Mastering in pixel art image : dramatic image translation converting pixel art image to realistic image

Team Name : 디비디비딥이실

1. Team details

a. 이성현 / 2020114010 / sheepswool@yonsei.ac.kr / team leader, writing report, paper research

b. 남세현 / 2022149028 / daniel5253@yonsei.ac.kr / modeling, analysis, paper research, presentation

c. 박수연 / 2019161011 / suyonpark00@naver.com / modeling, writing report, paper research

d. 이우흥 / 2020147586 / wuheng@yonsei.ac.kr / data preparation, modeling, editing report

2. Problem Statement

Pixel art, a digital art form characterized by its low-resolution and stylistic simplicity, holds a significant place in various cultural industries such as gaming, animation, and digital art. It is favored for its (1) ease of creation, (2) unique aesthetic appeal, and (3) ability to encapsulate complex features in a minimalistic format. This accessibility allows both amateur and professional artists to effectively convey their creative intentions. However, there exists a potential to further realize these artistic visions by transforming pixel art into more detailed and realistic imagery. In other words, this art style offers high accessibility for drawing the creator's idea, so bridging the pixel art image to a realistic image will offer a high chance for realizing the creator's intention effectively.

The primary objective of this study is to explore the translation of pixel art into realistic images using unsupervised image-to-image or text-to-image translation methods. This initiative involves more than just enhancing the image resolution. It encompasses a comprehensive reconceptualization of pixel art into a three-dimensional style, while maintaining the essence of the original art.

This approach seeks to introduce a new level of detail and realism, focusing on either image-based or text-based translation methodologies. The study aims to determine which of these methods is more effective in accurately capturing and rendering the intricacies of pixel art in a realistic form, thereby providing valuable insights into the potential of these technologies in the realm of artistic image translation.

Translating pixel art to realistic images can assist artists in producing high-quality visuals while preserving the core features and intentions of the original art. This process not only benefits professional artists in refining their work but also enables amateurs to expand their creative scope, potentially enriching the overall artistic landscape.

In the realm of computer vision, this project is primarily focused on the task of unsupervised two-domain image-to-image translation, employing cycle-consistency constraints. This complex task will be approached by harnessing the capabilities of various advanced generative models, which are foundational in their generative architectures. Notably, models like cycleGAN, UNIT, and Stable Diffusion have

each shown proficiency in specific facets of image-to-image translation. However, there remains a significant research gap in understanding their comparative effectiveness, especially concerning the criteria of image quality, fidelity, and the preservation of essential features during domain translations.

The core aim of this study is to meticulously compare the performance of these models through a series of controlled experiments. The focus will be predominantly on the images generated by cycleGAN, UNIT, and Stable Diffusion. The comparison will not be limited to assessing the visual quality of the generated images but will extend to evaluating how well each model preserves the structural and contextual integrity of the original images after translation. An additional objective is to investigate how each model navigates the intricacies of different image domains, particularly variations in texture, color, and patterns, and to understand the impact of these factors on the accuracy of the translation.

A crucial aspect of this project is to delve into the underlying mechanisms of how these models encode and decode images into and from their latent spaces. This investigation is expected to provide insights into the strengths and limitations of each model's approach to image translation.

This study proposes a hypothesis that passing through two steps for generating the realistic images performed better than only one step, which only passes through the input image-embedded latent space. This counterintuitive hypothesis includes the approach of using an Image-to-Text-to-Image model and Image-to-Image-to-Image model. This new approach for embedding the images will be rigorously tested to ascertain its potential in enhancing image translation processes.

By addressing these challenges, the project aims to contribute significantly to the advancement of unsupervised image-to-image translation. The insights derived from this research are expected to be pivotal in guiding future developments in the field, particularly in applications where accurate and high-fidelity image translation is essential. Ultimately, this study seeks to pave the way for generating higher quality images, thus making a substantial contribution to the fields of computer vision and digital art.

Importantly, this study posits a notably counterintuitive idea: the process of embedding the input image through textual descriptions, as opposed to the conventional approach of utilizing latent space, may significantly enhance the likelihood of successfully decoding or generating a realistic output image. This hypothesis challenges prevailing norms and opens up new avenues for exploration in the field of image translation.

3. Related Work

Task

This section delves into the various existing methodologies and studies pertinent to the enhancement of pixel art image resolution. Unlike conventional super-resolution tasks, this project's objective diverges significantly, necessitating a profound comprehension of pixel art imagery and an acute interpretation of the artist's underlying intent.

A notable precedent in this domain is the study "Generating Images from Pixel Art Images" Gonzalez et al.(2020)[1]. This research embarked on an ambitious endeavor to synthesize pixel art images through a deep learning model trained on a corpus of pixel art. Its distinction lies in the recognition that each pixel in an art piece is not merely a color unit but a carrier of significant intent and meaning, as envisioned by the artist. However, the study

presented a potential avenue for transcending beyond mere dimensional replication of pixel art, suggesting the feasibility of generating images with enhanced dimensional complexity.

In the broader context of computer vision, this project situates itself within the unsupervised two-domain image-to-image translation task framework. The methodologies central to this task are primarily founded on the principles of Generative Adversarial Networks (GANs) and Unsupervised Image-to-Image Translation Networks (UNIT). These approaches represent the forefront of current research in the field, offering innovative pathways for image generation that are both sophisticated and adaptive to the nuanced characteristics of pixel art imagery.

Model

The current project contemplates the application of several advanced models to address its specific tasks, namely cycleGAN, UNIT, and Stable Diffusion, each contributing uniquely to the realm of image-to-image translation.

The inception of Generative Adversarial Networks (GANs) by Goodfellow et al.[2] marked a significant advancement in image synthesis. However, the application of GANs to image-to-image tasks revealed inherent limitations, such as non-convergence, mode collapse, and diminished gradients, despite their state-of-the-art performance. The challenge is amplified by the scarcity of labeled data, prompting extensive research into unsupervised methodologies. Unsupervised image-to-image translation, therefore, stands as a critical frontier in computer vision.

Zhu et al.'s introduction of cycleGAN[3] significantly advanced the application of GANs in this domain. Unlike traditional GANs, cycleGAN facilitates image-to-image translation without the need for paired data, employing dual sets of generators and discriminators for each domain. A distinctive feature of cycleGAN is its incorporation of a cycle consistency loss, ensuring that an image translated from one domain can be reverted to its original domain with minimal distortion. This mechanism is vital for maintaining the integrity of key attributes in the translated images.

Furthering the field, Liu et al.'s research on UNIT[4] amalgamates the architectural principles of Variational Autoencoders (VAE) and GANs within a shared-latent space framework. This model learns the joint distribution of images across different domains, enabling translation by encoding an image into this shared latent space and subsequently decoding it into the target domain. This approach yields more fluid and coherent translations, bridging the gap between disparate image domains.

The recent development of Rombach et al's Stable Diffusion models[5] represents a leap in image synthesis and translation efficiency. Central to this model is a diffusion process that incrementally introduces noise into an image and learns to reverse this process, effectively capturing and replicating complex patterns and textures of the source images. Stable Diffusion's core principle offers vital insights for unsupervised translation tasks, particularly highlighting the stability and scalability of generative models in handling two-domain translations.

4. Specific Objectives

a. Task description

This task is, as mentioned, unsupervised image-to-image translation from two different domains. Realistic art images are generated from pixel art images. To achieve the goal, the task demands an exploration of various methodologies

within the constraints of available resources. The benchmarks for success in this task are threefold: (1) the generation of realistic images that are high in fidelity and exhibit a profound visual understanding, (2) the accurate reflection of the creator's intention embedded within the pixel art, and (3) the consistent generation of diverse images.

b. Dataset

For the purpose of this study, our team has meticulously gathered a dual dataset comprising pixel art images and realistic images. The first dataset, sourced directly from the internet, encompasses a total of 900 images in each category - pixel art and realistic images. Further, through various methodologies which will be elaborated upon subsequently, an additional set of 100 three-dimensional realistic images was procured to enhance the understanding and representation of realism in the generated images.

c. Evaluation metric

The Frechet Inception Distance (FID) stands as the predominant metric in evaluating Generative Adversarial Networks (GANs), measuring the feature distance between real and generated images. The rationale behind selecting the FID score is its alignment with human subjective evaluation in assessing the quality of generated images. Given the project's objective to generate images that not only resemble pixel art but also embody realistic features, the FID score emerged as the apt metric for evaluation. In our experiments, the performance of cycleGAN and UNIT was assessed by measuring both the semantic and quantitative distances between randomly selected cartoon-realistic images and their corresponding images generated from pixel art. The FID score is computed using the following formula:

$$FID = ||\mu_X - \mu_Y||^2 - Tr(\sum_X + \sum_Y - 2\sqrt{\sum_X \sum_Y})$$

Here, X and Y represent the embeddings of the pixel and realistic images, presumed to follow normal distributions. μ_X and μ_Y are the magnitudes of the vectors. Tr denotes the trace of a matrix. $\sum_X$ and $\sum_Y$ are the covariance matrix of the vectors. This formula provides a robust framework for quantifying the divergence between the generated image and its realistic counterpart, thus serving as a critical tool in the evaluation of our image translation task.

d. Approach

To address the task at hand, our team implemented a structured approach encompassing 5 distinct methodologies. The initial phase of the project saw the concurrent implementation of the first 2 approaches. Subsequent to the analysis and evaluation of these initial models, the latter 3 methodologies were employed, each designed to build upon the insights gained from the preceding steps.

1. cycleGAN

The first approach utilized in this project was the Cycle Generative Adversarial Network (cycleGAN). This choice was driven by the need for a model capable of effectively learning from a limited dataset. cycleGAN's unique architecture, which includes two translator networks for two different domains, is guided by the cycle-consistency loss. This loss function plays a pivotal role in preventing mode collapse, a common pitfall in GAN training. It ensures that the model does not merely learn exact matching pairs, but rather understands the distribution of

each domain. This approach is particularly effective in training the model stably with a smaller quantity of images.

2. UNIT

UNIT, or Unsupervised Image-to-Image Translation Network, represents the second approach and shares conceptual similarities with cycleGAN. However, it introduces a nuanced difference in understanding image distributions. UNIT operates under the assumption that 2 distinct image domains can share a common latent space. In this project, UNIT encodes both pixel art and realistic images into this shared latent space. Subsequently, it generates realistic images from the learned distribution, offering a different perspective on image translation. The evaluation of the UNIT model, akin to cycleGAN, is conducted using the Frechet Inception Distance (FID) metric, ensuring a consistent and objective assessment of its performance in translating pixel art into realistic images.

- **Interim Check**

During the initial phase of the project, both cycleGAN and UNIT were deployed to assess their efficacy in the task of image translation. An interim evaluation revealed that both models exhibited similar performance trends, with UNIT achieving a marginally better initial and final Frechet Inception Distance (FID) score in fewer epochs.

