**Theoretical paper:**

**Unsupervised Full-Image-to-Image Translation**

Lim Zhen Yang

Singapore Polytechnic (SP)

ST1504 Deep Learning

Author Note

Singapore Polytechnic, ST1504 Deep Learning, February 2024

Abstract

The field of image-to-image (I2I) translation has seen substantial progress, particularly in applications requiring the transformation of images from one domain to another without direct paired examples. This paper introduces a novel unsupervised full-image-to-image translation (UFI2IT) approach that addresses the challenges posed by large domain shifts and the translation of complex scenes with multiple objects. By integrating attention mechanisms and cross-domain adaptation strategies into Generative Adversarial Networks (GANs), our model demonstrates significant improvements in maintaining the integrity and realism of translated images across diverse domains. Our methodology combines self-attention GANs for capturing long-range dependencies, spatial attention for focusing on discriminative regions, and cross-domain adaptation to bridge the domain gap effectively. The use of cycle consistency loss ensures the preservation of content during translation. This paper reviews related works, including advancements in GANs, attention mechanisms, and domain adaptation, and presents a comprehensive SPA-SA-CDA GAN architecture that has the potential to surpasses current state-of-the-art models in unsupervised I2I translation tasks. Our findings highlight the model's potential in various applications, despite limitations such as computational demands and training complexities. Future research directions are suggested, including the exploration of datasets and industry applications, to further advance the field of I2I translation.

Keywords: Image-to-Image Translation, Unsupervised Learning, Generative Adversarial Networks, Self-Attention, Spatial Attention, Cross-Domain Adaptation, Cycle Consistency Loss.

Introduction

Image-to-image (I2I) translation is a process of converting an image from one category to another while keeping its core content intact. For instance, it can turn a picture of a horse into that of a zebra. However, these two image categories have distinct characteristics, such as zebras having stripes and horses having larger tails than zebras. Many I2I translation methods rely on supervised learning, where pairs of source and target images are available. Nevertheless, obtaining such paired data can be expensive or infeasible for various applications. Hence, there is a growing interest in unsupervised approaches, where source and target image sets are unrelated, lacking any paired examples between them.

In general, an effective image-to-image translation technique must identify specific regions of interest within the input image and learn how to transform these regions into the target domain. In unsupervised settings without paired images, it becomes crucial to focus on these regions subject to transformation. This task of locating areas of interest is particularly important in applications where translation should only be applied to objects within the image rather than the entire image. For example, to convert an "orange" image into the target category "apple," one must first identify the oranges in the input image and then change them into apples. Many image-to-image translation tasks share this characteristic, like the well-known horse-to-zebra dataset, apple-to-orange dataset, and lion-to-tiger dataset. These tasks involve translating a single specific object, which tends to be simpler compared to full-image-to-image translation tasks, such as the GTA-to-cityscapes dataset or the day-to-night dataset. The latter involves multiple diverse objects that require careful attention. As a result, current unsupervised image-to-image translation (UI2IT) networks often produce results where the translated images are visually similar to the target domain in terms of style but may contain inaccuracies or corrupted objects, as shown in Figure 1. The sign board on the top right is missing in CycleGAN (Jun-Yan Zhu, 2017) and the wrong colour in SPA-GAN (Hajar Emami, 2020)

A road with cars on it

Description automatically generatedFigure 1 Cityscapes-to-GTA dataset. Input image on the left, CycleGAN output in the middle, SPA-GAN output on the right.

In this paper, we are proposing a solution for the task of full-image-to-image translation with many objects. Recent research shows that attention mechanism is helpful to improve the performance of generative adversarial networks (GANs) in image-to-image translation applications. A task like this also involves a large domain shift. As stated earlier, current models do not perform well in these scenarios, so we will be looking at how we can use domain adaptation in the field of GANs.

# Related Works

Recently, there have been great interest in GANs. In 2017, Zhu et al. introduced CycleGAN [6], a network able to translate an image from one domain to another using cycle consistency. Nevertheless, despite this method breaking the paired constraint, it requires a lot of training iterations and suffers from instability. Since then, many models have emerged trying to improve this result. One increasingly popular technique is the use of attention.

