IMAGE-TO-IMAGE TRANSLATION: A LITERATURE REVIEW

Yuanwen Yue
D-BAUG
ETH Zurich
yuayue@student.ethz.ch

December 18, 2020

ABSTRACT

Image-to-image translation aims to learn the mapping between an input image from a source domain and an output image from a target domain. The recent advances in deep learning have greatly improved the quality of image-to-image translation methods for many applications. This report discusses and analyses some representative papers of image-to-image translation in both supervised setting and unsupervised setting, and their applications in the field of geomatics. For each reviewed paper, a critical evaluation about its strength and weakness is also provided. Finally, the reviewed papers are compared and open challenges are discussed.

1 Introduction

Image-to-image (I2I) translation refers to the task of translating one image from one domain into another domain. Many problems in image processing and computer vision can be viewed as "translating" an input image into a corresponding output image, such as semantic maps to real images, sketches to photos, gray-scale to color images, aerial photos to maps, etc. This problem can be studied in supervised and unsupervised settings. As shown in Figure 1, in the supervised setting, we are given pairs of corresponding images from the two domains. In the unsupervised setting, we only have two independent data sets of images where one consists of images in one domain and the other consists of images in another domain. The correspondence between the two data sets is not known. Due to the lack of paired images, the unsupervised I2I translation task is much harder because there are infinitely many mappings existing between the two unpaired image domains. However, unsupervised I2I translation is more practical because for many tasks it is not easy, or even possible, to obtain such paired data in the two domains, like cross-city street view translation or male-female face translation.

In this report, representative works of I2I translation in both supervised setting and unsupervised setting are discussed and analyzed. There have been a lot of works on I2I translation and most of them are based on generative adversarial networks (GANs)[1]. In the supervised I2I translation works, Isola et al. [2] proposed pix2pix to learn the mapping from input images to output images in an adversarial way. Zhu et al. [3] extended pix2pix to BicycleGAN, which can achieve multimodal translation and produce diverse translation results. Instead of using GAN, Chen et al. [4] proposed a single convolutional network for photographic image synthesis from pixelwise semantic layouts. Recently, several unsupervised I2I translation methods have been proposed to learn the mappings between two image data sets without paired training data. Zhu et al. [5] proposed CycleGAN to regularize the learning process with cycle-consistency. Going one step further, Liu et al. [6] proposed UNIT with a shared latent space constraint. It assumes that a pair of images from both domains can to be mapped to the same representation in the latent space. Although CycleGAN and UNIT can achieve unsupervised I2I translation, they are one to one mapping. Ma et al. [7] proposed the exemplar guided & semantically consistent I2I translation (EGSC-IT) network which achieves multimodal translation.

I2I translation have been widely applied in the field of geomatics. This report mainly discusses two translations, i.e. aerial photos to maps, synthetic aperture radar (SAR) images to optical images. Gu et al. [8] explored the bidirectional translation of aerial photos and maps using a GAN-based network. Reyes et al. [9] adopted

CycleGAN to generate alternative SAR image representations based on the combination of SAR images and optical images for training.

This report's structure can be summarized as follows. Representative works of I2I translation in supervised setting and unsupervised setting are discussed and analyzed in section 2 and section 3, respectively. Section 4 reviewed two applications of I2I translation in geomatics, i.e. aerial photos to maps, SAR images to optical images. For each reviewed paper, a critical evaluation about its strength and weakness is provided. Section 5 consists of the comparison of reviewed papers and the outlook of open challenges.

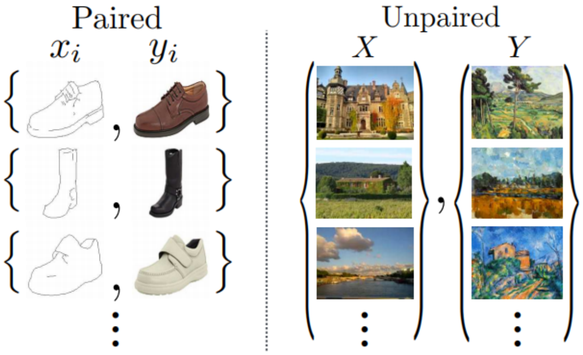


Figure 1: Paired vs. Unpaired. [5]

