CS 594 Project Report :
GANs with Unpaired Dataset

Chieh-Hsi Lin 670777349
Hyeonghwan Kwon 676630235

I. Abstract

Image-to-image translation is the task of mapping an image from a source domain to a target domain, and it is applicable in different fields like colorization, super-resolution, style transfer, domain adaptation, and data augmentation making this a popular topic these days. In this paper, we did the research review on solving this task with generative adversarial networks(GAN), and for the training dataset instead of using an easily accessible paired dataset, we focused on the circumstance when encountering the unpaired dataset. Upon these conditions, we chose three GANs: CycleGAN, AttentionGAN, and U-GAN-IT among all kinds of GAN models. The detailed analysis supports that CycleGAN is capable to deal with image-to-image translation with the unpaired dataset, and AttentionGAN uses the same idea as CycleGAN and introduced the built-in attention mechanism in generators part to address the pitfalls from previous model architecture, and for U-GAN-IT, it has the best results among three models. U-GAN-IT proposed a novel classifier and normalization function so that it is capable of dealing with various datasets using the identical network and hyperparameters. The model architecture, optimization function, and analytic results of these three models will be described in this report.

II. Introduction

- **Task**

Image-to-image translation has been applied and studied in many ways these days and the application of it can be widely implemented, therefore, we chose image-to-image translation as our main task, and the generative adversarial network (GAN) model is the backbone dealing with this task. Hence we did lots of research about image-to-image translation with GAN and found that most of the papers are dealing with the translation with paired datasets. For a paired dataset, one-to-one correspondence between input and output is needed, however, in many cases, paired training data will not be available. So, we concluded, “Image-to-Image translation with Unpaired dataset” for this project and worked with horse2zebra dataset.

- **Models**

Among all kinds of GAN models, our approaches are CycleGAN, AttentionGAN, and U-GAN-IT. CycleGAN is the baseline in image-to-image translation with unpaired dataset.

And Attention GAN is the upgraded version of CycleGAN so this model is more robust at geometric changes compared to CycleGAN. As for U-GAN-IT, it performs the best among these three models. This model is good at not only geometric changes but also large shape change translations like cat2dog, selfie2anime.

- **Challenges**

GAN models are built for better performance on appearance changes. In many cases, models did well at appearance changes, on the other hand, they did poorly at geometric changes. Moreover, datasets are unpaired between input and output so it is very hard to train compared to paired datasets. Some problems are caused by the distribution of the training dataset, like the training dataset in the paper we read consists of only horse images. So, when using input images including people and horses, the model is highly likely to change whole features to zebra.

III. Approaches

This section introduces three GAN models that we chose for doing image-to-image translation with the unpaired dataset. In each approach, we discussed the motivation, framework, and results from the paper that was published.

A. CycleGAN

CycleGAN is the model to suggest the baseline of image-to-image translation with the unpaired dataset. Before CycleGAN, many papers tried image translation with unpaired datasets but their performance was not good. On the other hand, CycleGAN got better performance than other previous models by introducing a new loss function.