**GANs**

Generative Adversarial Networks (GANs) have been highly successful in various image-related tasks, such as image-to-image translation, image super-resolution, and text-to-image synthesis. However, GAN training is known for its instability and sensitivity to hyperparameters. Many efforts have aimed to address these issues, including designing new network architectures, modifying learning objectives, new loss functions (Bottou, 2017), adding regularization techniques, and using heuristic methods. Recently, (Takeru Miyato, 2018) proposed a method that limits the spectral norm of discriminator weight matrices to improve class-conditional image generation on ImageNet, especially when combined with a projection-based discriminator.

**Image to Image translation with attention GANs**

Recently, attention mechanisms have become an integral part of models that must capture global dependencies for image to image translation. Vaswani et al. (Ashish Vaswani, 2023) demonstrated that machine translation models could achieve state-of-the-art results by solely using a self-attention model. Wang et al. (Xiaolong Wang, 2018) formalized self-attention as a non-local operation to model the spatial-temporal dependencies in video sequences. Then, self-attention was introduced (Odena, 2019) and achieved state-of-the-art results. It enables both the generator and the discriminator to efficiently model relationships between widely separated spatial regions. Then to introduce attention to the whole GAN network, there is SPA-GAN (Hajar Emami, 2020) which computes the attention in its discriminator and use it to help the generator focus more on the most discriminative regions between the source and target domains.

**Cross Domain Adaptation**

A common problem with deep learning is that even state-of-the-art deep learning models still suffer from significant performance drops even when the testing distribution slightly drifts from the training distribution. In our case instead of testing, when the input domain images and target domain images are very different, GANs are unable to perform well. Unsupervised domain adaptation is used to reduce this domain shift. AugGAN (Sheng-Wei Huang, 2018) proposes a structure-aware image-to-image translation network, which allows us to directly benefit object detection by translating existing detection RGB data from its original domain other scenarios.

**Methodology**

**A diagram of a method

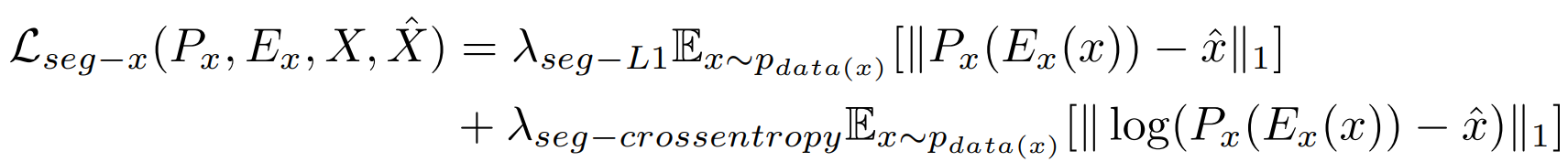
Description automatically generated**The goal of image-to-image translation is to learn a mapping G from a source domain to a target domain , where and are the number of samples in domains X and Y, respectively. In an unpaired setting, two inverse mappings are learned simultaneously through the cycle consistency loss (Jun-Yan Zhu, 2017)

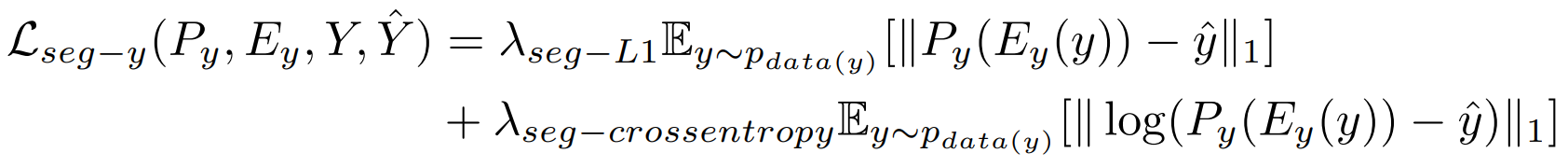
Figure 2 Proposed SPA-SA-CDA GAN

The model begins with taking in input images X/Y’ which refers to images from domain X with a target domain of Y. Similarly, Y/X’ refers to images from domain Y with a target domain of X’. Next, it is passed to residual blocks and then converted to a vector Z. This part of the model was inspired by AugGAN, where it actively guides the encoder networks to extract context-aware features. Z is then passed into the generator made up of self-attention blocks (Odena, 2019) and is up sampled to reconstruct the image. The generated images are compared with the real images by calculating the cycle consistency loss, and passed into the discriminator, made up of self-attention blocks and down sampling. The discriminator then calculates the GAN loss, as well as generate the attention map which is then passed into the input of the generator other. The same applies to the other input image of the other domain. This double GAN architecture is commonly used in UI2IT tasks.