2 Supervised I2I translation

2.1 Pix2pix: conditional GANs

2.1.1 Summary

Isola et al. [2] investigated conditional adversarial networks (cGANs) as a general-purpose solution to I2I translation problems. Their main idea is to condition on an input image and generate a corresponding output image using GAN. Unlike an unconditional GAN which learns a mapping from random noise vector z to output image y, $G : z \rightarrow y$, conditional GANs learn a mapping from observed image x and random noise vector z, to y, $G : \{x, z\} \rightarrow y$. In this case, both the generator and discriminator observe the input edge map. For the network architectures, they use a "U-Net"-based architecture [10] as the generator and a convolutional "PatchGAN" classifier [11] as the discriminator. Their method is proven to be effective on a variety of I2I translation tasks.

2.1.2 Strength

The main novelty of this paper is to apply conditional GANs to I2I translation. The proposed method is simple and applicable to a variety of I2I translation tasks. Both qualitative and quantitative experiments are carefully designed and well explained. The authors also released a software called pix2pix associated with this paper, which has been widely explored in the Twitter community. Overall, the paper is well written with clear contribution on demonstrating the generality of cGANs on various I2I translation tasks.

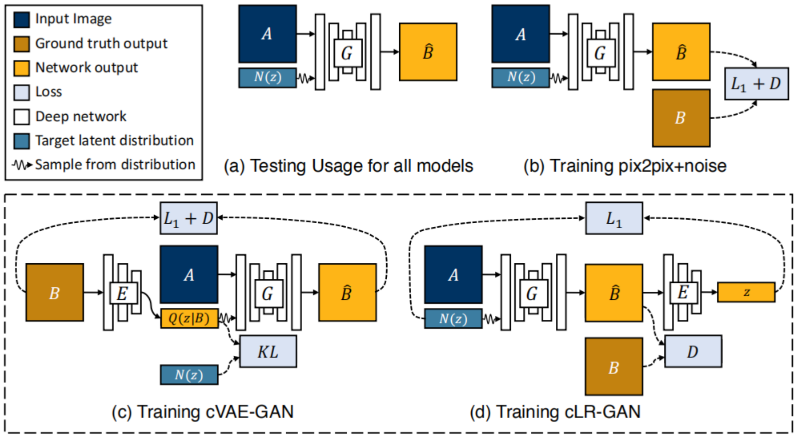
2.1.3 Weakness

The authors tested the effect of varying the patch size N of their discriminator receptive fields, from a 1 × 1 "PixelGAN" to a full 286 × 286 "ImageGAN", it might have been helpful to explain more details about the reason for setting the size interval to 1 × 1, 16 × 16, 70 × 70, and 286 × 286. Besides, in the perceptual validation, the authors only compared their results on the task of map $\leftrightarrow$ aerial photograph with a simple baseline, it would be more convincing if more other methods are compared.

2.2 BicycleGAN

2.2.1 Summary

Many of the mappings in I2I translation tasks are one-to-many in nature. However, pix2pix can only produce a single translation answer. To achieve multimodal translation, Zhu et al. [3] extended pix2pix to BicycleGAN. BicycleGAN can model multimodal distributions and produce both diverse and realistic results. The key idea is to combine multiple objectives for encouraging a bijective mapping between the latent and output spaces. As shown in Figure 2, it first encodes the ground truth image into a latent space which is then used by the generator to help reconstruct the ground truth image. On the other hand, a conditional latent regressor model starts with a randomly drawn noise vector, produces an output and then uses an encoder to attempt to recover the original latent vector. The two procedures blend together to form BicycleGAN. The method is tested on several image-to-image translation problems.



(a) Testing Usage for all models

(b) Training pix2pix+noise

(c) Training cVAE-GAN

(d) Training cLR-GAN

(e) Training BicycleGAN

Figure 2: Overview of BicycleGAN. [3]

2.2.2 Strength

The paper is clearly written and the Figure 2 is helpful for understanding the overall architecture. This paper alleviates the mode collapse problem, which commonly occurred in the prior work. The main novelty of this paper is to add a latent code to generate diverse samples and use latent cycle-consistency to prevent the model to ignore the latent code. The experimental results are convincing and consistent with the aims of the research.

2.2.3 Weakness

The authors built their model on the Least Squares GANs (LSGANs) variant [12], which uses a least-squares objective instead of a cross entropy loss. It would have been useful to know any specific reasons or intuitions behind this decision. The authors only performed human-perception experiments on the Google maps ↔ satellites task. However, it would be more convincing if the human-perception experiments are also performed on other tasks such as sketches to real images, because this task can generate more diverse images in terms of colors, textures etc., than aerial photos.

