

Unsupervised Image-to-Image Translation with Self-Attention Networks

Taewon Kang
Sejong Academy of Science and Arts
Sejong, Korea
itschool@itsc.kr

Kwang Hee Lee*
Boeing Korea Engineering and
Technology Center (BKETC)
Seoul, Korea
kwanghee.lee2@boeing.com
* indicates corresponding author

Abstract—Unsupervised image translation aims to learn the transformation from a source domain to a target domain given unpaired training data. Several state-of-the-art works have yielded impressive results in the GANs-based unsupervised image-to-image translation. It fails to capture strong geometric changes between domains, or it produces unsatisfactory results for complex scenes, compared to local texture mapping tasks such as style transfer. Recently, SAGAN [35] showed that the self-attention network produces better results than the convolution-based GAN. However, the effectiveness of the self-attention network in unsupervised image-to-image translation tasks have not been verified. In this paper, we propose an unsupervised image-to-image translation with self-attention networks, in which long range dependency helps to not only capture strong geometric change but also generate details using cues from all feature locations. In experiments, we qualitatively and quantitatively show superiority of the proposed method compared to existing state-of-the-art unsupervised image-to-image translation task. The source code and our results are online: https://github.com/itsss/img2img_sa and http://itsc.kr/2019/01/24/2019_img2img_sa

Keywords—Generative Adversarial Networks, Image-to-Image Translation, Self-Attention Networks

I. INTRODUCTION

In computer vision and graphics there are many image-to-image translation tasks, including inpainting [17], [26], super resolution [10], [19], colorization [36], [37], style transfer [11], [15], [25] and so on. This cross-domain image-to-image translation topic has become a major concern of researchers.

In many cases, given a paired dataset, it is possible to solve the problem with conditional image translation [18], [22], [30]. However, it is difficult and expensive to obtain the paired samples. In addition, there are cases where supervision is not possible.

The goal of the unsupervised image translation is to learn the transformation from a source domain to a target domain given unpaired training data. Recent works have yielded impressive results in the GANs-based unsupervised image-to-image translation [1], [8], [16], [20], [23], [27], [29], [34], [38]. It can be largely classified into two types. The first is the style transfer task. This problem is to change low-level information such as color or texture while maintaining high-level information such as content or geometric structure.

Style transfer and conditional GANs-based methods have yielded excellent results in this research area.

The second is the object transfiguration task. Unlike the style transfer task, this focuses on changing high-level information while keeping the low-level information. CycleGAN [38], the most representative unsupervised image translation method, failed to change the high-level semantic meaning due to the network structure specialized for style transfer.

To solve the unsupervised image-to-image translation problem, UNIT [23] made a shared-latent space assumption. It assumes a pair of corresponding images in different domains can be mapped to a same latent code in a shared-latent space. MUNIT [16] proposed a multimodal unsupervised image-to-image translation framework.

To achieve many-to-many cross domain mapping, it mitigates a fully shared latent space assumption in UNIT by decomposing a shared-latent space across domains and each domain-specific part for the style code. UNIT and MUNIT experimentally showed impressive animal image translation from a cropped dataset centered on the head. When the training image dataset is spatially unnormalized, it makes the problem more difficult because the absence of correspondences between the shared semantic parts.

In our experiments, we show that these methods often fail in various image-to-image translation applications with strong geometric change. Recently, SAGAN [35] showed that the self-attention module is complementary to convolutions and helps with modeling long range, multi-level dependencies across image regions. Despite the success of the self-attention module in non-conditional GANs, the effectiveness of the self-attention module for unsupervised image-to-image translation has not been validated.

In this paper, we propose a unpaired image-to-image translation model with self-attention networks which allows long range dependency modeling for image translation task with strong geometry change. In experiments, we show superiority of the proposed method compared to existing state-of-the-art unsupervised image-to-image translation tasks.

The source code and our results are online: https://github.com/itsss/img2img_sa and http://itsc.kr/2019/01/24/2019_img2img_sa.

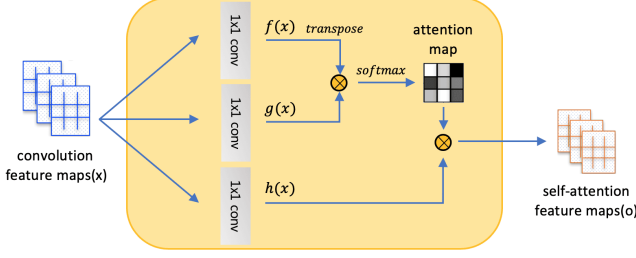


Figure 1. Self Attention Networks. [35] \otimes means Matrix multiplication.