**Why cross domain adaptation**

As mentioned earlier we want to close the domain gap between input domain X and target domain Y. This is to get our model more “comfortable” by generating images from the target domain given images from the input domain. We actively guide the encoder networks to extract context-aware features by regularizing them via segmentation subtask so that the extracted 256-channel feature vector contains not only mutual style information between X and Y domains, but also the intricate low-level semantic features of the input image that are valuable in the preservation of image-objects during translation. The segmentation loss is formulated as:



****

**Why self-attention**

A close up of a date

Description automatically generatedIn the SAGAN paper by Odena (2019), they conducted a comparison between SAGAN and state-of-the-art GAN models by Takeru Miyato (Takeru Miyato, 2018) for generating class-specific images on the ImageNet dataset. Table 2 presents the results, showing that SAGAN outperforms the other models in terms of Inception score, intra FID, and FID. Specifically, SAGAN achieves a substantial improvement in the Inception score, raising it from 36.8 to an impressive 52.52. Additionally, SAGAN demonstrates lower FID (18.65) and intra FID (83.7) scores, indicating its superior ability to capture the original image distribution. This improvement is attributed to the incorporation of the self-attention module, allowing SAGAN to model long-range dependencies among different regions of the images more effectively.

Table 1 Comparison of the proposed SAGAN with state-of-the-art GAN models

A collage of different birds and animals

Description automatically generatedFigure 3 shows some comparison results and generated images for representative classes of ImageNet. We observe that our SAGAN achieves much better performance (i.e., lower intra FID) than the state-of-the-art GAN model (Takeru Miyato, 2018) for synthesizing image classes with **complex geometric or structural patterns**, such as goldfish and Saint Bernard. This additional improvement of synthesizing image classes with complex geometric or structural patternsis of utmost importance in our solution for unsupervised full image-to-image translation (UFI2IT).

Figure 3 128x128 example images generated by SAGAN for different classes. In the leftmost column, the intra FID of our SAGAN (left) and the state-of-the-art method (Takeru Miyato, 2018) (right) are listed.

**Why spatial attention**

Incorporating the attention mechanism into image-to-image translations can help the generative network to attend to the regions of interest and produce more realistic images. SPA-GAN achieves this by explicitly transferring the knowledge from the discriminator to the generator to force it to focus on the discriminative areas of the source and the target domains. We know that the addition of explicitly transferring the attention map to the generator help because in their ablation study, this ablation study showed that feature map loss works better when spatial attention is employed and vice versa, and when comparing the removal of both feature map loss and spatial attention (which are the two main innovations of SPA-GAN) to SPA-GAN, SPA-GAN performed significantly better; accuracy of 71.80% and 87.21% respectively. This certainly displays the significant improvement spatial attention has.

Looking at the qualitative results, as shown in the fourth column of Fig. 3, in the orange→apple translation, the SPA-GAN attention map computed in the discriminator focuses on both the shape and texture of the generated and real apple images to correctly classify the inputs. It has higher values around the boundaries and on the top part of the oranges while AGGAN (Yan, 2019) attends on the whole oranges. This higher level of detail would be essential for images with multiple objects requiring attention, the main problem this paper want to tackle.

A collage of images of animals

Description automatically generated

Figure 4 Comparison between the attention maps generated by the attention network in AGGAN [7] (the fourth row) and the attention maps computed in SPA-GAN (the second row) on different image samples. SPA-GAN attention maps have higher activation values in the most discriminative regions between the source and target domains. Note, for example, in column one AGGAN generates a disconnected attention map for zebra while SPA-GAN attends on all the zebra patterns. In column four, AGGAN attends on the whole oranges while SPA-GAN has higher attention values around the boundaries and the top part of the oranges. Also note that the attention weights are generally greater than 0.5 as shown in the weight distributions in the third row. Thus, the generator will not produce agnostic outputs for the non-discriminative regions.