2.3 Photographic image synthesis without GAN

2.3.1 Summary

Instead of using GAN, Chen et al. [4] proposed a single convolutional network for photographic image synthesis trained in a supervised fashion on pairs of photographs and corresponding semantic layouts. The problem that this paper focuses on can be seen as the inverse of semantic segmentation. Their approach does not reply on adversarial training and synthesizes photographic images by a single feedforward network, trained end-to-end with a regression objective. A special loss is also introduced to train the network to produce a diverse collection of images in one shot. The proposed method is tested on datasets of outdoor and indoor scenes.

2.3.2 Strength

This paper provides an alternative to GAN to realize photographic image synthesis, a typical task of I2I translation. The proposed method is simple and straightforward, and is clearly explained in the paper. The baselines are properly selected and the experiments are well set up, in which the time-limited pairwise comparisons are very interesting.

2.3.3 Weakness

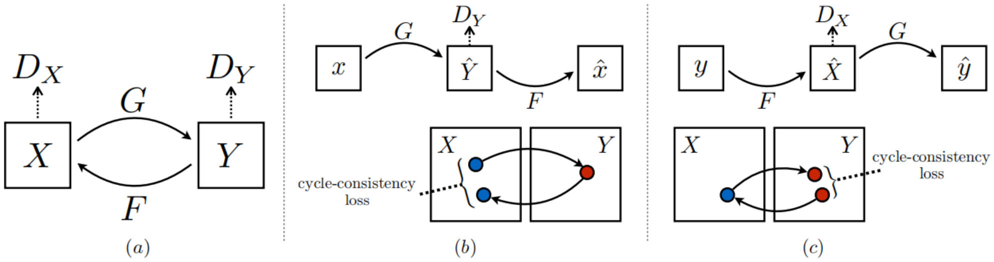
In the figures of qualitative comparison, no ground truth images are provided. It would be better to show the ground truth images so that the results can be better evaluated. Besides, the authors claimed that evaluating "inverse semantic segmentation" on the Cityscapes dataset is impossible, while similar evaluation can be found in pix2pix [2] and CycleGAN [5].

3 Unsupervised I2I translation

3.1 CycleGAN

3.1.1 Summary

In this paper, CycleGAN was proposed by Zhu et al. [5] for I2I translation in the absence of any paired examples. The key idea is based on a cycle consistency, which means if we translate a dog image to a cat image and translate it back, we should reconstruct the original dog image. As shown in Figure 3 (a), their model includes two mappings $G : X \rightarrow Y$ and $F : Y \rightarrow X$ associated with two adversarial discriminators D_X and D_Y. Given an input image x, as shown in Figure 3 (b), they first apply mapping G to translate it into domain Y, then they apply inverse mapping F to reconstruct the input image x and simultaneously minimize the reconstruction error. They do the same thing in the opposite direction as shown in Figure 3 (c). Both qualitative and quantitative tests are conducted to prove the superiority of the proposed method.



(a) (b) (c)

Figure 3: Overview of CycleGAN. [5]

3.1.2 Strength

The idea of cycle consistency behind the method proposed in this paper is novel and simple to understand. The introduction and related work are written to perfection, well explaining the motivation and intuition for combining cycle consistency with GAN. The method is evaluated both qualitatively and quantitatively, also with an ablation study, and convincing results are presented. The paper pushed the boundaries of unsupervised I2I translation and served as a step stone for many subsequent works.

3.1.3 Weakness

Similar to the weakness of BicycleGAN, it would be more interesting to see that the Amazon Mechanical Turk (AMT) "real vs fake" study is also conducted on other tasks besides maps ↔ satellites.

3.2 UNIT

3.2.1 Summary

Another successful model for unsupervised I2I translation was proposed by Liu et al. [6], which is called UNIT, the short name of unsupervised I2I translation. They propose a different assumption which is called shared latent space. As shown in Figure 4, they assume that a latent space Z can be shared by two different domains χ_1 and χ_2. Each domain has an encoding function, i.e. E_1 and E_2 to map images to latent codes. G_1 and G_2 are two generation functions, mapping latent codes to images. To translate an image from one domain to another domain, the input image is simply encoded using the encoder of the source domain and then decoded using the decoder of the target domain to achieve a cycle consistency. The method is evaluated on various challenging unsupervised I2I translation tasks and domain adaptation.