II. SELF-ATTENTION GANS

SAGAN [35] showed that the self-attention module is complementary to convolutions and it helps with long range modeling, multi-level dependencies across image regions. Attention mechanisms have become a important part of models that must capture global dependencies [2], [7], [13], [24], [32], [33].

Self attention networks adapt a non-local block [31] to introduce the self-attention to the GAN networks, can enable both the generator and discriminator to efficiently model relationships between widely separated spatial regions. The non-local mechanisms also have become a important part of image generation [3]–[6], [9], [12].

In the self attention module (Figure 1.), image features from the previous hidden layer x are firstly transformed into two feature spaces f and g to calculate the attention.

$$\beta_{j,i} = \frac{\exp(s_{ij})}{\sum_{i=1}^N \exp(s_{ij})}, \text{ where } s_{ij} = f(x_i)^T g(x_j),$$

where $f(x) = W_f x$, $g(x) = W_g x$ and $\beta_{j,i}$ indicates the extent to which the model attends to the i^{th} location when synthesizing the j^{th} region. Then the output of the attention layer is $o = (o_1, o_2, \dots, o_j, \dots, o_N)$, where,

$$o_j = \sum_{i=1}^N \beta_{j,i} h(x_i), \quad h(x_i) = W_h x_i$$

In the above formulation, W_g , W_f and W_h are the learned weights parameters, which are implemented as 1×1 convolutions.

III. METHODS

A. Unpaired Image-to-Image Translation with Self Attention Networks

We propose an unsupervised image-to-image translation model with self-attention networks that allows long range dependency modeling for image translation tasks with strong geometry change. Combined with self-attention, the generator can translate images in which fine details at every position are carefully coordinated with fine details in distant portions of the image. Furthermore, the discriminator can

also more accurately enforce complicated geometric constraints on the global image structure.

In this paper, our network architecture is devised by combining several self-attention blocks into the generator and discriminator of the Multimodal Unsupervised Image-to-Image Translation [16](MUNIT) model.

To explore the effect of the proposed self-attention mechanism, we built several SAGAN blocks by adding the self-attention mechanism to different stages of the generator and discriminator. For the generator, the self-attention layers are placed before the downsampling layer in the encoder and before the upsampling layer in the decoder, respectively. For the discriminator, it is added before the downsampling layer. Figure 2. shows architecture of our autoencoder model with self-attention networks.

B. Loss Function

The full objective of our model comprises a bidirectional reconstruction loss function and an adversarial loss function. Same as in [16], our model consists of an encoder E_i and a decoder G_i for each domain. The latent code of each autoencoder is divided into a content code c_i and a style code s_i , where $(c_i, s_i) = (E_i^c(x_i), E_i^s(x_i)) = E_i(x_i)$. Image-to-image translation can be performed by exchanging encoder-decoder pairs.

Bidirectional Reconstruction Loss Bidirectional reconstruction loss includes image reconstruction loss and latent reconstruction loss. The image reconstruction loss formula is as follows:

$$\mathcal{L}_{recon}^{x_1} = \mathbb{E}_{x_1 \sim p(x_1)} [\|G_1(E_1^c(x_1), E_1^s(x_1)) - x_1\|_1].$$

We should be able to reconstruct an image sampled from the data distribution after encoding and decoding.

The latent reconstruction loss formula is as follows:

$$\begin{aligned} \mathcal{L}_{recon}^{c_1} &= \mathbb{E}_{c_1 \sim p(c_1), s_2 \sim q(s_2)} [\|E_2^c(G_2(c_1, s_2)) - c_1\|_1] \\ \mathcal{L}_{recon}^{s_2} &= \mathbb{E}_{c_1 \sim p(c_1), s_2 \sim q(s_2)} [\|E_2^s(G_2(c_1, s_2)) - s_2\|_1] \end{aligned}$$

Given a latent code (content and style) from the latent distribution, we should be able to reconstruct it after decoding and encoding.

Adversarial Loss The adversarial loss formula is as follows:

$$\begin{aligned} \mathcal{L}_{GAN}^{x_2} &= \mathbb{E}_{c_1 \sim p(c_1), s_2 \sim q(s_2)} [\log(1 - D_2(G_2(c_1, s_2)))] \\ &\quad + \mathbb{E}_{x_2 \sim p(x_2)} [\log D_2(x_2)]. \end{aligned}$$

To match the distribution between the translated and target domain, we employ the adversarial loss.