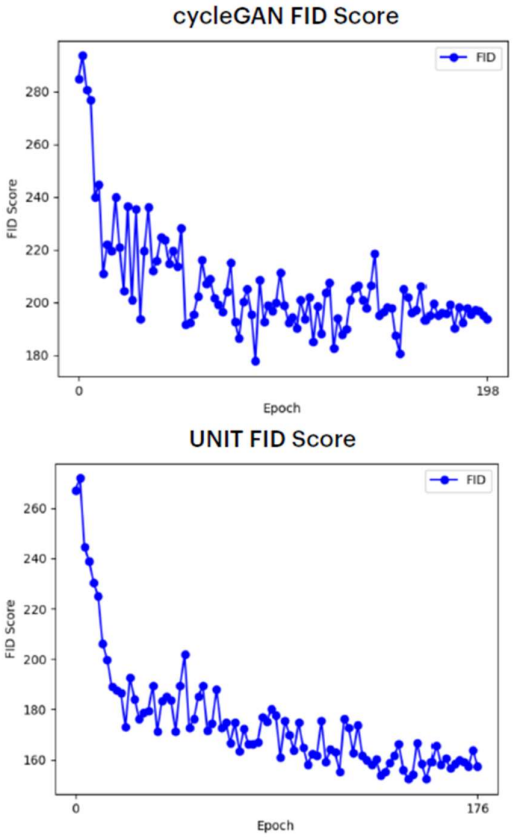
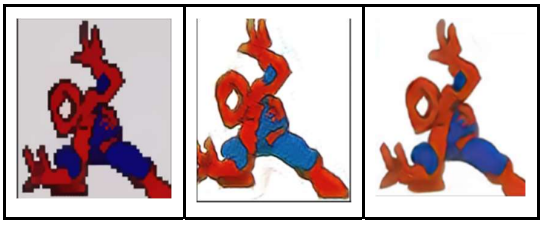


Fig 1: validation score plot of cycleGAN and UNIT. UNIT outperforms cycleGAN.

The analysis of test sample images generated by both models indicated discernible differences in output quality. UNIT demonstrated a superior ability to produce smoother and more realistic images compared to cycleGAN. This observation leads to the inference that UNIT possesses a more refined capability in understanding and interpreting the nuances of the input images, transcending the performance of cycleGAN in this regard.



original image	cycleGAN	UNIT

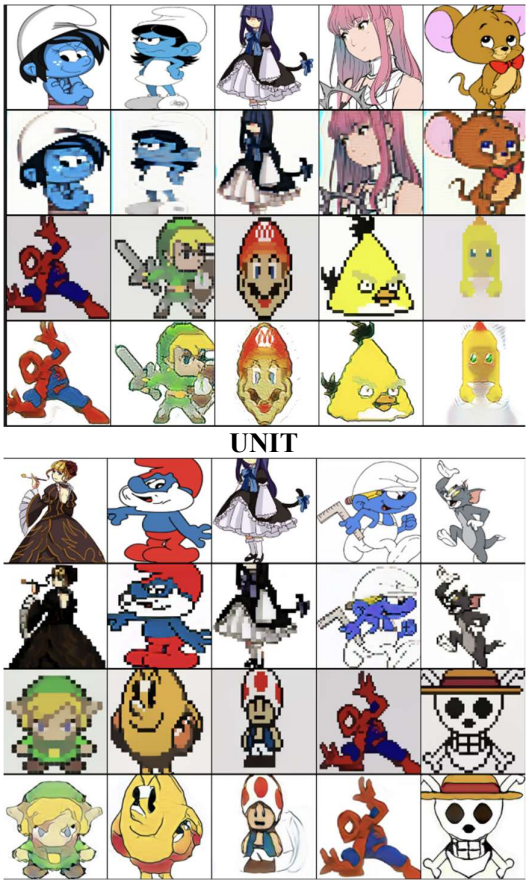
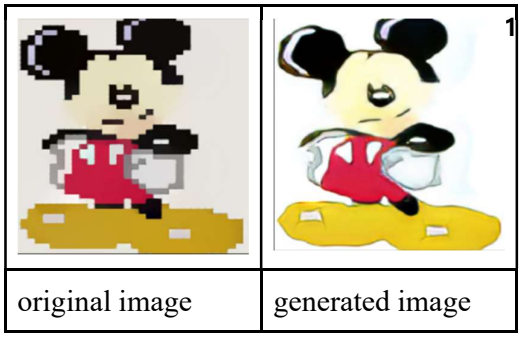


Fig 2: Translation cases of cycleGAN and UNIT. Samples were drawn from best-performing checkpoint of each model. UNIT's translation smooth edge better.

Despite UNIT's comparatively better FID score, it is important to note that neither model achieved the desired proficiency level for the task at hand. The results predominantly aligned with those typical of super-resolution tasks. A critical observation during this evaluation was the loss of essential information in the translated images. For instance, in some instances, vital details such as the eyes were inadequately represented or entirely lost in the translation process.



original image	generated image

Fig 3: Translation result of mickey mouse. Character's eyes are erased in the translation process.

The outcomes of this interim check underscore the necessity of identifying a model with an enhanced visual understanding, particularly attuned to the intricacies of pixel art images. A model with a more profound comprehension of the input image is pivotal in avoiding the generation of images with awkward or incomplete representations. Such an advancement is crucial for producing images that are not only aesthetically pleasing but also carry practical and coherent information, thereby fulfilling the objectives of this project more effectively. This realization serves as a strong motivation to explore and develop models that can bridge this gap in image translation fidelity.

3. U-GAT-IT

In response to the insights gleaned from the evaluations of cycleGAN and UNIT, our team progressed to the implementation of Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization for Image-to-Image Translation (U-GAT-IT). This model was adopted to meet the specific needs identified from the results of the previous models.

U-GAT-IT innovatively integrates an attention mechanism into the UNIT

framework, augmenting both the discriminator and generator components of Generative Adversarial Networks (GANs) with an attention head. This architecture enables the real domain generator to focus on attributes that confer realism to an image, while the pixel domain generator concentrates on pixel-specific attributes. Similarly, this structured focus is applied to the discriminators, allowing the model to hone in on critical features that distinguish the two domains.

The images generated by U-GAT-IT exhibited a marked improvement in recognizing and rendering small, yet pivotal attributes of the source images. Notably, U-GAT-IT achieved enhanced detail in areas such as the eyes in pixel art images, a significant advancement over its predecessors. This improvement indicates a superior capability of the model to interpret and translate the nuances of the input images more effectively.

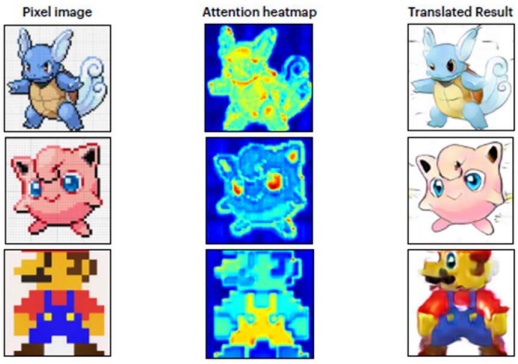


Fig 4: Attention map extracted from pixel-to-real generator (translator) and Translation result of U-GAT-IT. Attention block captured pixel-like regions well. Furthermore, detailed features like eyes were reinforced in translation.

It is a remarkable improvement for letting the model understand the images better than before. However, two limits with the current approach are still analyzed through the result. First, the images are not generated in the context this project is aiming for. The model for this project needs to generate a 3D-like output.

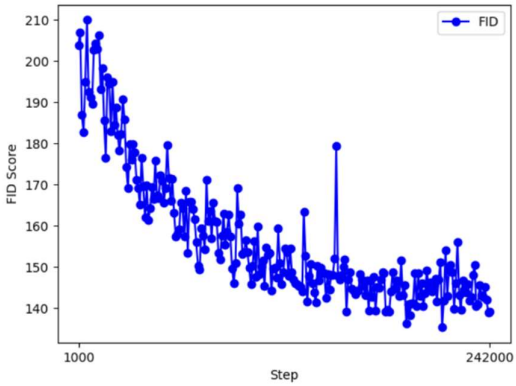


Fig 5: validation score plot of U-GAT-IT.

It is observable that the FID score is better than the performances of cycleGAN and UNIT. With the last sequence of training for each model (either epochs or steps), the U-GAT-It got the best score among the three models.

	cycle GAN	UNIT	U-GAT-IT
Trained Epoch	86	154	305
FID Score	177.6	152.2	135.2

Fig 6: test FID Score comparison matrix. U-GAT-IT best performed among three models.

Second, the quality of the dataset is not appropriate for generating the wanted outputs. There is a difficulty in generating a 3D image without training enough of such images. Additionally, there is a risk of generating an image without stepping on dramatically different domains of image. It is not ensured that the low-resolution pixel art image can match itself with the high-resolution realistic image well.

In conclusion, while U-GAT-IT represents a significant step forward in terms of attention to detail and image understanding, the task of generating realistic images from pixel art requires further exploration and possibly additional steps. Future work might involve enhancing the dataset quality, focusing on 3D image representation, and refining the model to bridge the gap between low-resolution pixel art and high-resolution realistic images more effectively. The pursuit of these objectives will be critical in achieving stable and realistic image generation that aligns with the project's vision.

4. Dataset Reconstruction

Given that the images generated from the initial datasets failed to capture realistic features, a strategic decision was made to reconstruct the dataset to better align with our goal of generating 3D realistic images. To this end, an additional 100 datasets were specifically curated, focusing on 'realistic images that exhibit 3D characteristics'. This collection included a diverse array of images that exemplified the desired 3D realism. Here are some examples of new datasets:

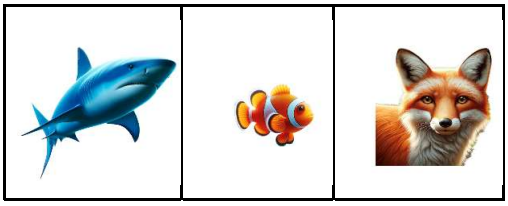


Fig 7: 3D real image dataset examples.

However, despite these efforts, the resultant images generated from these datasets did not meet our expectations of success. One of the primary challenges encountered was the limited size of the dataset, constrained by both time and resource availability. Given the vast and diverse nature of realistic 3D imagery, the small dataset size proved inadequate for accurately capturing and learning the essential features of these images. This limitation was further compounded by the inherent complexity of translating from low-information pixel art to intricate 3D images, a task that was rendered nearly impossible due to the significant disparity between the two domains.

Consequently, this situation necessitated the exploration of alternative methodologies capable of incorporating and leveraging prior knowledge about 3D realistic images. The recognition of this requirement marked a pivotal point in our project, underscoring the need for a more robust and sophisticated approach that could effectively bridge the substantial gap between simplistic pixel art and complex 3D realistic imagery. This shift in strategy aimed to overcome the limitations posed by dataset constraints and domain disparities, paving the way for more successful image generation outcomes.

5. Image-text-image

Pixel image to text and text to image is a way to create 3D realistic images with a model with sufficient ability to interpret images represented in text. Because it is not a direct translation from image to image, we tried and compared the image caption model BLIP(Bootstrapping Language-Image Pre-training) and the VisualQA model LLAVA(Large Language-and-Vision Assistant) to minimize information loss in the caption.