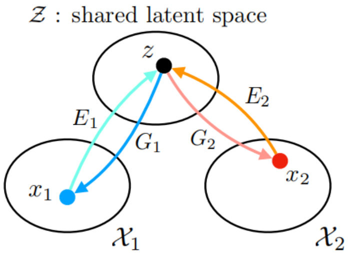


Figure 4: Assumption of shared latent space. [6]

3.2.2 Strength

This paper investigated the problem of unsupervised I2I translation from a probabilistic modeling perspective. Results on several image datasets demonstrate good performance of the proposed UNIT method. Overall, the paper is clearly written with sufficient details including both qualitative and quantitative analysis. The authors also provided an interpretation of the roles of the subnetworks in the proposed framework, which is helpful to understand the whole architecture.

3.2.3 Weakness

The authors performed an ablation study measuring impact of the weight-sharing and cycle-consistency constraints to the translation performance only on satellite images → to maps. It would be a nice extension to see these ablation studies also carried out on other tasks, such as SVHN → MNIST.

3.3 EGSC-IT

3.3.1 Summary

Although CycleGAN and UNIT works fine on unsupervised I2I translation, they are one to one mapping. Ma et al. [7] proposed the exemplar guided & semantically consistent I2I translation (EGSC-IT) network which achieves multimodal translation. They assume that an image is composed of two representations, a domain-shared representation that models the content in the image, like objects' category, shape and spatial layout, and a domain-specific representation that contains the style information, like color and texture. To learn the content component of an image pair that is shared across source and target domains they employ the weight sharing strategy proposed in UNIT [6]. To learn the style component, they apply adaptive instance normalization to the shared content component before the decoding stage. In particular, the exemplar from the target domain is fed to another network to compute a set of feature maps which are expected to contain the style information. Finally, to tackle the unsupervision problem, they compute feature masks which can be used to retain the semantic consistency during translation. The method is evaluated on three tasks: 1) single-digit translation; 2) multi-digit translation; 3) street view translation.

3.3.2 Strength

First, the paper is well organized and easy to follow. Second, the paper is based on a clear assumption that an image comprises of a content component and a style component, which is used by others as well. The main novelty of this paper is to propose feature masks to retain semantic consistency during the translation process without using any semantic labels. Besides, the introduction and related work clearly identify the research gap and explain the novelty of the proposed method. Overall, this paper is solid with interesting ideas and good implementation.

3.3.3 Weakness

First, the discussion and analysis of experiments are biased. The experiment of single-digit translation, the simplest task among the three tasks, is overly descriptive. It would be more convincing if the controlled experiment is also conducted on street view translation. Moreover, it would be more helpful for future researches if the authors provide some comments on the limitation of their approach and recommendations in the conclusion.

4 I2I translation in Geomatics

4.1 Aerial photos $\leftrightarrow$ maps

4.1.1 Summary

As discussed before, some models have been evaluated on the task of aerial photos $\leftrightarrow$ maps translation, such as pix2pix [2], CycleGAN [5], BicycleGAN [3] and UNIT [6]. Due to hardware limitation and large detection distance, scene distribution of aerial photos is more complicated than ordinary image. Gu et al. [8] explored the bidirectional translation of aerial photos and maps using a GAN-based network. They decompose the generator into two sub-networks, a global generator and a local generator. The global generator is proposed to generate the basic structure of the images. The local generator outputs an image with a high resolution. Correspondingly, they use multi-scale discriminators, that is, three discriminators to distinguish real and synthesized images at the three different scales, respectively.

4.1.2 Strength

In general, this paper is well organized and easy to follow. Research gaps in aerial photos $\leftrightarrow$ maps translation are well identified in the introduction. The multi-scale generators and discriminators proposed in this paper are proven effective to deal with the aerial photos $\leftrightarrow$ maps translation task.

4.1.3 Weakness

The authors did not provide reasons for the hyperparameters setting in the implementation details. Besides, the authors claimed that their network can also be used in other image style translation tasks. However, no experiments are conducted in other tasks in the paper. The authors should provide more details of the performance of their model to make this statement more convincing.