Full objective The total loss formula is as follows:

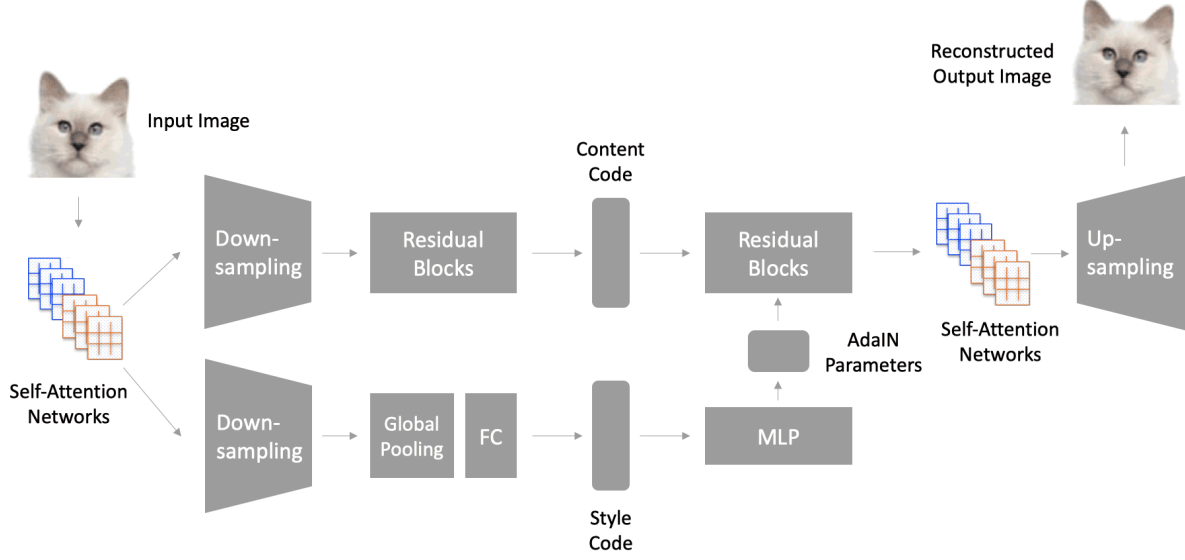


Figure 2. Architecture of our Network Autoencoder Model

$$\min_{E_1, E_2, G_1, G_2} \max_{D_1, D_2} \mathcal{L}(E_1, E_2, G_1, G_2, D_1, D_2) = \mathcal{L}_{GAN}^{x_1} + \mathcal{L}_{GAN}^{x_2} + \lambda_x(\mathcal{L}_{recon}^{x_1} + \mathcal{L}_{recon}^{x_2}) + \lambda_c(\mathcal{L}_{recon}^{c_1} + \mathcal{L}_{recon}^{c_2}) + \lambda_s(\mathcal{L}_{recon}^{s_1} + \mathcal{L}_{recon}^{s_2}).$$

IV. EXPERIMENTAL RESULTS

In this section, we compared the performance of our model against various unsupervised image-to-image translation models (CycleGAN [38], DRIT [21], UNIT [23], MUNIT [16]). In order to evaluate visual quality of translated images, we performed a user study.

A. Implementation Details

We used the MUNIT default setting for experiments. We used the Adam optimizer with $\beta_1 = 0.05$, $\beta_2 = 0.999$. Initial learning rate of 0.0001 and the learning rate is decreased by half every 100,000 iterations. We used a batch size of 1 and set the loss weights to $\lambda_x = 10$, $\lambda_c = 1$, $\lambda_s = 1$. We trained our networks on four TITAN X accelerators. We trained it over 1,000,000 epochs for around 5 days.

B. Datasets

We used cat2dog, face2dog, face2cat, portrait and edges2shoes for test our network.

cat2dog: This datasets are used in DRIT [21]. This dataset contains cat(871) and dog(1,364).

face2dog: This dataset contains faces (CelebA dataset, 202,599) and dog(cat2dog dataset, 1,364).



Figure 3. Examples of Unsupervised image translation from cat(cat2dog Dataset, Domain A) to dog(cat2dog Dataset, Domain B) using various network structures. CycleGAN, DRIT, UNIT, MUNIT are all trained to 64×64 resolution using the default settings from the official implementations.

face2cat: This dataset contains faces (CelebA dataset, 202,599) and cat(cat2dog dataset, 871).

portrait: This datasets are used in DRIT [21]. This dataset contains portrait(1,814) and face photo(6,452).

edges2shoes: This dataset are used in MUNIT [16]. This dataset contains edges(50,025) and shoes(50,025).

