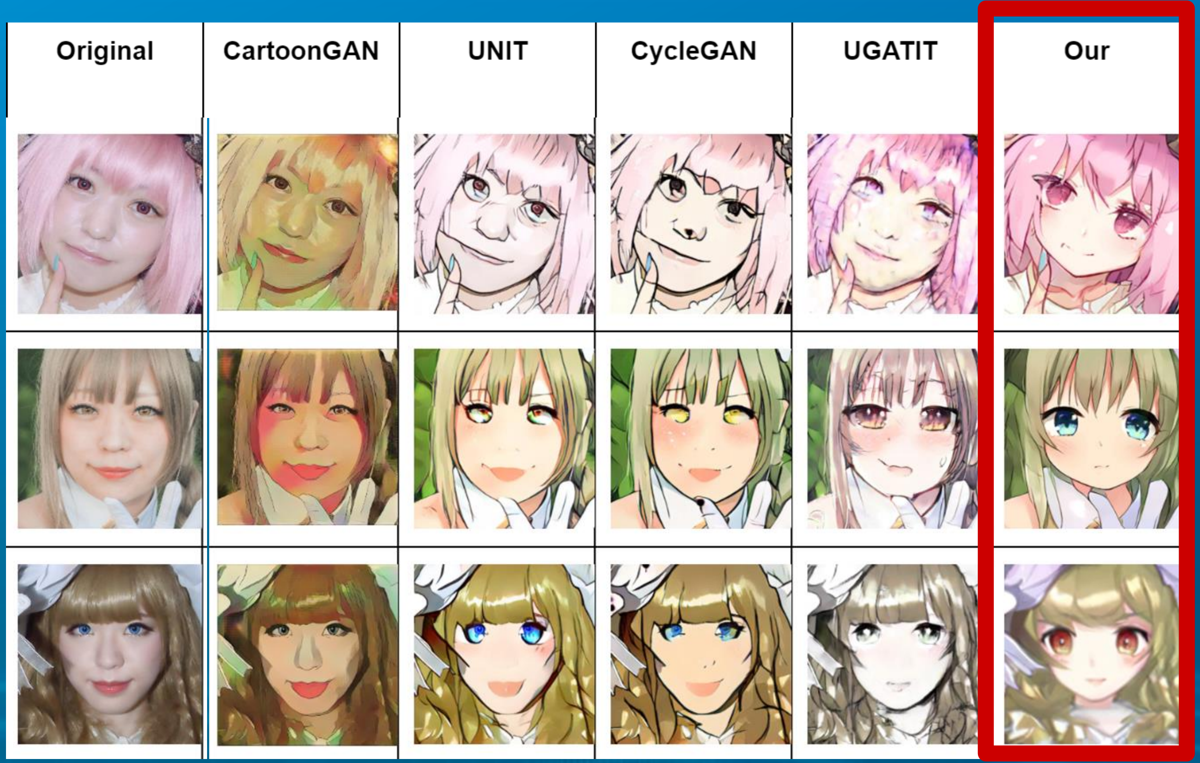


High Quality