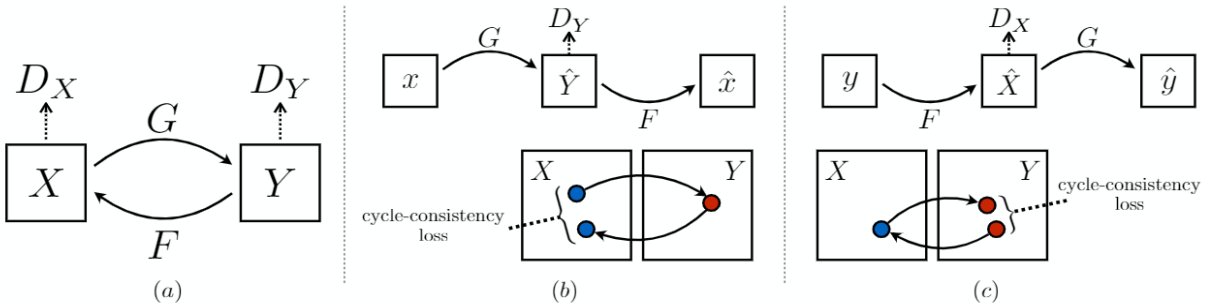


Fig. 1: CycleGAN Framework

In the paper [1], at the beginning, they used two generators and discriminators for the model as you can see in Fig. 1. The key point here is that they used only adversarial loss function for two generators at the beginning. However, it caused a problem that by only using adversarial loss function, two generators generate almost identical images called “mode collapse problem”. Therefore, they added a new loss function for generators which is called “cycle-consistency loss function” as you can see Fig. 1(b), Fig 1(c) above. For cycle-consistency loss function, if we use an image for input X, x is translated to $\widehat{Y}$ by the generator G and then $\widehat{Y}$ is translated to $\widehat{X}$ by generator F. So, the input X image should be

the same as $\widehat{X}$. In the opposite way, input Y is translated to $\widehat{X}$ by the generator F, and then $\widehat{X}$ is translated to $\widehat{Y}$ by the generator G. And this $\widehat{Y}$ image should be the same as the input Y. By adding this “consistency loss function”, the model could resolve the “mode collapsed problem”, showing pretty good results in their paper. We think that it is very meaningful because the model can be built on unpaired datasets since it is not always easy to have available paired datasets, on the other hand, collecting unpaired datasets is relatively easy.

Fig. 2: Limitations on CycleGAN

First, the model is built to focus on appearance changes. As you can see in Fig. 2(A) above, many horses are translated to zebras successfully, however, there are some errors in geometric changes on the house in Fig. 2(A). We also tried the model with other images and still found many errors in geometric changes. Second, they said that if the input image includes other features, not horses, it could cause errors as Fig. 2(B). The reason is caused by the distribution of the training dataset. The training dataset used in the paper does not have any other features except for horses. Therefore, this problem could be resolved by changing the distribution of the training dataset including other features. Third, as we said at the beginning, models with unpaired datasets are very hard to train, and this issue also introduced inside the paper saying its difficulties and the need for supervision in the dataset.

B. AttentionGAN

CycleGAN has shown promising results, but there are some circumstances where it shows visual defects and could not generate expected images. One possible reason is that while doing image translation, the existing model changes unwanted parts of the image. Therefore, Tang et al. [2] has proposed a new method called attention-guided GAN to tackle this problem. The generators in the model are equipped with a built-in attention module so that they can learn both the foreground and background of the input image. It uses two types of attention systems. One is foreground attention to focus on foreground regions and the second is background attention to focus on background regions. By doing so, attention GAN can focus on the most discriminative foreground objects and minimize the change of the background of the image. In this paper, they

presented two attention-guided generation schemes; scheme I is the original model, and scheme II is the refined version showing more precise results.

1. **Scheme I**

The model learns two mappings between domain X and Y via two generators. Each generator takes a three-channel image as the input and generates one content mask and one attention mask to find the discriminative semantic objects from the image. And fuse the attention and the content masks to obtain the final generation. Moreover, an attention-guided discriminator is also introduced, which is structurally the same as the vanilla discriminator but also takes the attention mask as the input.