BLIP is a model that generates synthetic captions through a captioner and removes noise captions through a filter. BLIP, where captions and filters work together to bootstrap captions, was chosen because of its good understanding of language and

images at the same time. However, pre-trained BLIP, which failed to perform fine tuning due to lack of paired data sets of pixel images and captions, did not show the desired performance. Due to the severe loss of information, most of the captions about the specific features of the image were missing. In addition, there was a problem that we had to add the "realistic image" on every caption because we had to use this caption as an input for the generative model.

Therefore, we used a VQA captioning model that can induce us to generate the desired caption for the original pixel image. We used LLaVA (Large Language-and-Vision Assistant), which is easy to access and has a code public as a VQA capturing model. As input from LLaVA, we used the original pixel image and the question for the image, which requires the model to describe as much information as possible, including color or specific features. In addition, we also added a question like "generate prompt for generating realistic image of the character" to induce it to be used as an input for the generated model immediately.

The detailed caption generated by LLaVA were then inputted into the diffusion model DALL-E to create the desired 3D images. This approach, leveraging the richness of the captions, successfully produced images that were both realistic and reminiscent of the original pixel art.

"The image is a close-up of a smiling face with a frowning mouth, giving it a humorous appearance. The face is predominantly yellow, with the mouth being blue. crying, and the face is positioned in the center of the image. The overall scene is a playful and lighthearted representation of a smiling face with a frowning mouth."
 
original image generated image
"The image features a skeleton character holding a sword and a shield, standing in a defensive stance. The skeleton is positioned in the center of the scene, with the sword held in its left hand and the shield in its right hand. The character appears to be a warrior or a knight, ready to face any challenges that come its way. The scene is set against a dark background, which adds to the dramatic atmosphere."
 
original image generated image

Fig 8: Generation examples using T2I generation models.

Despite these advancements, some challenges persisted. In certain instances, the final generated images did not fully align with the original pixel art image information. This discrepancy arose from the generative model's reliance solely on the text captions, leading it to arbitrarily

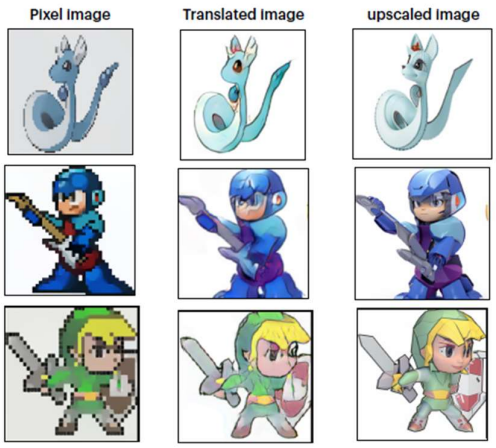
generate features not explicitly mentioned in the description. For example, if the direction in which a character was facing in the original pixel image was not specified in the caption, the generative model might produce an image with the character facing a different direction. This limitation highlights the need for even more precise and comprehensive captioning to guide the generative process effectively.

"The image features a video game character, Mario, riding on top of a green and white horse, which appears to be a Yoshi. The character is wearing a cape and is positioned in the center of the scene. The horse is prominently displayed, occupying a large portion of the image. The scene is set against a blue background, which complements the vibrant colors of the characters and the horse. The overall composition of the image showcases the iconic characters from the Mario franchise, creating a nostalgic and playful atmosphere."	
	
original image	generated image

Fig 9: Generation example using T2I model. Direction of character is not aligned.

6. Conditioned Upscaling

After exploring various methodologies, our final approach focused on utilizing a model with a rich generative prior. ControlNet, which leverages the generative capabilities of the diffusion model, was employed for various conditioned generation tasks. This model takes an intermediate image, one that has undergone resolution enhancement, and combines it with simple text prompts to output a 3D realistic image. Remarkably, the process does not require complex, image-specific prompts, underscoring the efficiency of the model in generating high-fidelity 3D images.



positive prompts : 3D
negative prompts : ((unrealistic)), low quality, blur, digital art, worst quality

Fig 10: Translation examples with Controlnet. Controlnet upscales translated image injecting '3D' attribute.

ControlNet offers a more nuanced and enriched guidance to the final generative model compared to direct input of pixel art images. This translation model acts as an intermediary, enhancing the understanding and interpretation of the pixel art before it is transformed into a 3D image. However, it is crucial to note that the effectiveness of this model architecture significantly depends on various factors: the extent and depth of the translation model's understanding of the pixel art, and how well it can interpret and translate these insights.

An illustrative example of this approach's limitations can be seen in the output image of an Iron Man character. When the translation model fails to adequately comprehend the nuances of the pixel art, the resulting image can diverge significantly from both the original pixel art and the collective human understanding of the character. This particular example underscores the challenges inherent in translating and generating images that accurately reflect the essence of the source material, especially when dealing with complex characters or concepts.

In summary, while the use of ControlNet with a generative prior marks a significant advancement in our project, it also highlights the critical need for precise and nuanced understanding of the source pixel art. The success of this model hinges on its ability to accurately interpret and translate the pixel art into a realistic 3D image that resonates with both the original artwork and the viewer's expectations.

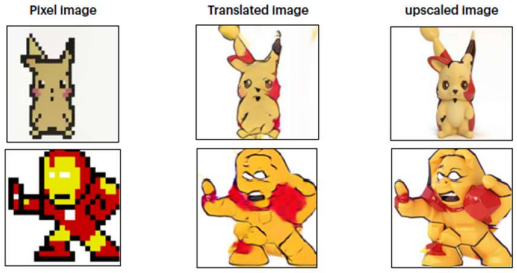


Fig 11: Unsuccessful translation examples with Controlnet. Controlnet relies on translation results of I2I translator, hence pixel art to low-quality image translation is required.

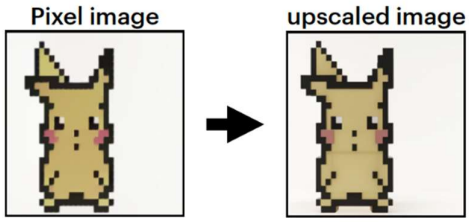


Fig 12: Short Ablation result of Translation model. We expected the '3D' prompt of Controlnet will utilize 3D appearance of character, but model captures pixelated feature of input as important and amplifies it. This reveals our translation model plays a significant role in eliminating 'pixel-like features' before upscaling to 3D image.

5. Final Results and Discussion

The exploration of five distinct approaches in this study has yielded a wealth of insights, each contributing to our understanding of image translation from pixel art to realistic 3D images.

Initially, cycleGAN and UNIT provided a foundational understanding of the potential for image translation. The marginally superior performance of UNIT over cycleGAN highlighted the possibility of refining and modifying UNIT to achieve enhanced visual understanding. This comparison laid the groundwork for further exploration in image translation techniques.

Secondly, the integration of an attention layer in the U-GAT-IT model marked a significant advancement. This addition proved instrumental in enhancing the model's visual understanding, enabling it to retain minor yet crucial details from the input images. U-GAT-IT not only demonstrated the efficacy of attention-based models in improving visual comprehension but also underscored the need for models capable of transitioning across dimensions.

Thirdly, the process of dataset reconstruction underscored the importance of having a sufficient and representative dataset. Adequate data is paramount in enabling the model to accurately interpret and understand the images, a crucial factor in the successful translation of pixel art to realistic imagery.

Fourthly, utilizing a well-trained generative model with a realistic three-dimensional image

as a prior, and embedding the image with text before regenerating it, yielded promising results. This Image-to-Text-to-Image approach generally produced images closely resembling the original input while simultaneously achieving realism. Although some information loss was observed, this method opened avenues for generating more dynamic and adaptable output images.

Finally, the approach of transforming the image into a smoothed line image, rather than embedding text, and then regenerating a three-dimensional realistic image also showed remarkable performance. This Image-to-Image-to-Image method demonstrated superior quality in image upscaling and exhibited higher consistency in the generation process. This consistency suggests a greater likelihood of producing output images that more accurately reflect the creator's intentions and understanding.

In conclusion, this comprehensive study presents a nuanced understanding of various approaches to image translation, each with its unique strengths and limitations. The insights garnered from these approaches lay a solid foundation for future exploration in the field of image translation, especially in the context of transforming pixel art into realistic 3D imagery.

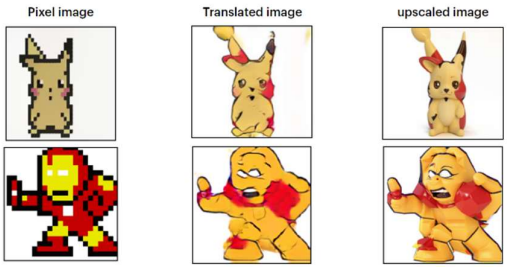
6. Limitation & Future Work

This study has made significant strides in generating realistic images from low-resolution pixel art. However, there remain several areas for improvement in enhancing image quality.

Firstly, both quantitatively and qualitatively, there is a critical need to enhance the dataset. The initial three approaches faced challenges in transitioning to an upscaled image output, largely due to the limited ability to understand realistic images and to learn the distribution of these images sufficiently. Therefore, future efforts should focus on acquiring a larger and higher-quality image dataset. Moreover, when employing two-step models (Img2Txt2Img or Img2Img2Img), an enriched dataset at the intermediate stage is essential for improved visual understanding and realization.

In addition, Subsequent research should explore methods to generate images that not only reflect the creator's intentions but also exhibit dynamic and varied features. For instance, a model receiving an image of Pikachu should be capable of generating a three-dimensional realistic image that can vary in posture or emotional expression.

Observations from the conditional upscaling model highlighted that the final image quality heavily depends on the intermediate translated image. An example with an Iron Man pixel art image revealed that inadequate translation led to a final image that deviated from the creator's intended portrayal. This outcome suggests a need for a deeper understanding of the pixel art, as the information derived solely from the pixel image may be insufficient for fully capturing the artist's vision.



Finally, there is a requirement for a more robust metric to evaluate the models effectively. The Frechet Inception Distance (FID) currently used relies on comparing the distributions of output and input images. However, the datasets collected thus far do not adequately represent the full spectrum of image distributions. Future research must focus on gathering a

comprehensive dataset that facilitates a more thorough understanding of image distribution, enabling a more accurate and meaningful evaluation of the models' performance.

In conclusion, while this study has advanced the field of image translation from pixel art to realistic imagery, the path forward involves refining datasets, enhancing model capabilities to reflect artistic intentions more accurately, and developing more sophisticated evaluation metrics. These efforts will be pivotal in realizing the full potential of image translation technologies.

7. Reference

[1] Adrian Gonzalez, Matthew Guzdial, Felix Ramos (2020). Generating Gameplay-Relevant Art Assets with Transfer Learning. In arXiv.

[2] Goodfellow, Ian, et al. "Generative adversarial nets." *Advances in neural information processing systems* 27 (2014).

[3] Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223-2232).

[4] Liu, M. Y., Breuel, T., & Kautz, J. (2017). Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, *30*.

[5] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10684-10695).