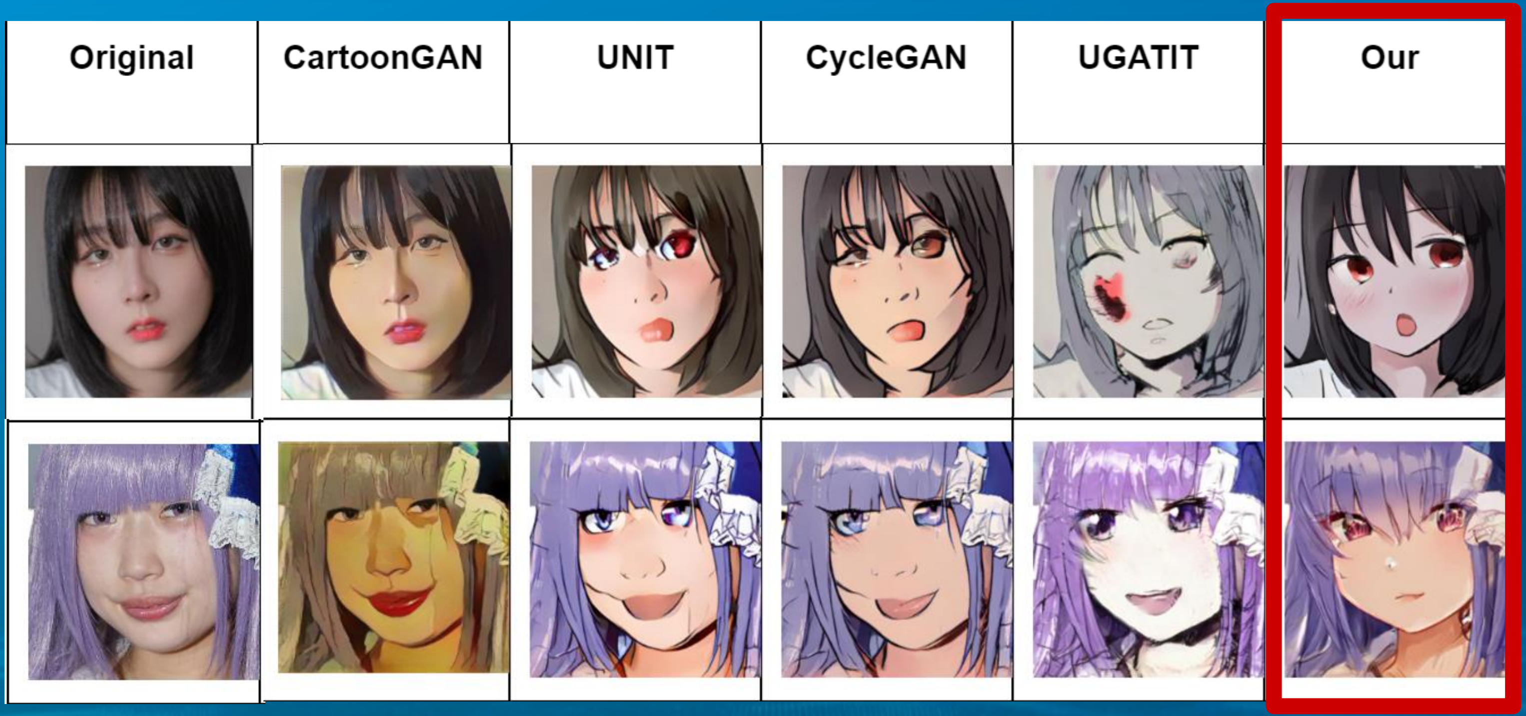


Low Computation

Example (Human -> Anime)