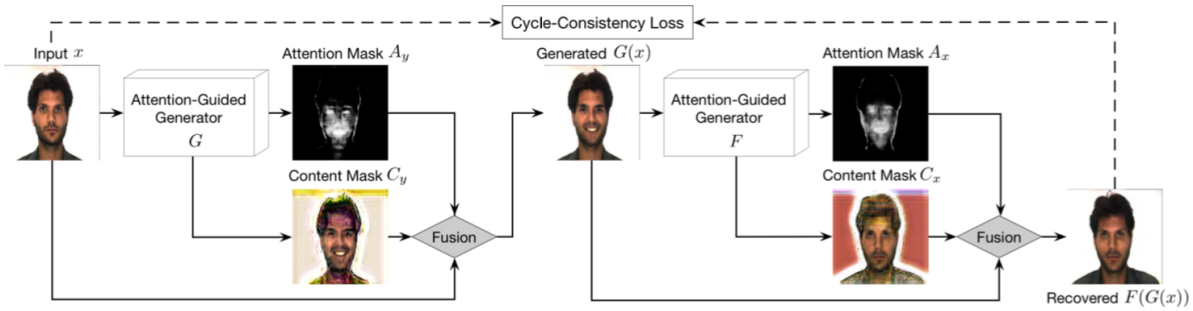


Fig. 3: Framework of the attention-guided scheme I took from [2]

The framework of scheme I can be found in Fig. 3 where attention masks A_x and A_y define intensity per a pixel specifying to which extent each pixel of the content masks C_x and C_y will contribute to the final rendered image. For input x, the higher intensity in the attention mask A_y means the larger contribution for changing the input image, and the content mask C_y creates images with the clear dynamic area and unclear static area. Therefore, the static area can be enhanced by $x * (1 - A_y)$ and the target image $G(x)$ is obtained using the following formula, and generator F is vice versa:

$$G(x) = C_y * A_y + x * (1 - A_y) \quad (1)$$

To regularize the mappings, same as CycleGAN, the cycle-consistency loss and adversarial loss are calculated based on the assumption that translates from one domain to the other and back again, we should arrive at where we started. In addition, a novel attention-guided GAN loss (L_{AGAN}) for training the attention-guide discriminators is measured as well as attention loss and pixel loss to get a complete optimization objective. Since there is no ground truth for annotation masks, they are learned from the resulting gradients of both generators, discriminators, and the rest of losses, attention loss is needed. And pixel loss takes L1 distance as loss measurement and is adopted between the inputs and generated images to reduce changes and constrain the generator in scheme I. In

Fig. 5 scheme I has shown good results when dealing with the task where source and target domain have largely overlapped similarity like facial expression, but it did not show expected results applying horse2zebra dataset.

2. Scheme II

To deal with the drawbacks in scheme I, scheme II has two separate sub-networks so that attention and content masks do not generate from the same network using the same parameters. Unlike scheme I only generates foreground attention masks, in scheme II foreground and background attention masks are both generated and learned simultaneously. Moreover, instead of learning a single mask, here also produced multiple attention masks and content masks. The framework of scheme II shows in fig. All intermediate masks are taken into consideration and fused to get the target image using the following formula:

$$G(x) = \sum_{f=1}^{n-1} (C_y^f * A_y^f) + x * A_y^b \quad (2)$$

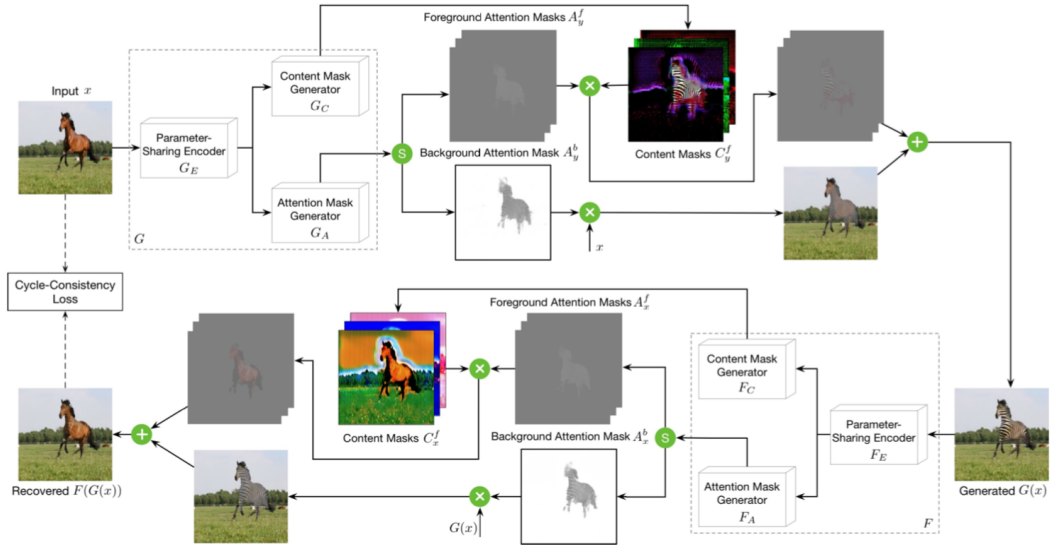


Fig. 4: Framework of the attention-guided scheme II taken from [2]

Likewise, to regularize the mappings, scheme II is also using cycle-consistency loss and adversarial loss. However, the attention-guided discriminator did not use scheme II since there is no improvement after applying it. The possible reason is that the updated generators have enough ability to learn the most discriminative content between the source and target images. Instead, identity loss is calculated to ensure the color distribution of the input image and the output image are similar.

The results in Fig. 5 display the progress of using the second model, although there are still some circumstances, like Fig. 6, where the attention-guided GAN model

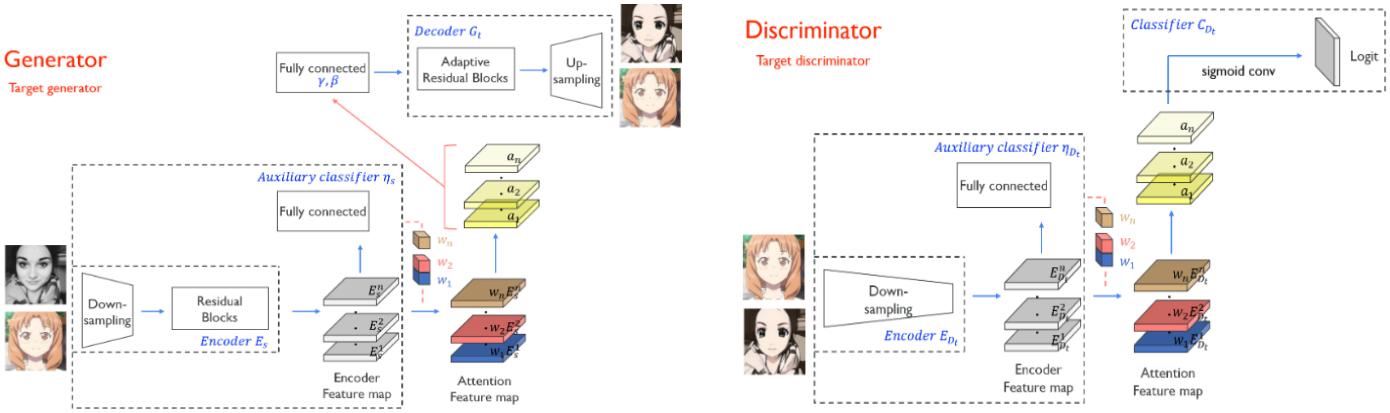
generates confusing attention masks leading to a false result. In conclusion, the outcome of this model still shows promising ability.

Fig. 5: Comparison Results of the Proposed Scheme I and II Taken from [2]

(a) Input Image (b) Attention Mask (c) Generated Image

Fig. 6: Circumstance that Attention GAN has False Attention Masks and False Generated Image

C. U-GAN-IT



(a) Generator (b) Discriminator

Fig. 7: Overview of U-GAN-IT Taken from [3]

Our previous approach, AttentionGAN, has improved in geometric changes by preserving the background of the input image, however, this model is still bad when

encountering a task that includes huge amounts of changes between domains. In the case of horse2zebra, it has to change the texture and color of their features. On the other hand, if the model has to do a big amount of changes like cat2dog and selfie2anime, AttnetionGAN is bad at doing it. Kim et al. [3] introduced the novel U-GAN-IT model to address this problem.