A close-up of a number

Description automatically generated

It is important to note that cycle consistency loss will be used in this model as all the papers reviewed mention how proven and effective it is in ensuring that the generator does not generate random images, forgetting about the translation task at hand. The SPA-GAN modified cycle consistency loss is as follows:

**Discussion**

This paper has reviewed past works and has detailed a potential solution for this novel problem. It does not however have the specific implementation details such as layers, encoders, hyperparameter specifics. But by reviewing the results of these incredible papers, with the right execution this model would have the potential to surpass existing state-of-the-art models.

**Further Research**

A collage of different images of a house

Description automatically generated There have not been any popular full image to image translation tasks that involve massive domain shifts and multiple important objects that require transfer. (Bellaaazzzzz, n.d.) is a dataset on hugging face with 5000 images of interior design wireframes, paired with their finished product (refer to Fig 5). I recognized that this dataset would be applicable to a supervised I2IT task whereas this paper is all about unsupervised. This is where it comes in. 5000 images for such a task is way too little. One would need to make use of existing unlabeled, unpaired interior design wireframe datasets and completed design datasets. The success of such a research has direct industry applications, further pushing the boundaries of machine learning.

Figure 5 (Bellaaazzzzz, n.d.) Interior design wireframe dataset

**Limitations**

The introduction of a model of such magnitude, as presented in this paper, brings with it a set of inherent limitations. Notably, its immense size and complexity result in significant computational expenses, demanding access to high-performance GPUs or TPUs. Training times for such models are substantially longer, hindering rapid experimentation and prototyping; wall-clock time would be longer than state-of-the-art model (Takeru Miyato, 2018). Moreover, the colossal memory requirements restrict their deployment on memory-constrained devices. Fine-tuning and adapting these models to specific tasks become arduous endeavors due to the vast number of parameters involved. Furthermore, their ability to generalize to new data may be limited, necessitating extensive data augmentation and adaptation efforts. These limitations, coupled with challenges in reproducibility, interpretation, and deployment, underscore the trade-offs involved in employing such computationally intensive models for various applications and research endeavors.

# Conclusion

In this paper, we have presented an innovative SPA-SA-CDA GAN architecture, leveraging a comprehensive review of state-of-the-art models. Our research has been dedicated to addressing a previously underemphasized challenge within the realm of image-to-image translation tasks. The potential of our architecture is substantial, offering promising and versatile applications in various domains. Our work contributes to the advancement of image generation and translation techniques, opening exciting opportunities for future research and practical applications.

# References

Ashish Vaswani, N. S. (2023). *Attention Is All You Need.*

Bellaaazzzzz. (n.d.). *Hugging Face*. Retrieved from Hugging Face: https://huggingface.co/datasets/Bellaaazzzzz/wireframe

Bottou, M. A. (2017). *Wasserstein GAN.*

Chintala, A. R. (2016). *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks.* ICLR.

Hajar Emami, M. M. (2020). *SPA-GAN: Spatial Attention GAN for Image-to-Image Translation.*

Jun-Yan Zhu, T. P. (2017). *Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks.*

Kim, Y. A. (2018). *Unsupervised Attention-guided Image to Image Translation.*

Odena, H. Z. (2019). *Self-Attention Generative Adversarial Networks.*

Rui Wang, Z. W.-J.-G. (2022). *Cross-domain Contrastive Learning for.* IEEE.

Sheng-Wei Huang, C.-T. L.-P.-Y.-H.-H. (2018). *AugGAN: Cross Domain Adaptation with GAN-based Data Augmentation.* ECCV.

Takeru Miyato, T. K. (2018). *Spectral Normalization for Generative Adversarial Networks.* ICLR.

Tao Xu, P. Z. (2017). *AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks.* CVPR.

Xiaolong Wang, R. G. (2018). *Non-local Neural Networks.* CVPR.

Yan, H. T. (2019). *Attention-Guided Generative Adversarial Networks for Unsupervised Image-to-Image Translation.*