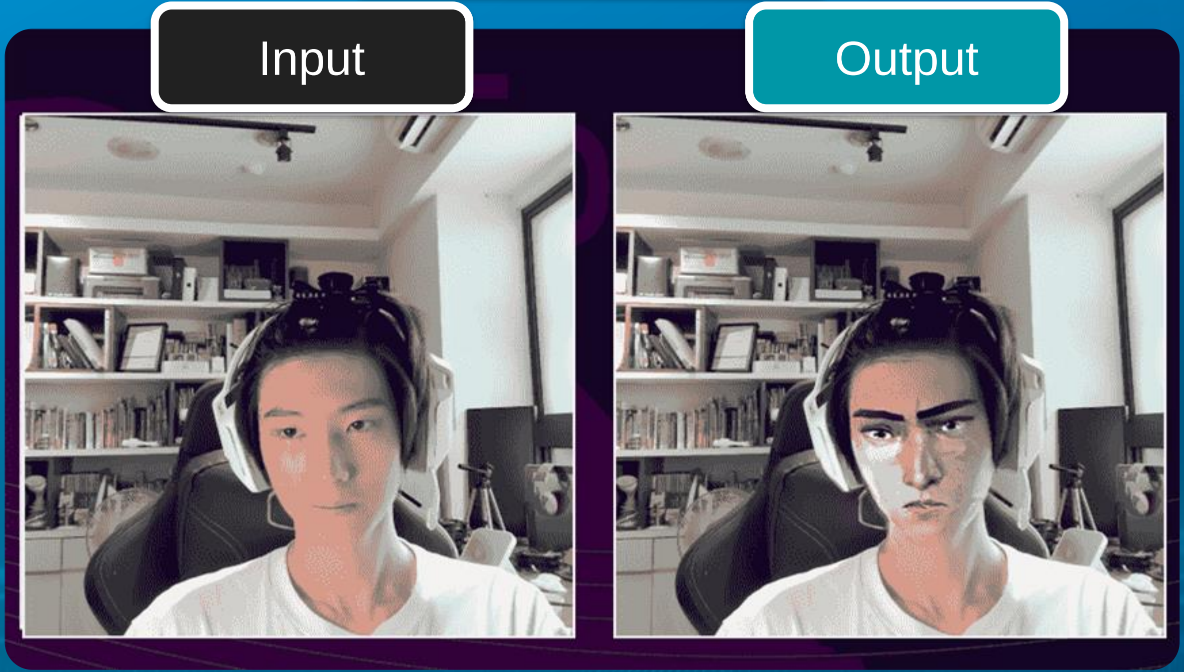


Example (Human -> Anime)

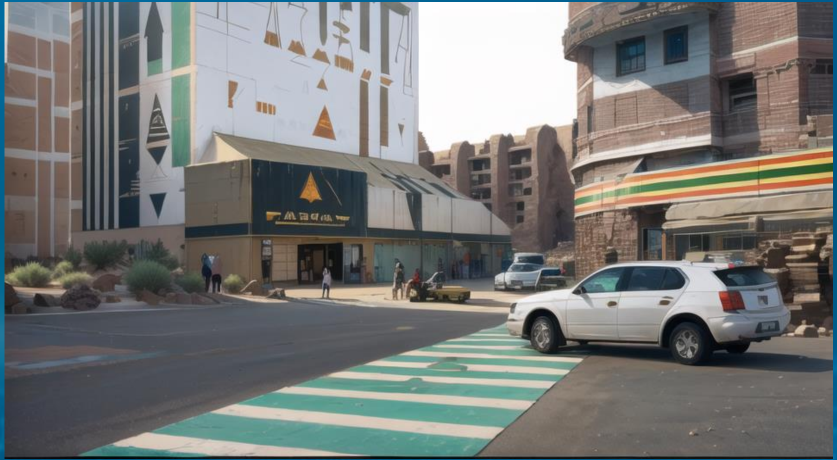


Low Computation

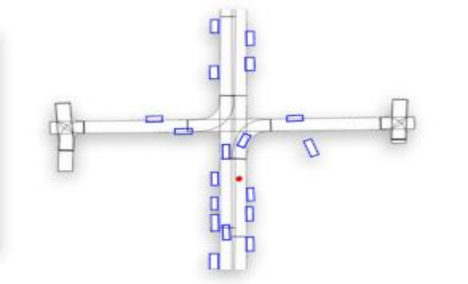
Working in Real-Time



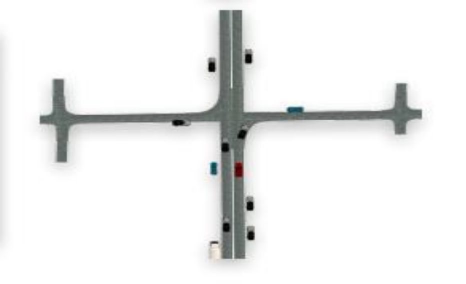
Example (Street -> Desert)



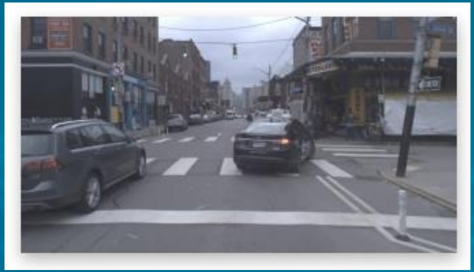
Application Scenario Example (Training Data Generation for AD)