Contributions for the team project

이성현 – 주제 선정, 데이터셋 수집, 모델 선정, cycleGAN 논문 강독 및 팀 내부 발제, 중간 보고서 및 최종 보고서 공동 작성

박수연 - Blip fine tuning, Image to text to Image task 수행 (LLaVA->DALLE / Stable diffusion), 중간 보고서 & 최종 보고서 작성 기여

이우흥 - cycleGAN, 데이터셋 수집, 중간 보고서 작성 기여 & 최종 보고서 수정

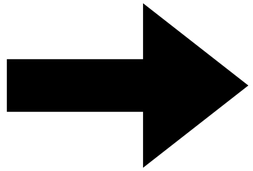
남세현 – 모델 선정, Scoring method, UNIT, U-GAT-IT 구현, 실험, 데이터셋 수집, Controlnet pipeline 도입, 발표자료 제작 및 발표, 주제 구체화 및 조정에 적극참여. 최종 보고서 수정.

MasterPixel

pixel image to realistic image

이우흥, 이성현, 박수연, 남세현

Problem Statement



Task

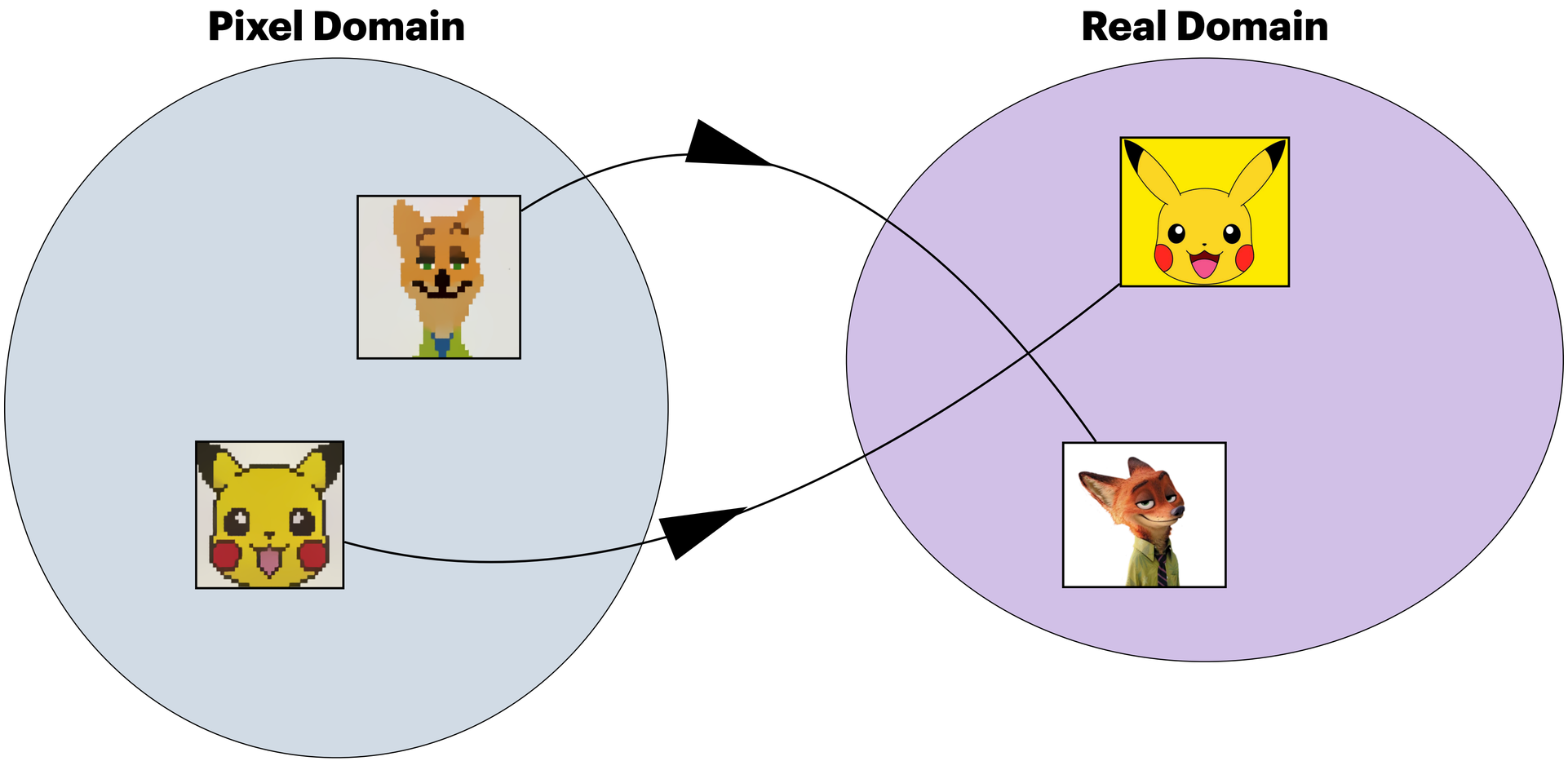
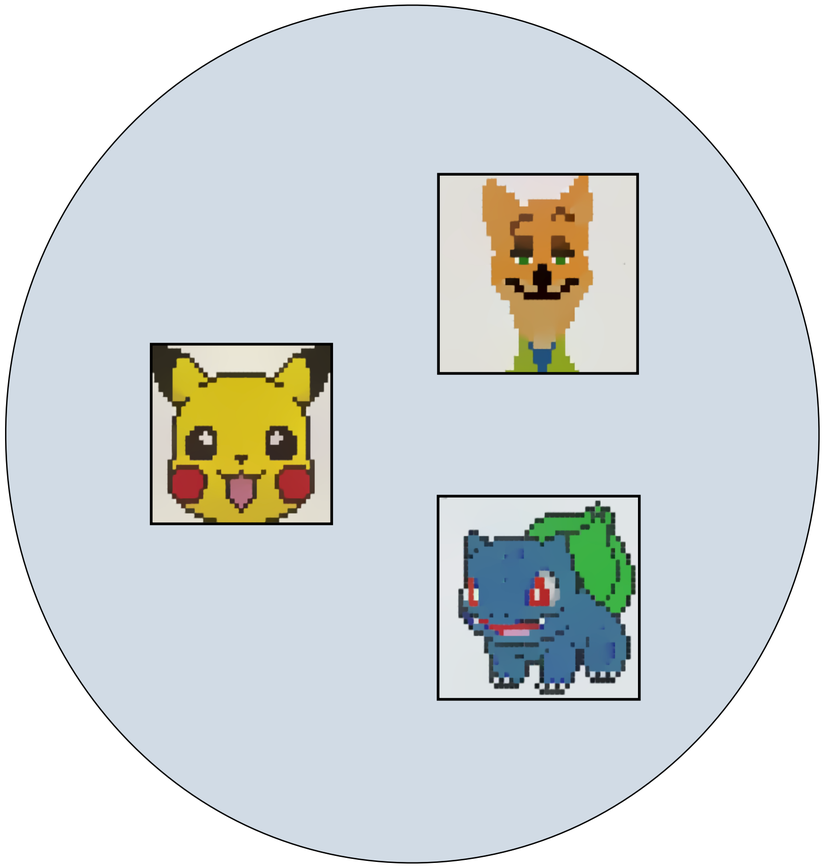