For the generator part, this model introduces a new attention module and a new normalization function. As in Fig. 7(a), for the Encoder part, this model uses Down-sampling and Residual Blocks here. They want to apply each feature map from Residual Blocks with corresponding weights. To do this, they added a new neural network called Auxiliary Classifier and trained the weights of this auxiliary classifier and the output of the classifier is used for the CAM Loss function, eq. 3. The more the model trains, the output of the classifier should be close to 1.

$$L_{cam}^{s \to t} = -\left(E_{x \sim X_s}[log(\eta_s(x))] + E_{x \sim X_t}[log(1 - \eta_s(x))] \right) \quad (3)$$

From the paper, with CAM Loss function, the model can apply Encoder feature map with different weights which could help the model distinguish which part should be focused on and which part should be ignored. Moreover, they used two other fully connected layers to apply a new normalization function. Generally, each model uses one normalization function for each layer, but in their Decoder part, the model uses their normalization function, adaptive layer instance normalization (AdaLIN), eq. 4, which is a special function using both Instance Normalization and Layer Normalization. By using these FCNs, this model could get gamma and beta, which are the output of FCNs, for this normalization function. And their model uses these values for AdaLIN normalization.

$$AdaLIN(\alpha, \gamma, \beta) = \gamma \cdot \left(\rho \cdot \widehat{\alpha_I} + (1 - \rho) \cdot \widehat{\alpha_L} \right) + \beta \quad (4)$$

The discriminator part, like Fig. 7(b), has no big difference from the generator model. Instead, this model uses only Down Sampling for Encoder and only FCN classification model to get an output. The important thing is that the CAM loss function for the Auxiliary classifier in the Generator is different from the loss function for the Auxiliary classifier in the Discriminator.

For evaluation, we couldn't find the pre-trained model for U-GAN-IT using horse2zebra dataset. So, we trained the model ourselves with our horse2zebra dataset and compared the results with other models using our results and the results in the paper [3]. By looking at the results in the paper, although they said that the model has good performance on big shape image translation like cat2dog, it still has many errors on geometric changes. And it is true that it has better performance on doing big shape image translation, but this model uses many FCNs for both the Discriminator and Generator, which means that is very computationally expensive.

IV. Comparison

Section III detailed three chosen GAN models with their architecture, motivation, novelty, and some translated images presented in its original paper, and here combined three models together and made the comparison. For quantitative evaluation, we used the kernel inception distance(KID), which computes the squared Maximum Mean Discrepancy between the feature representations of real and generated images. KID has an unbiased estimator making it more reliable, especially when there are fewer test images than the dimensionality of the inception features. The lower KID indicates that the more shared visual similarities between real and generated images, therefore, if the image is well translated, the KID will have a small value. As shown in Table 1., where values are taken from [3], besides our mainly horse2zebra, they also implemented other datasets to have comprehensive analysis. From the results, CycleGAN is taken as the base standard compared with the other two models. As we mentioned, AttentionGAN aims to fix the defects for CycleGAN, and indeed it shows improvements while dealing with horse2zebra, however, when it comes to big shape changes, it does not have optimistic results. On the contrary, U-GAN-IT shows the lowest KID in all kinds of datasets meaning it is not only robust while dealing with style translation but also excellent in shape changes like selfie2anime. Most importantly, U-GAN-IT has applied the identical model structure and hyperparameters for all these datasets without any presetting making it the most appealing model compared to the other two models.

Table 1. Kernel Inception Distance * $100 \pm std$, values are taken from [3]

Model/Dataset	horse2zebra	selfie2anime	cat2dog	photo2potrait
CycleGAN	8.05±0.72	13.08±0.49	8.92±0.69	1.84±0.34
ATNGAN	7.58±0.71	14.63±0.55	9.84±0.79	2.33±0.36
U-GAN-IT	**7.06±0.80**	**11.61±0.57**	**7.07±0.65**	**1.79±0.34**

*ATNGAN = AttentionGAN

With all the visual and quantitative analysis, to have a better understanding of these three models, we made a table to conclude the pros and cons of each model, in Table 2. In the table, we listed the most discriminative difference between them, although there are still other things that can be mentioned.