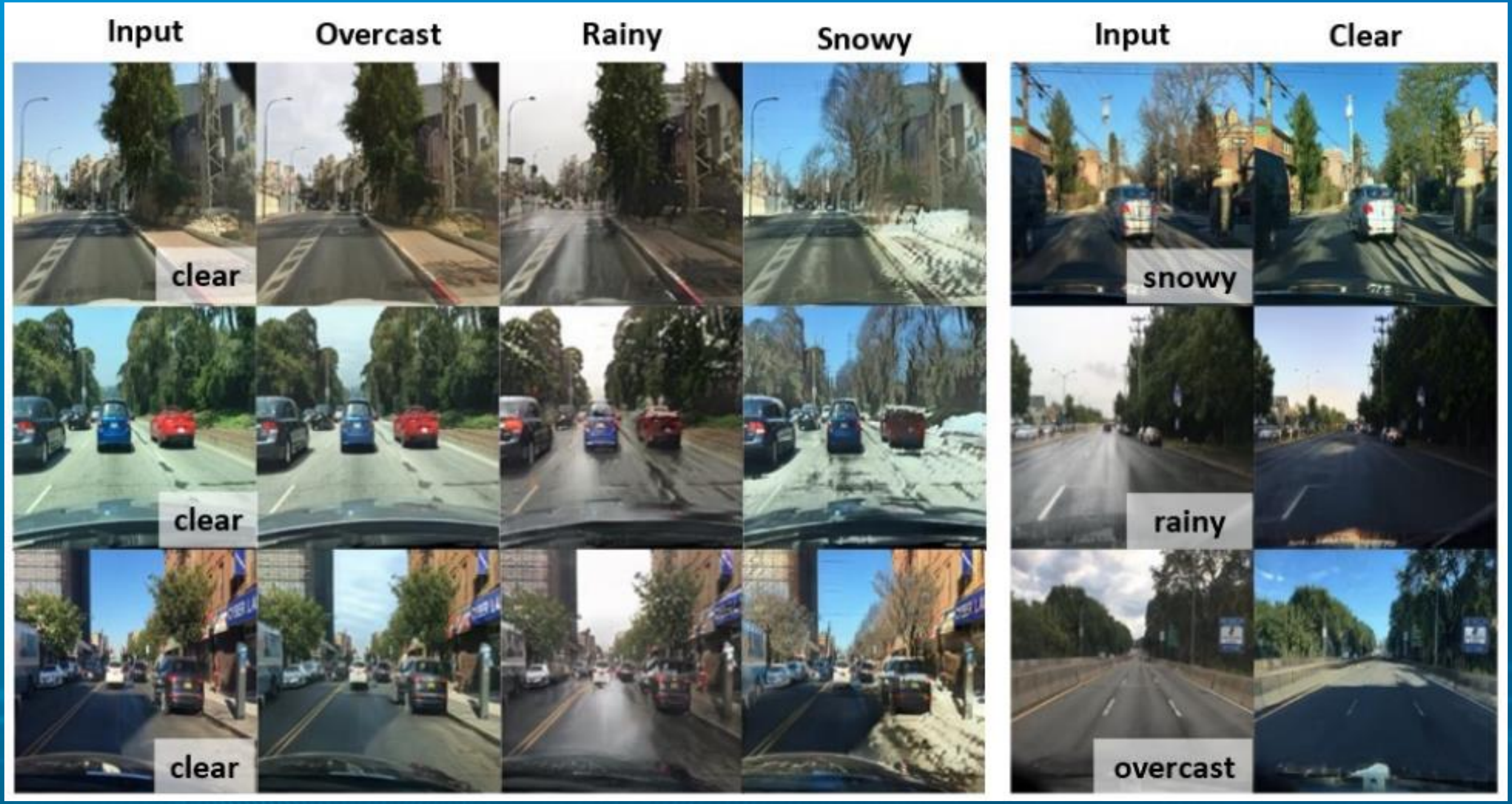
I2I



I2I



Diverse Scene Generation



Conclusion (1/3)

- **GenAI + VR/AR/MR**
  - The rise of GenAI, combined with XR, will revolutionize how we teach and interact with students
  - Virtual field trip brings brand new experience so that students can learn in an immersive environment
  - It opens up new ways of teaching, learning and stimulates creativity.

- **Mainstream DL Models for Image generation**
  - VAE (Variational Autoencoder)
  - NF (Normalizing Flow)
  - GAN (Generative Adversarial Networks)
  - Diffusion Model

Conclusion (2/3)

- **Image to Image Translation (I2I model)**
  - The I2I model can be used to convert style of the image while preserving the outline and semantics of every objects.
  - It can be used to create sci-fi like experience in the metaverse and teleport users to various worlds of imagination.
  - Another application of I2I is **Synthetic Dataset Generation**, which can be used to quickly generate image databases for Deep Learning model training

Conclusion (3/3)

- **PRO-U-GAT-IT**
  - We propose PRO-U-GAT-IT , a novel I2I model for efficient image style transfer
  - Through its advanced modular design, it is able to produce high-quality, superior images while reducing computation resources.
  - Application includes: avatar in metaverse, Sim2Real for AD ...etc.