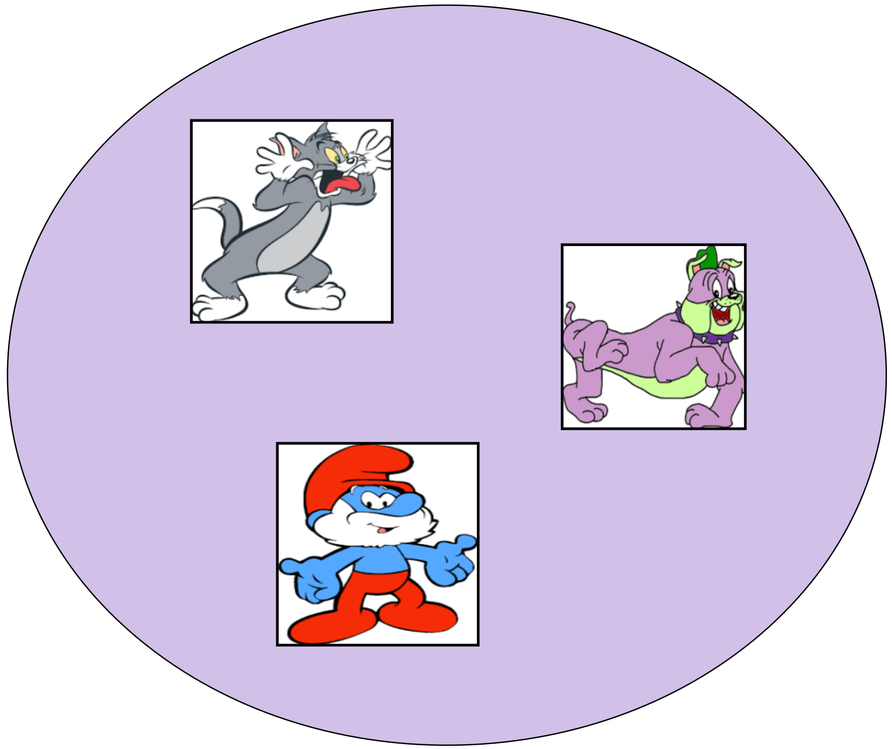
image - to - image translation

Dataset

Pixel Domain

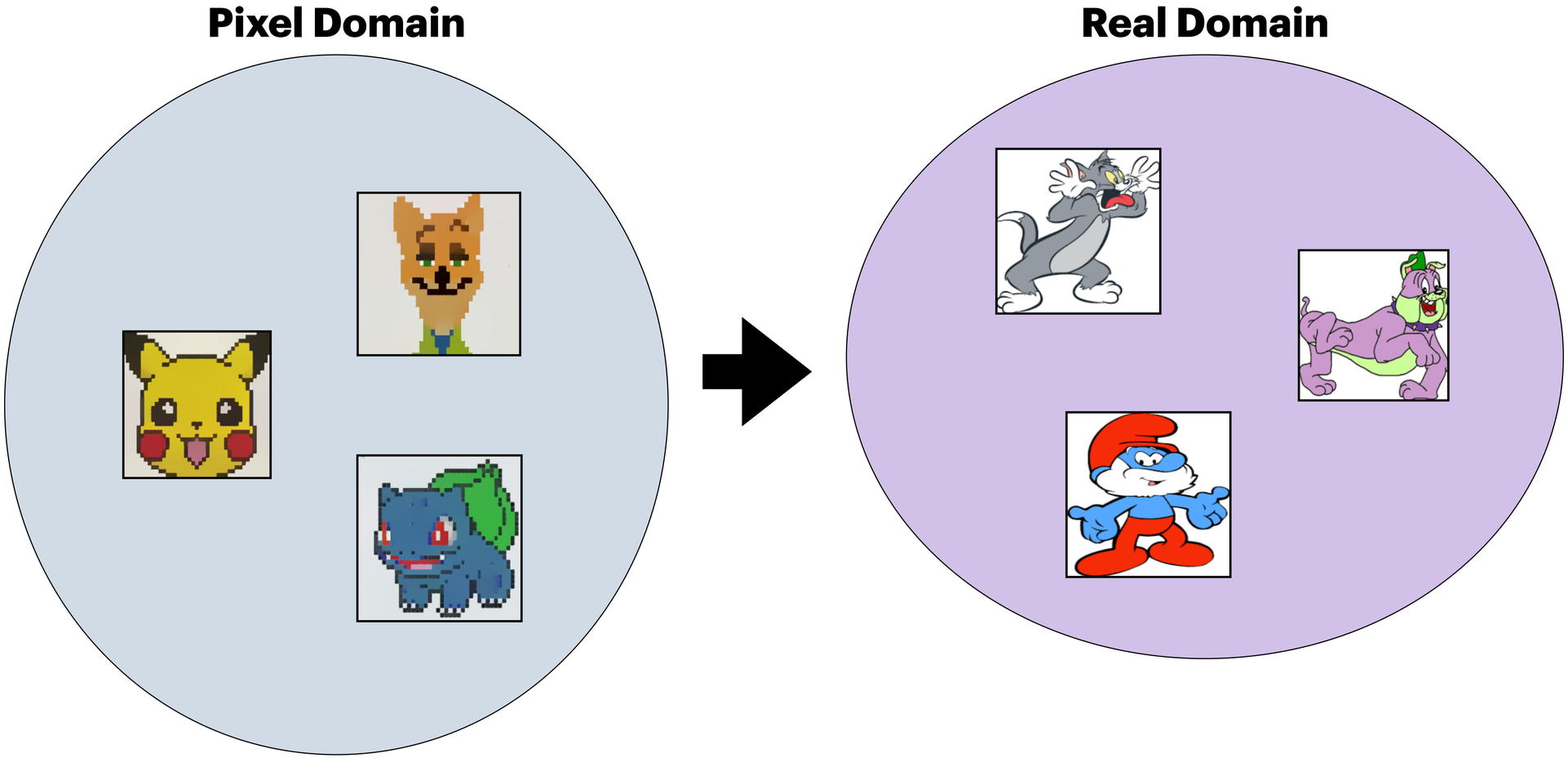


Real Domain



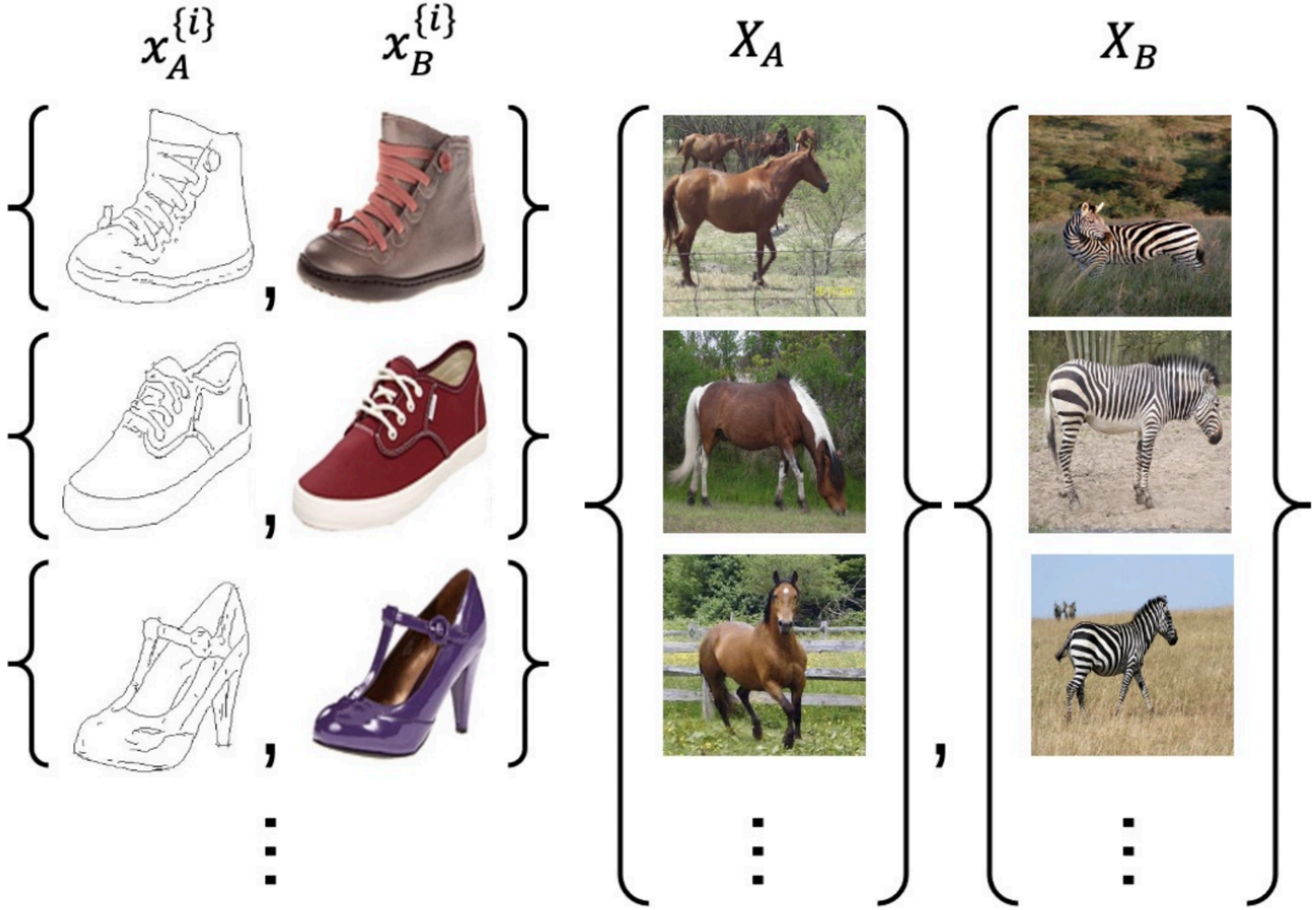
Crawled 900 images each

Task



Unsupervised image - to - image translation

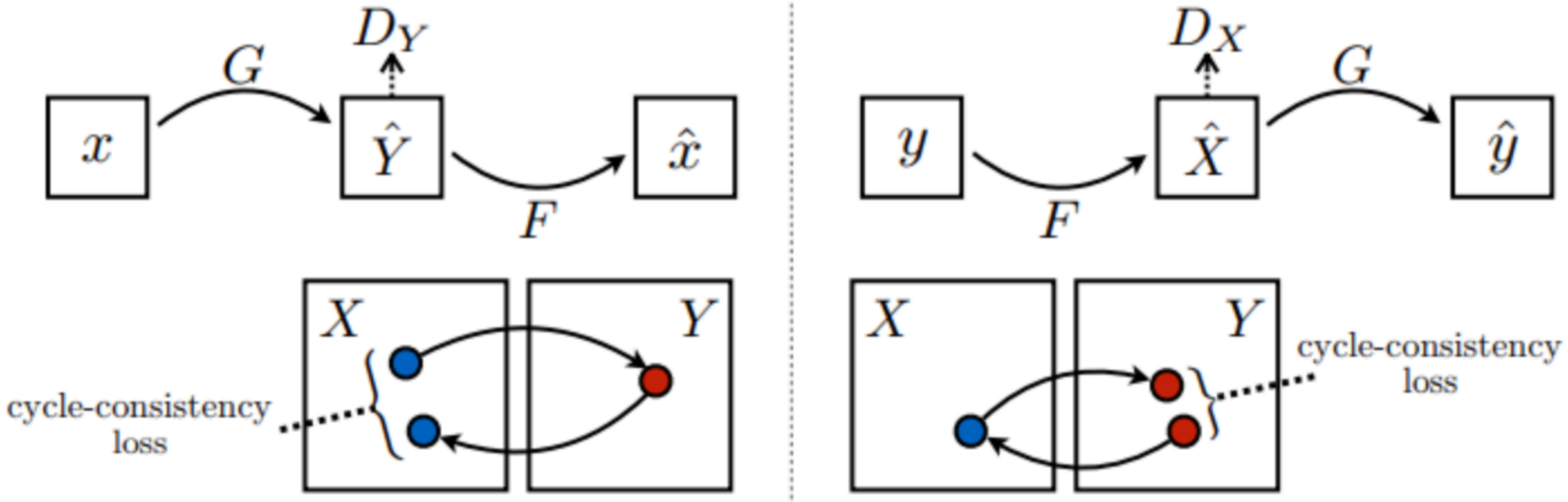
Unsupervised image - to image translation