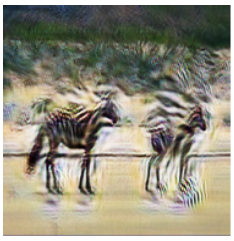
Table 2. Comparison of CycleGAN, AttentionGAN, and U-GAN-IT

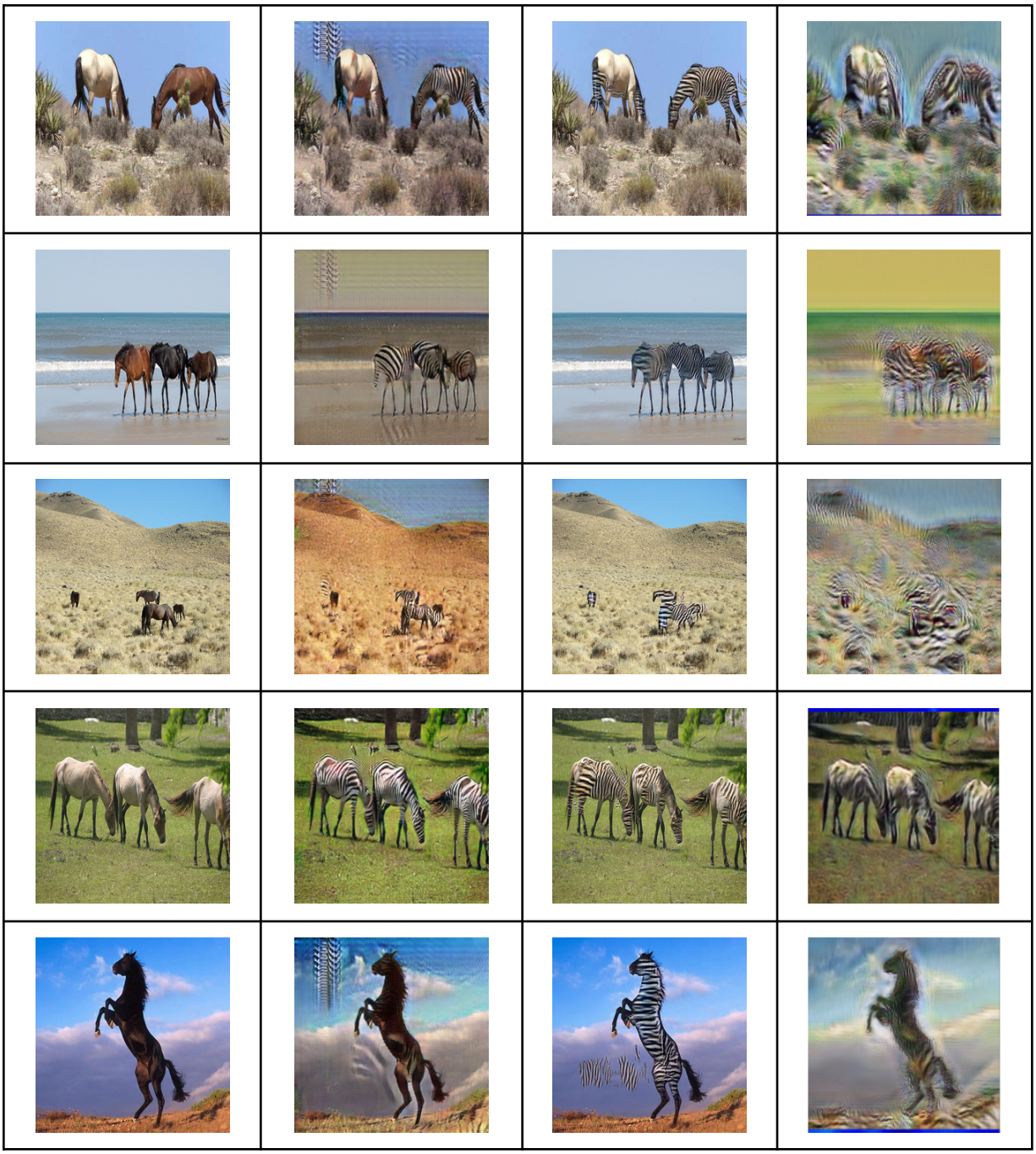
	Pros	**Cons**
CycleGAN	- Baseline model for img2image translation - Solution to Unpaired Dataset	- Many pitfalls can be improved like performance is not good as other models, poorly dealing with geometric changes, etc
AttentionGAN	- Generate good translated images while persevering the irrelevant regions or background features from input images	- Possibility of creating false attention masks leading to unexpected generated images - Poor performance with big shape changes task like cat2dog, selfie2anime
U-GAN-IT	- Outstanding performance with big shape changes - Capability of applying all kinds of datasets with reasonable outcomes	- Expensive Computation cost due to complex model structures (the use of FCNs)