Thank you

AI Research Center
Hon Hai Research
Institute (HHRI)

YouTube
立即訂閱

鴻海研究院

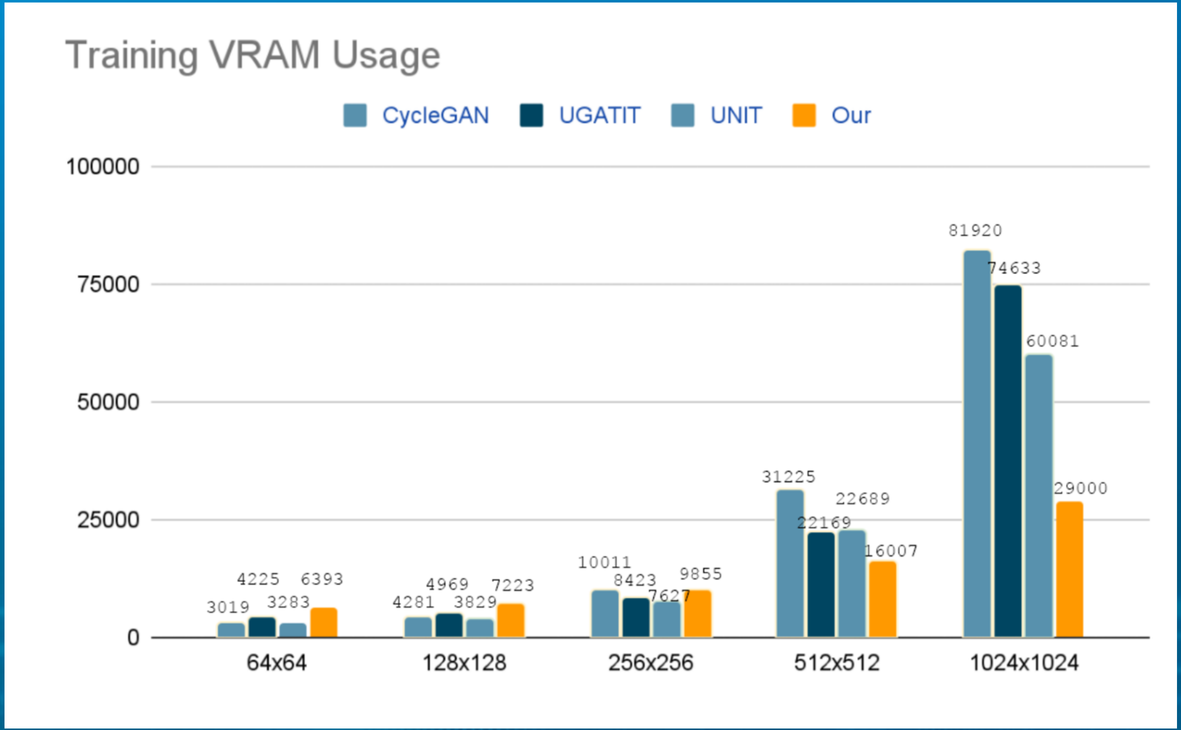
YouTube

PRO-U-GAT-IT

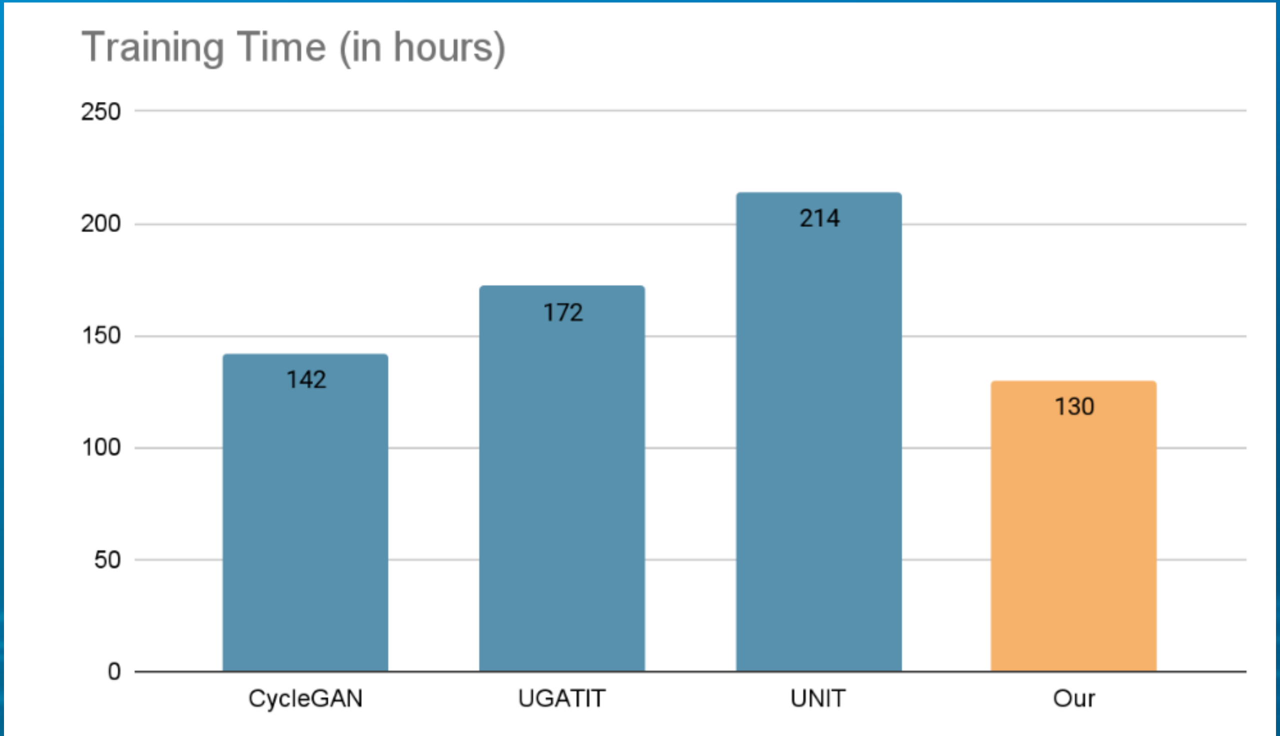
Progress in AR/VR

visionOS allows apps to fill the space around the user, move anywhere, and scale to the ideal size. Apps can even respond to light, creating shadows

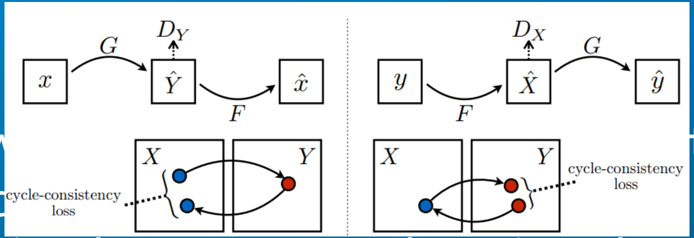
VRAM usage is low



Faster Training



Loss Function (1)



Cycle Loss

- In an unsupervised image translation task, w … ta, just separate sets of images from different c … infrared images). In order to learn the mapping between two domains, the model needs some kind of supervised signal to constrain the output
- Calculation :
  - First, decode the image from source domain A to target domain B
  - The image of target domain B that was just translated is then decoded back to source domain A
  - Calculate the difference between this reconstructed image and the original image as the cycle Loss:

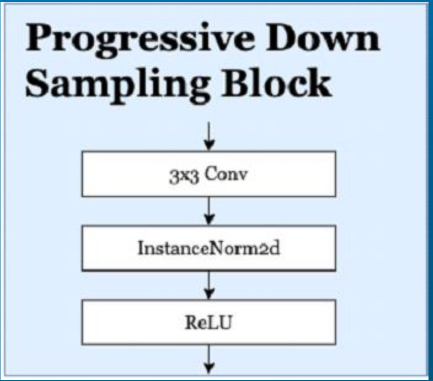
$$L_{Cycle}^{s\to t} = E_{x\sim X_s}[\,|x - G_{t\to s}(G_{s\to t}(x))\,|_1]$$