translation on unpaired condition

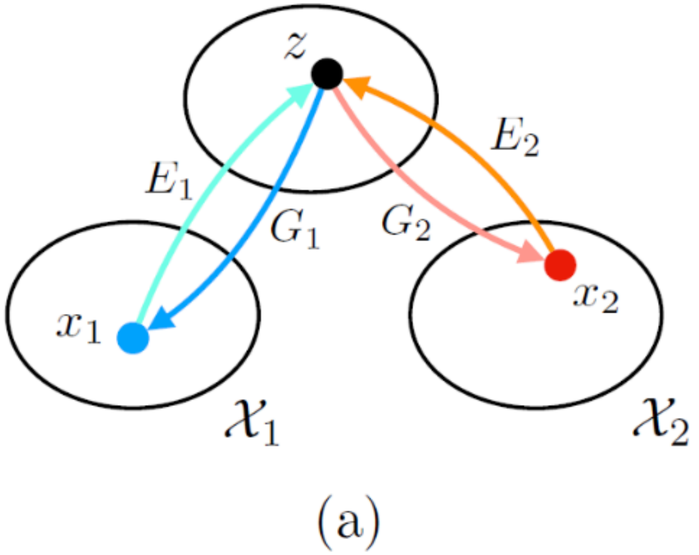
Approach 1 : cycleGAN

Adversarial loss

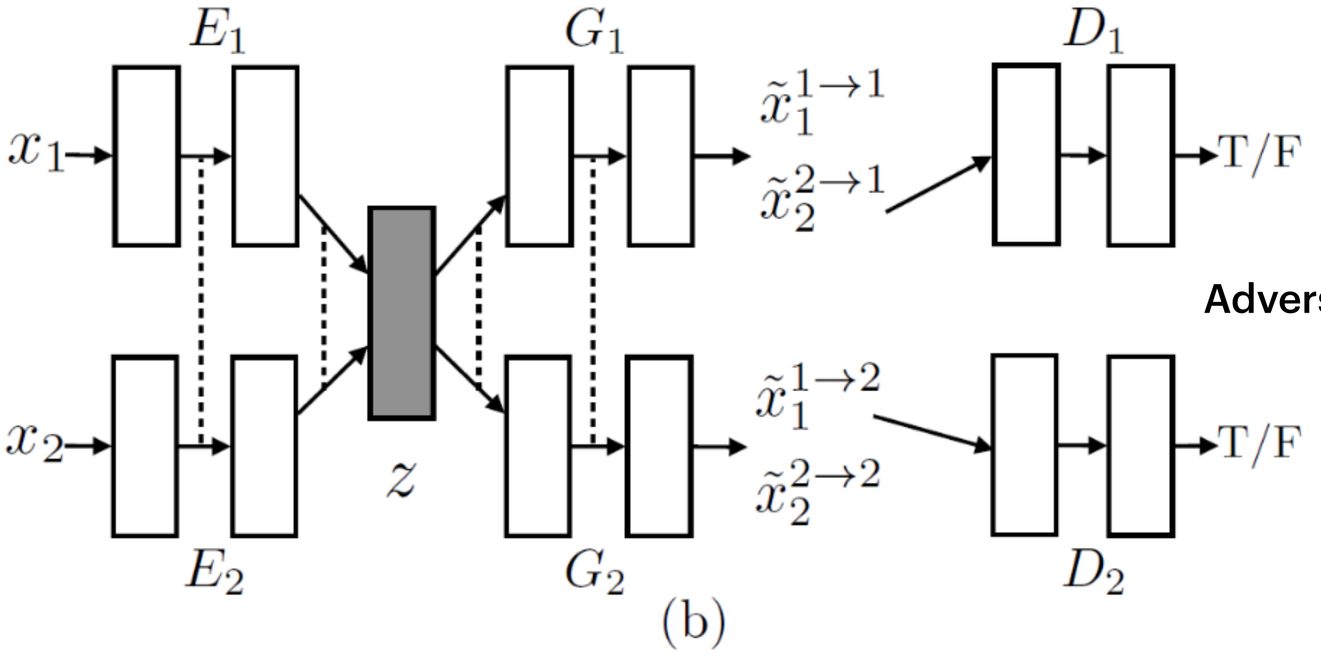


cycle-consistency loss

Approach 2 : UNIT

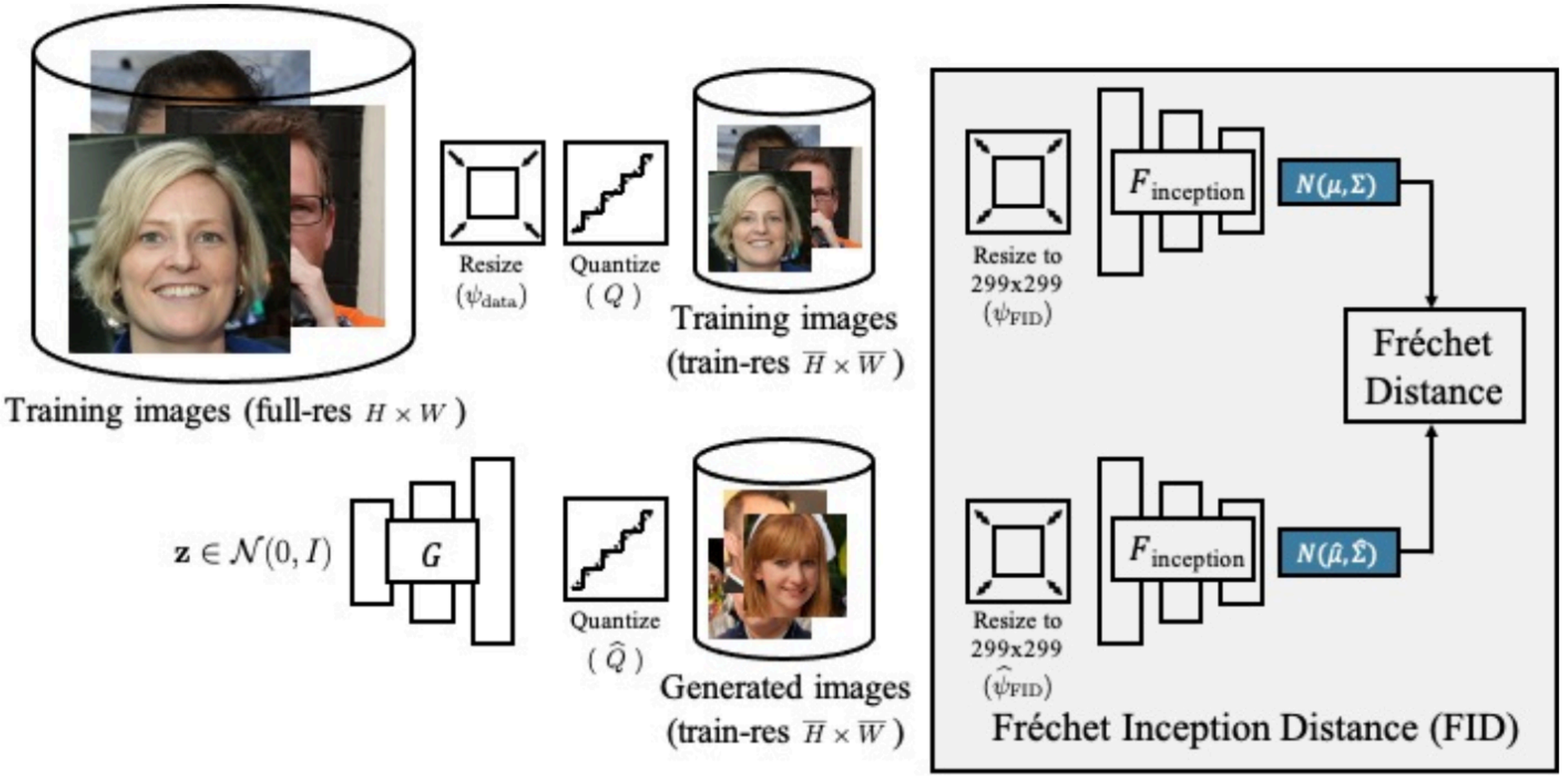


Variational loss



cycle-consistency loss

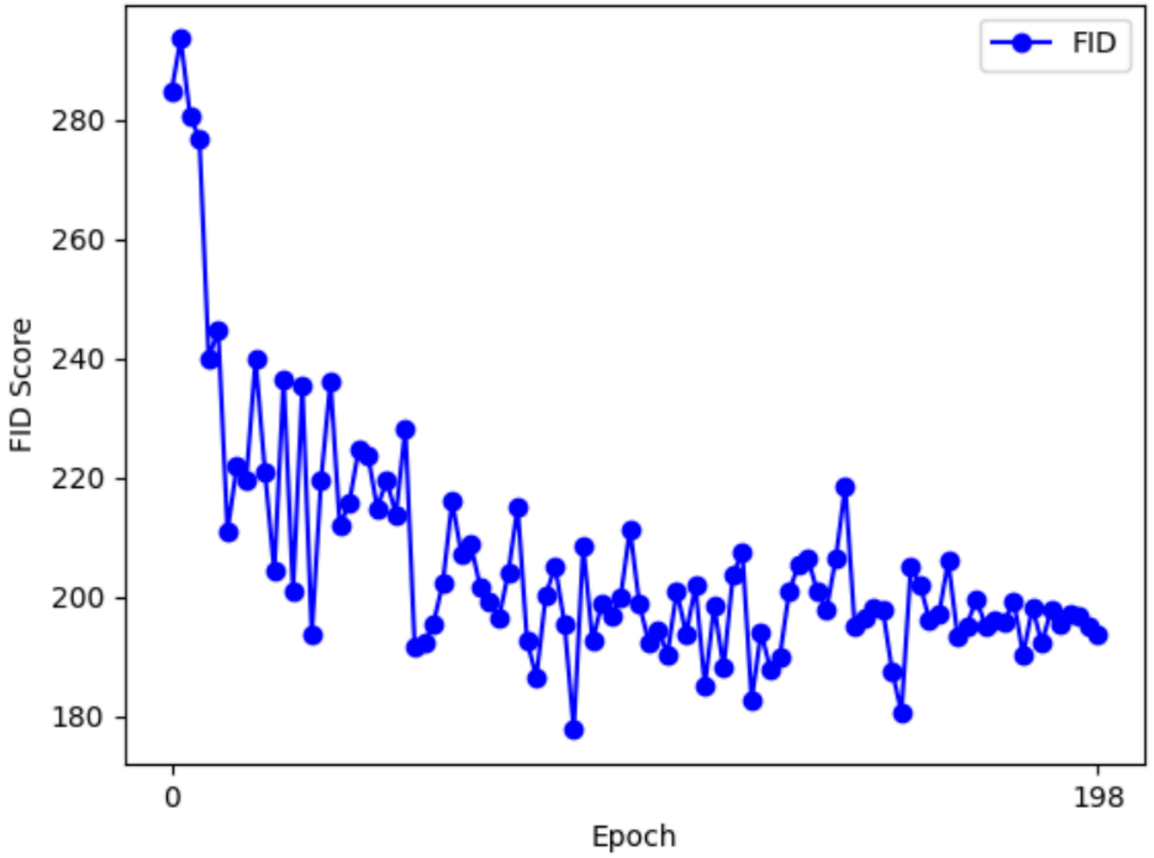
Evaluation : FID



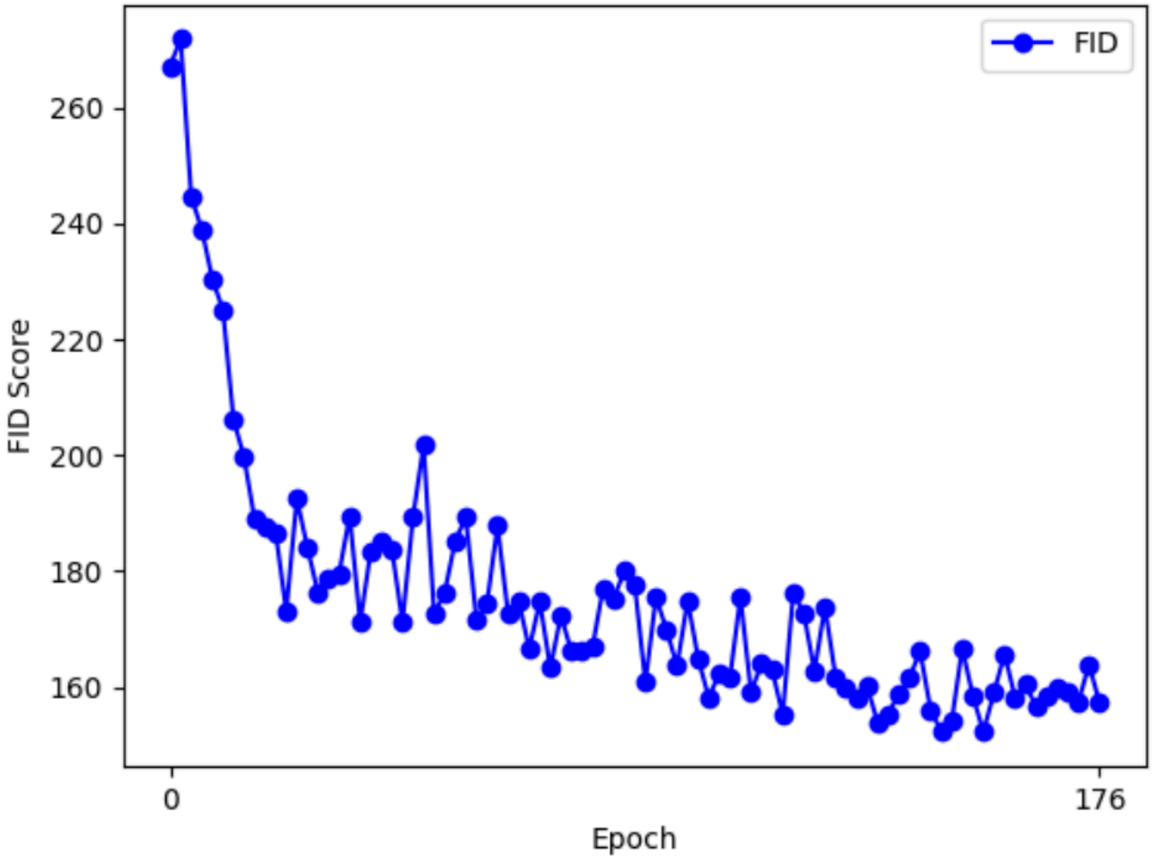
Frechet Inception Distance : semantic, qualitative distance between two image sets