V. Qualitative Results

Besides the output translated images obtained from its original paper, we provided some custom results in this section. Since GAN is a relatively complex model compared to other deep learning models requiring large computation cost during training, and we did not have efficient hardware to train all models from scratch, therefore, the results below used the pre-trained model provided by the authors who published the paper. However, for the U-GAN-IT model, we could not find the pre-trained model, so the translated images generated from the model were trained by ourselves. With limited hardware, our trained U-GAN-IT model was not the 'well-trained' model as presented in its original paper, hence the results in Table 3. were not able to show competitive results as described in the paper using our custom horse dataset.

Table 3. Generated Images Applied on Three Chosen GAN Models Dealing with Custom Horse2zebra Dataset

Original Image	Cycle GAN	Attention GAN	U-GAN-IT
			



VI. Conclusion

When dealing with image-to-image translation, CycleGAN is the first-come solution and it presented the first concept of cycle-consistency loss and adversarial loss function which are the essential step for further studies of this topics. Being the baseline model, although it has promising outcomes, the later models, like AttentionGAN and U-GAN-IT, show even better performance solving the issues CycleGAN has and enhances the ability and variability of their model. And AttnetionGAN presented an add-on mechanism to the

CycleGAN-based model to generate better results dealing with horse2zebra dataset, however, when it comes to the task dealing with big size changes, the model could not perform expected results. Lastly, U-GAN-IT introduced a new architecture to get better performance on big size image translation and also maintained good results on small size image translation in the meanwhile. Even though the complexity of the architecture is expensive in computation, considering the quality of the results and its accessibility in miscellaneous tasks, it is the top choice for most people.

Despite the strengths that GAN models have demonstrated in addressing image-to-image translation, there are still difficulties. A major issue is that GAN models are easy to suffer "Model collapse problem" generating the same translated image. In addition, GANs contain several sub-models inside, generators, and discriminators, making the model hard to train. Even though many GAN models including the models we mentioned have improved in many ways, improvements are still needed. And if we could get more comprehensive unpaired datasets, the model could be trained more thoroughly and provide better results.

VII. References

[1] Jun-Yan Zhu and Taesung Park and Phillip Isola and Alexei A. Efros, "*Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks"* arXiv 1703.10593, 2020

[2] Hao Tang, Hong Liu, Dan Xu, Philip H. S. Torr and Nicu Sebe, "*AttentionGAN: Unpaired Image-to-Image Translation using Attention-Guided Generative Adversarial Networks*", in CoRR, abs/1911.11897, 2019

[3] Junho Kim and Minjae Kim and Hyeonwoo Kang and Kwanghee Lee, "*U-GAT-IT: Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization for Image-to-Image Translation*" arXiv 1907.10830, 2020