$$L_{Cycle} = L_{Cycle}^{s\to t} + L_{Cycle}^{t\to s}$$

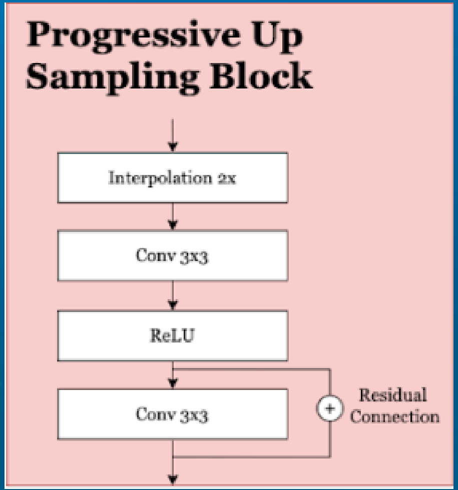
Introduction

FOXCONN 鴻海科技集團

- **Incremental Training Paradigm**
  - The purpose of incremental training with **PRO-DS Block & PRO-US Block** is to extract image feature adaptively according to the resolution of the image

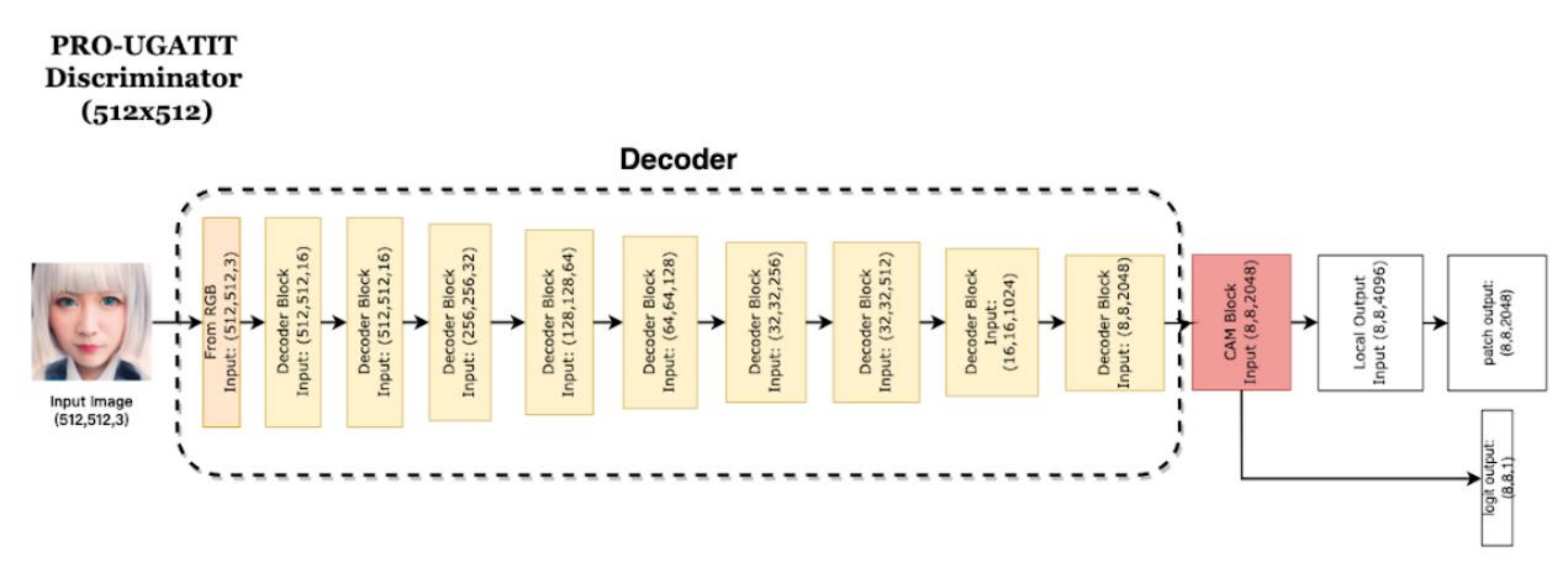


PRO-DS Block

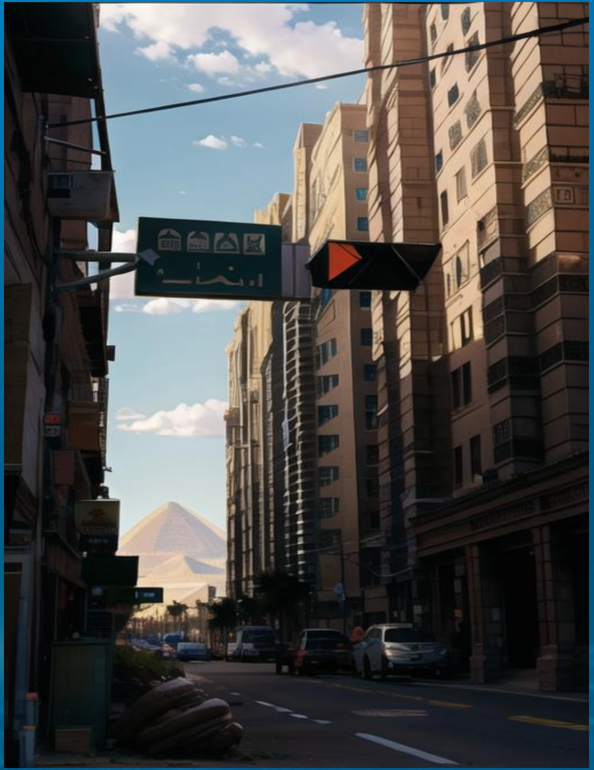


PRO-US Block

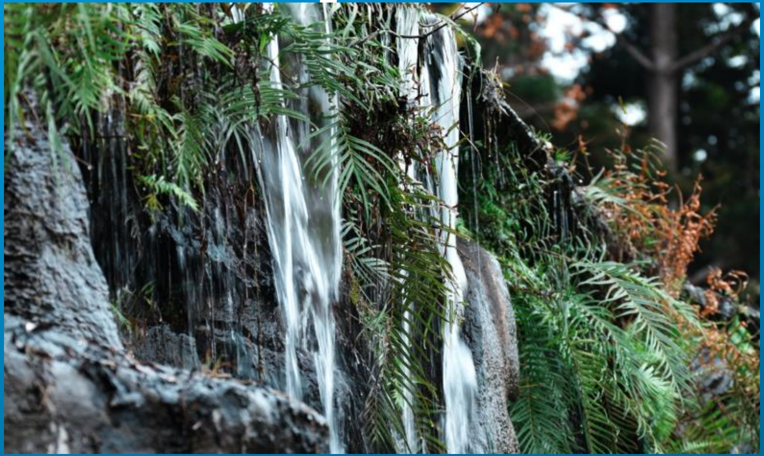
Discriminator Architecture

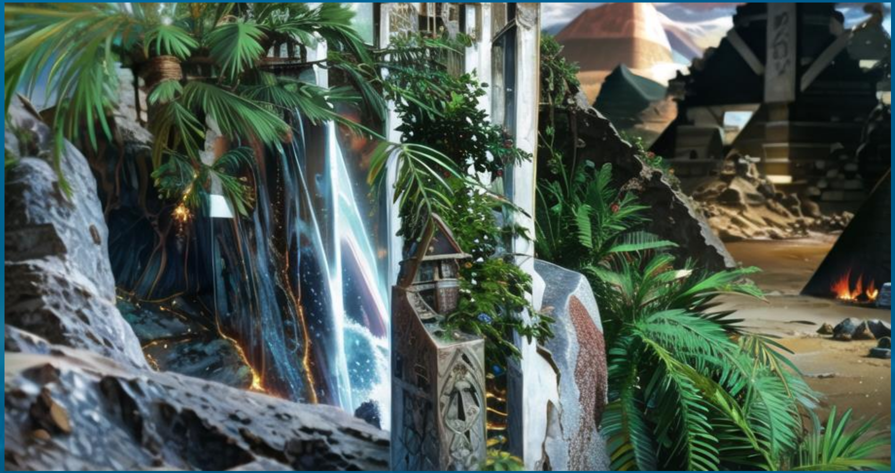


Example (Street -> Desert)



Example (Scene -> Desert)





Example (Scene -> Desert)