Evaluation

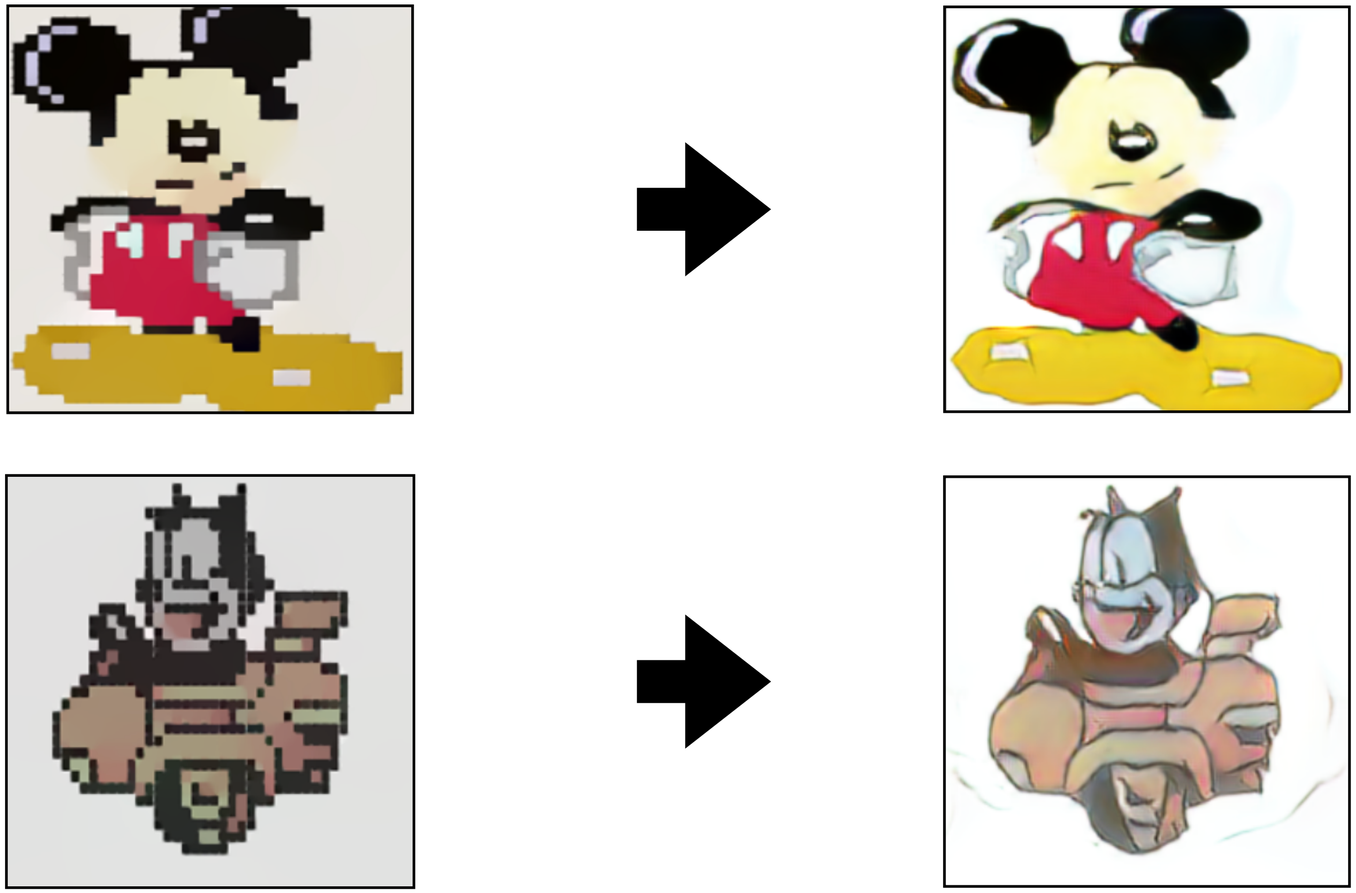
cycleGAN FID Score



UNIT FID Score

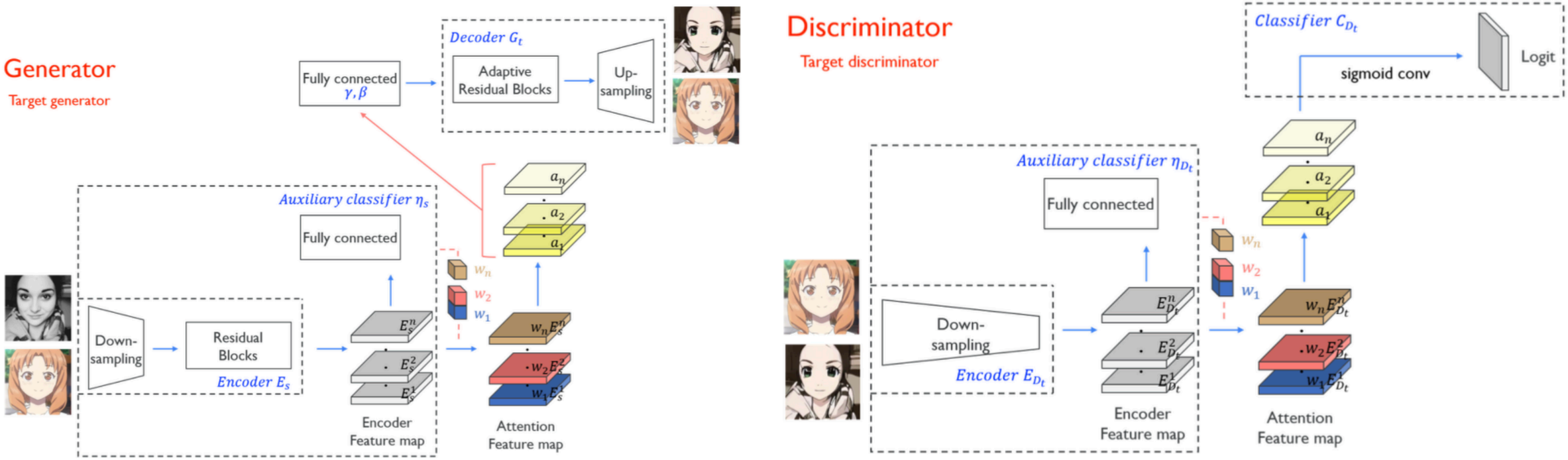


Limitation



looses detailed representation like eyes

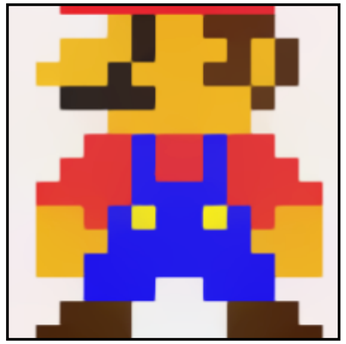
Approach 3 : attention based UNIT



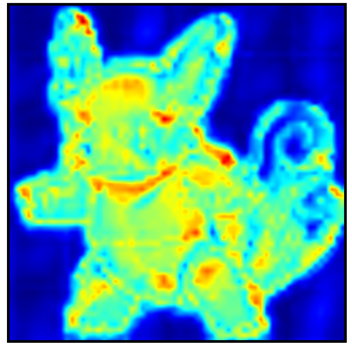
U-GAT-IT : attention mechanism captures domain-relevant areas

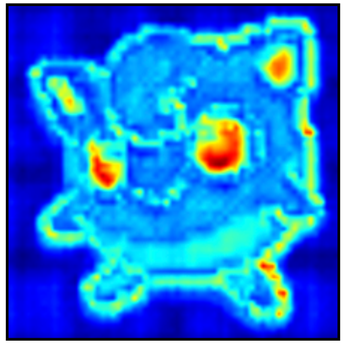
Result

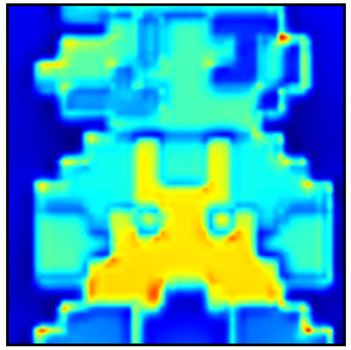
Pixel image



Attention heatmap



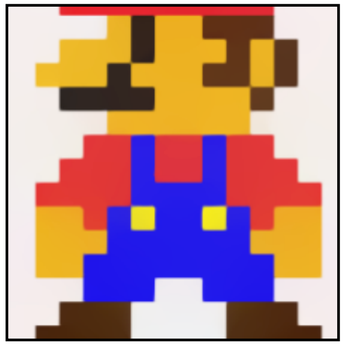




Translated Result

Limitation

Pixel image



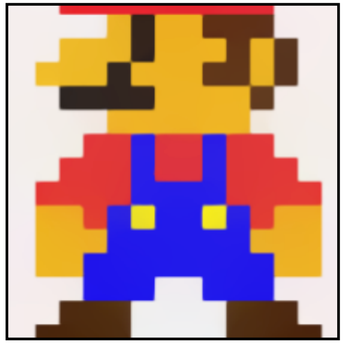
Translated Result

NOT Realistic Enough!