4.2 SAR images $\rightarrow$ optical images

4.2.1 Summary

Interpreting details in SAR images is a challenging task, even for experts. In order to facilitate the interpretation of SAR images, some researchers have investigated the potential of generative deep learning models in the context of SAR image interpretation. Reyes et al. [9] proposed an adapted version of CycleGAN for the SAR-to-optical image translation. They conducted several steps of optimization starting with basic CycleGAN to improve the results of SAR image translation. Some parameters (e.g. learning rate, number of epochs and size of the patches) are tuned to outperform the default conditions. Several residual layers are added to the original CycleGAN architecture to maintain the level of detail of the generated images. Besides, to reduce the creation of fictional objects in the generated images, a logarithmic scaling of SAR image intensity is applied and the intensity of the patches belonging to a same image is normalized.

4.2.2 Strength

This paper is well written with sufficient details. Although the paper did not propose a novel algorithm, it comprehensively analyzed the application opportunities of GAN in SAR-to-optical image translation and proposed effective optimization methods, which are beneficial to the field of remote sensing. The introduction and related work are clear and comprehensive, well explaining the opportunity and challenge in GAN-based SAR-to-optical image translation. The results are well evaluated based on feedback from experts in SAR remote sensing and the impact on road extraction as an example for follow-up applications.

4.2.3 Weakness

The authors proposed several optimization steps for CycleGAN, but no ablation study is presented in the paper. It would be more interesting to see how these optimization steps affect the model performance individually.

5 Summary

5.1 Comparison

The papers reviewed in this report are listed in Table 1. As can be seen, some early works [2, 3, 4] are in a supervised way, which replies on paired training data. Because acquiring paired data is expensive or sometimes impossible, more and more unsupervised methods [5, 6, 7] appeared, some of which can achieve comparable performance to the supervised approaches. These models can also be categorized into unimodal algorithms and multimodal algorithms. Unsupervised multimodal I2I translation maybe the mainstream in the future because it does not require paired training data and can produce diverse results. Another interesting thing to note is that only the method proposed by Chen et al. [4] is independent of GAN. The most prominent contemporary approach to I2I translation is based on GAN because of its inherent advantages in image synthesis. Finally, two papers focused on I2I translation tasks in geomatics are reviewed. These works usually adopt the model proposed by the computer community and propose optimization methods to make the model better applicable for specific tasks in the field of geomatics.

Table 1: Comparison of reviewed papers

Author(s), Year	Model abbreviation	Unsupervised	Multimodal	Based on GAN	Focus
Isola et al. (2017) [2]	pix2pix	×	×	✓	general translation
Zhu et al. (2017) [3]	BicycleGAN	×	✓	✓	general translation
Chen et al. (2017) [4]	-	×	✓	×	label map → photo
Zhu et al. (2017) [5]	CycleGAN	✓	×	✓	general translation
Liu et al. (2017) [6]	UNIT	✓	×	✓	general translation
Ma et al. (2018) [7]	EGSC-IT	✓	✓	✓	general translation
Gu et al. (2019) [8]	-	✓	×	✓	aerial photo ↔ map
Reyes et al. (2019) [9]	-	✓	×	✓	SAR image → optical image

5.2 Open challenges

Although many proposed methods have attempted to overcome the limitations of I2I translation, there are still some open challenges that have not been fully addressed.

Mode collapse I2I translation with GANs and GAN variants usually suffers from the mode collapse, which means the generator provides limited sample variety. This is caused by the inherent structure of GAN. The generator is always trying to find the output that seems most plausible to the discriminator.

Non-trivial training Another challenge for GAN based methods is non-trvial training, which means that it is not yet clear whether there is a point to which the training converges or not.

Evaluation metrics Most papers reviewed in this report evaluate their methods both qualitatively and quantitatively. However, almost all qualitative evaluation are based on human judgment, which is a subjective metric. As I2I translation is essentially a one-to-many mapping problem, there is no strict one-to-one correspondence between the input image and the generated image, which increases the difficulty of evaluation.

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[2] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[3] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in neural information processing systems*, pages 465–476, 2017.

[4] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1511–1520, 2017.

[5] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[6] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30:700–708, 2017.

[7] Liqian Ma, Xu Jia, Stamatios Georgoulis, Tinne Tuytelaars, and Luc Van Gool. Exemplar guided unsupervised image-to-image translation with semantic consistency. *Proceedings of the international conference for learning representations*, 2019.