We wanted 3D-like output

Limitation

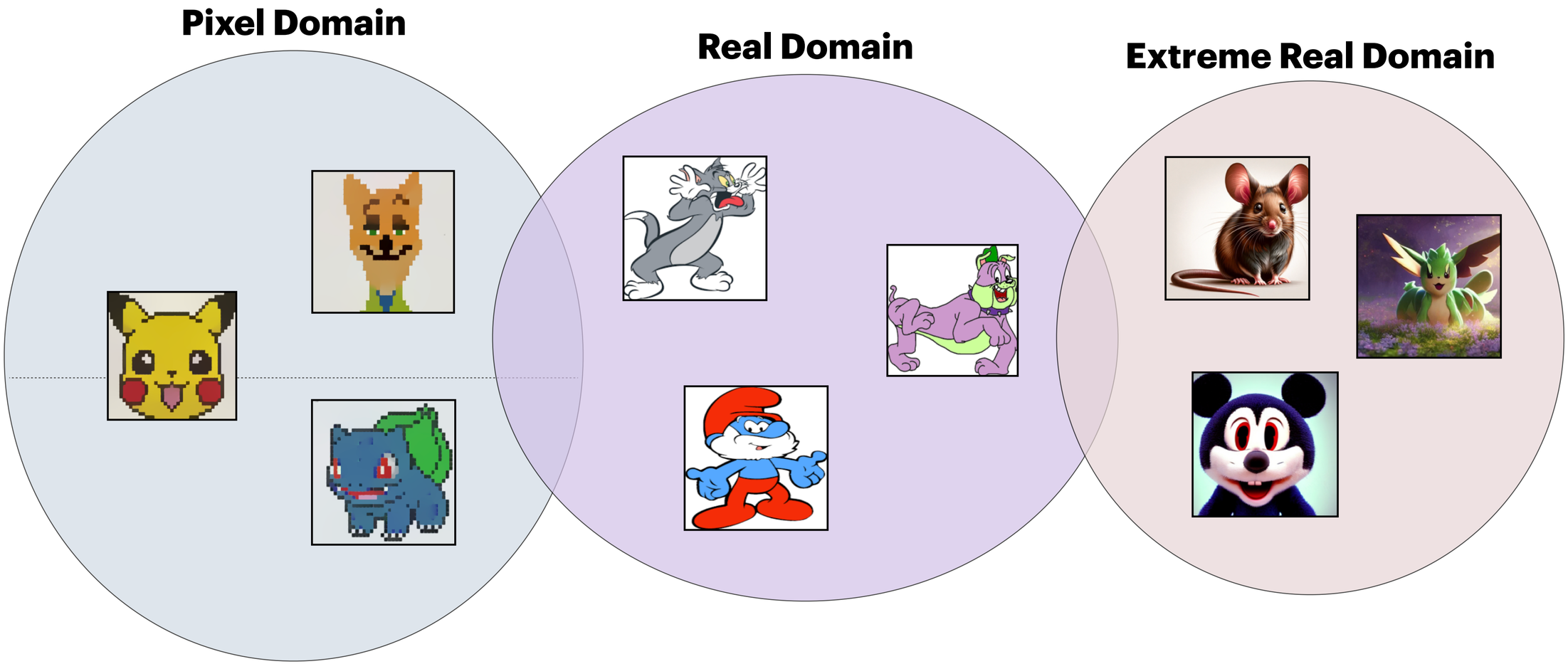
Pixel image



Translated Result

Maybe due to dataset

Approach 4 : Dataset Reconstruction



Result

Pixel image

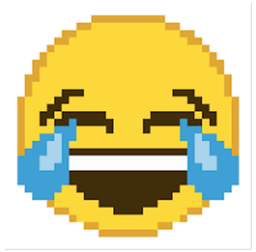
Translated Result

Data Insufficient
(about 100 images)

Domain Gap

background noise

Approach 5 : image-text-image



→

VQA Captioning

→

text2image

→

"describe the character in detail"

"generate prompt for generating realistic image of the character"

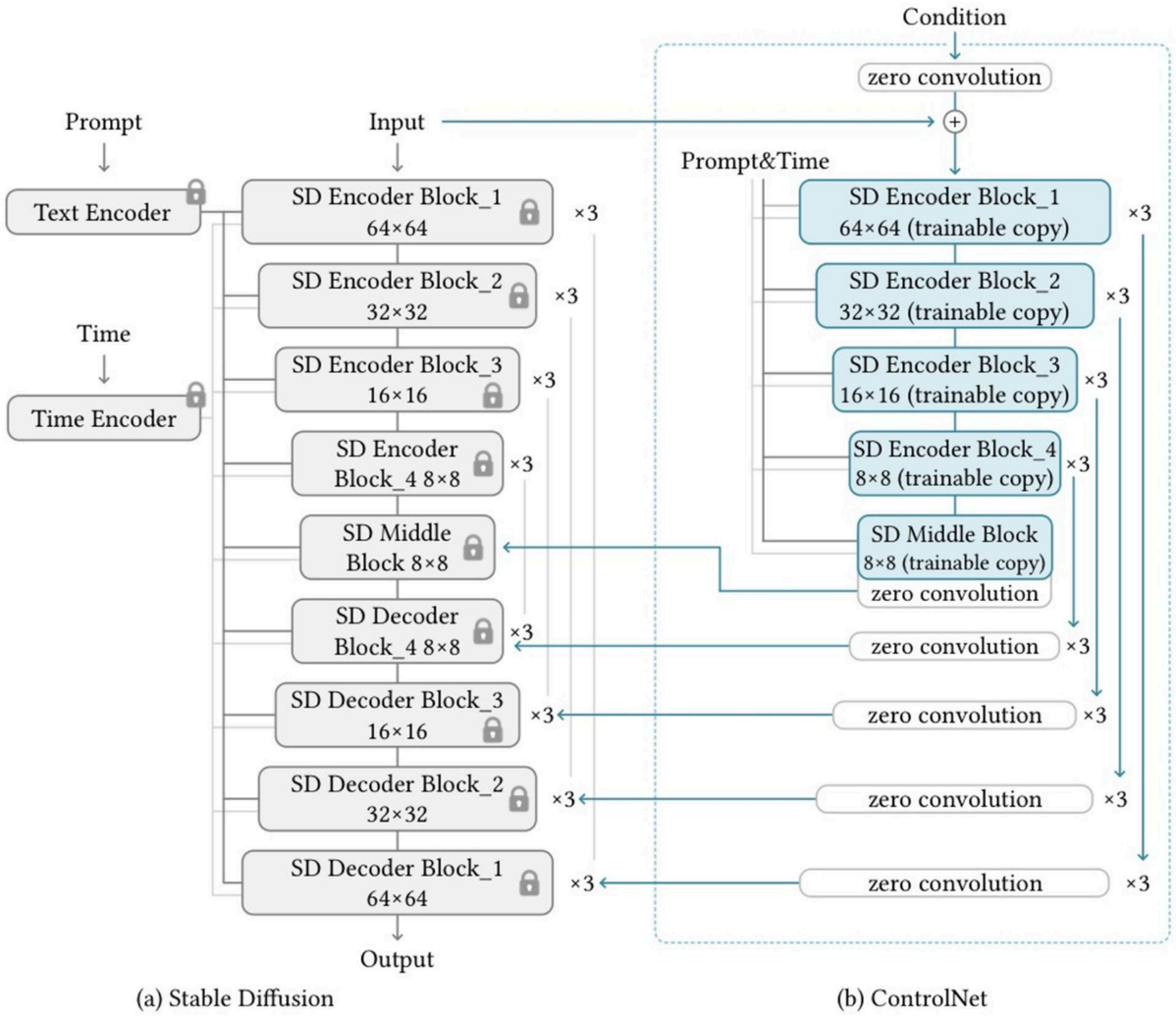
Result

" The image features a skeleton character holding a sword and a shield, standing in a defensive stance. The skeleton is positioned in the center of the scene, with the sword held in its left hand and the shield in its right hand. The character appears to be a warrior or a knight, ready to face any challenges that come its way. The scene is set against a dark background, which adds to the dramatic atmosphere. "

limitation

Input and output
not aligned

Approach 5 : conditioned upscaling

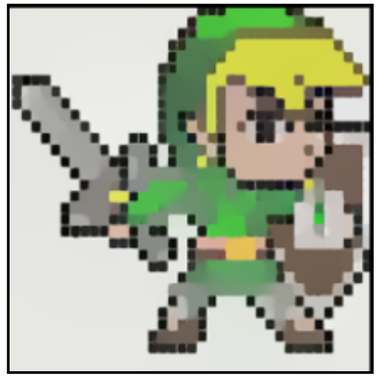


Adding Conditional Control to Text-to-Image Diffusion Models

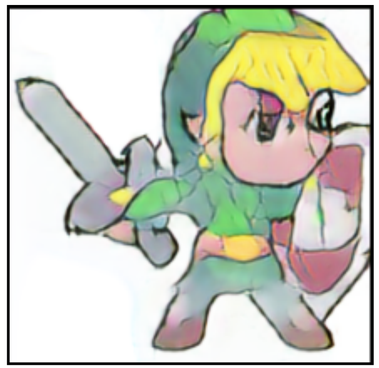
Controlnet utilizes generative prior of diffusion model in various conditioned generation tasks

Result

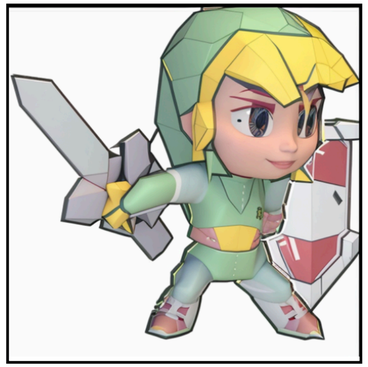
Pixel image



Translated image

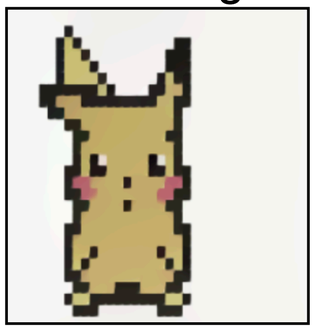


upscaled image

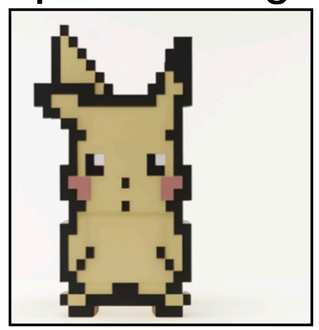


- positive prompts : 3D
- negative prompts : ((unrealistic)),
low quality, blur, digital art, worst quality

Pixel image

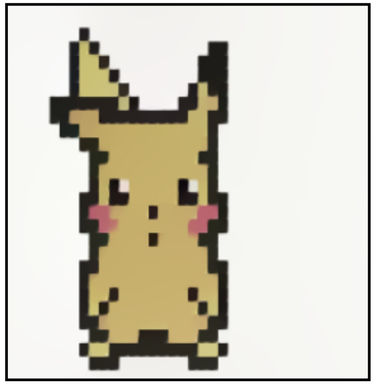


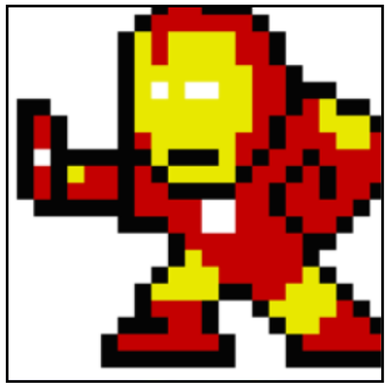
upscaled image



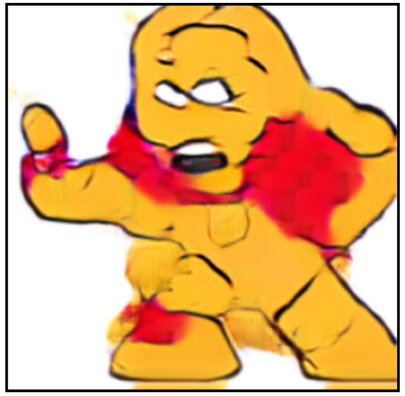
Result

Pixel image

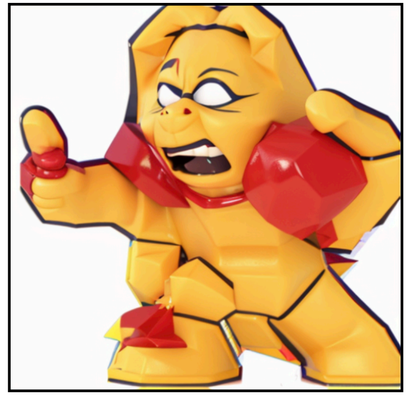




Translated image



upscaled image



quality highly relies on translation quality

Thank you!

디비디비딥이실 팀 : 박수연, 이성현, 이우흥, 남세현