[8] Jun Gu, Yue Zhang, Wenkai Zhang, Hongfeng Yu, Siyue Wang, Yaoling Wang, and Lei Wang. Aerial image and map synthesis using generative adversarial networks. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 9803–9806. IEEE, 2019.

[9] Mario Fuentes Reyes, Stefan Auer, Nina Merkle, Corentin Henry, and Michael Schmitt. Sar-to-optical image translation based on conditional generative adversarial networks—optimization, opportunities and limits. *Remote Sensing*, 11(17):2067, 2019.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[11] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European conference on computer vision*, pages 702–716. Springer, 2016.

[12] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

Appendix

Visited Scientific Presentations

Table 2: Visited scientific presentations

Time	Title	Speaker
22/10/2020 17:00	Measuring the Earth with Quasars	Prof. J. Böhm, TU Wien
29/10/2020 17:00	Enhancing Knowledge, Skills, and Spatial Reasoning through Location-based Mobile Learning	Dr. Ch. Sailer, Esri Schweiz
01/12/2020 03:00	CREATE Mobility Symposium: Singapore-MIT Alliance for Research and Technology Centre	Prof. J. Zhao, MIT

Checklist for reviewers

No. Manuscript: Ms XXXXXX
Date received for review: 01-10-2014
Responsible editor: X. Xxxxxxxxx
Title: Exemplar guided unsupervised image-to-image translation with semantic consistency
Author(s): Liqian Ma, Xu Jia, Stamatios Georgoulis, Tinne Tuytelaars, Luc Van Gool

Reviewer: Yuanwen Yue

The checklist below is supplied to aid in the critical evaluation of manuscripts submitted for publication.

1. Does the manuscript contain enough new material to warrant publication in the Journal of XXXXXXX? ☑ Yes □ No

2. Are the assumptions clearly and explicitly stated? ☑ Yes □ No

3. Are the conclusions, claims, etc., clear? ☑ Yes □ No

4. Does the author give proper credit to related work? ☑ Yes □ No

5. Is the English language satisfactory? ☑ Yes □ No

6. Is the length of the manuscript satisfactory? ☑ Yes □ No

7. Are any parts of the manuscript (text, tables, figures, mathematical operations) too short or unnecessarily long? If yes please state which. ☑ Yes □ No

<u>Experiments are not balanced</u>

8. Is the abstract of the manuscript self-contained without being too long? ☑ Yes □ No

9. Are new results emphasised for maximum effectiveness in abstract journals? ☑ Yes □ No

10. Are the figures clearly drawn? ☑ Yes □ No

11. Are computer programs commented and brushed up for publication? □ Yes ☒ No

12. Do you recommend outright rejection? □ Yes ☒ No

13. Do you recommend publication of this manuscript in its present form? □ Yes ☒ No

14. Does reviewer consider it useful to see the revised version before publication? ☑ Yes □ No

15. Do you give permission for your name to be known as the reviewer? ☑ Yes □ No

State, on ½ - ¾ page, the most important reasons for your "yes" or "no" answers to the above items.

All comments are based on this version of the paper: https://arxiv.org/pdf/1805.11145v3.pdf

Reasons for items 1-6 and 8-12:

- The paper is based on a clear assumption that an image comprises of a content component shared across domains, and a style component specific to each domain, which is used by others as well.
- The main novelty of this paper is to propose feature masks to retain semantic consistency during the translation process without using any semantic labels.
- The abstract provides a concise overview of the research and clearly defines the research problem of unsupervised multimodal image to image translation.
- The introduction and related work clearly identify the research gap and explain the novelty of the proposed method.
- Overall, the paper is well organized and easy to follow.

Reasons for items 7, 13-15:

- The discussion and analysis of experiments are not balanced. A large number of paragraphs are used to describe the experiment of single-digit translation, the simplest task among the three tasks. It would be more convincing if more details about the performance on complex tasks are given.
- There is no explanation for the absence of controlled experiment on the task of street view translation. It would have been useful to know why this was the case.
- It might have been helpful to explain more details about the figure of t-SNE embeddings visualization to support “Our method can match the distributions well”.
- There is no description of limitations and recommendations in the conclusion, which may be useful for future researches.

Based on above comments, I would recommend publication of this manuscript in a revised version.