C. Comparision with Previous Works

cat2dog In the process of changing the image of a cat(domain A) to a dog(domain B) image(Figure 3.), Cy-

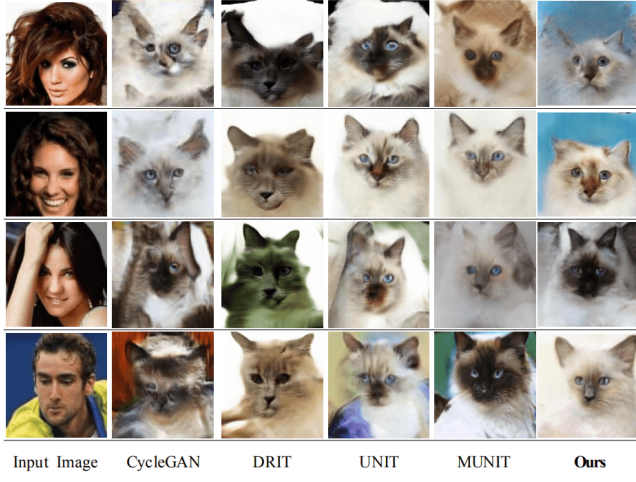


Figure 4. Examples of Unsupervised image translation from human face(CelebA Dataset, Domain A) to cat(cat2dog Dataset, Domain B) using various network structures. CycleGAN, DRIT, UNIT, MUNIT are all trained to 64×64 resolution using the default settings from the official implementations.

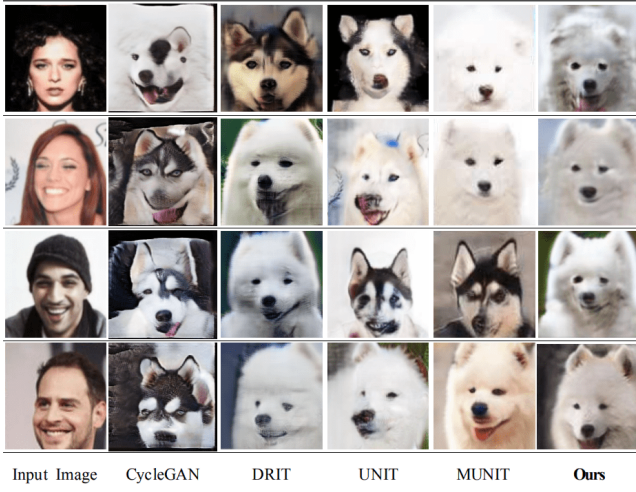


Figure 5. Examples of Unsupervised image translation from human face(CelebA Dataset, Domain A) to dog(cat2dog Dataset, Domain B) using various network structures. CycleGAN, DRIT, UNIT, MUNIT are all trained to 64×64 resolution using the default settings from the official implementations.

cleGAN is unable to generate a dog image, since it only takes the color from the image. In the case of DRIT, there is a problem that the image is broken, and it is hard to see it as a dog image reflecting the shape and direction of a cat and dog.

face2cat and face2dog In the process of translating the human face image(domain A) to cat and dog(domain B), CycleGAN and DRIT could not obtain the desired results. Most translated results were distorted. In the case of UNIT

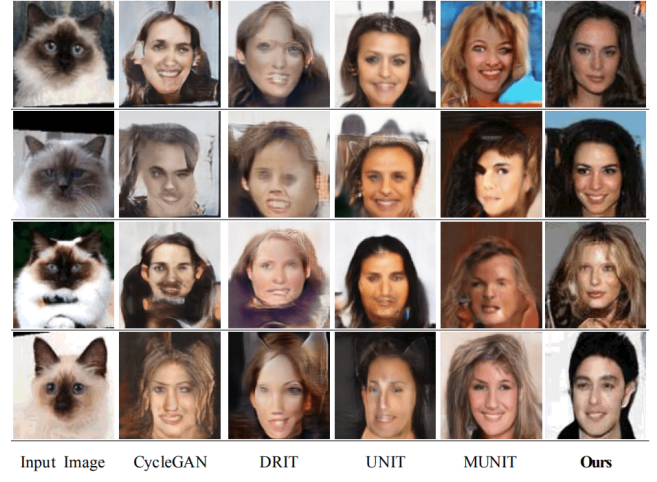


Figure 6. Reverse of Figure 4. Examples of Unsupervised image translation from cat(cat2dog Dataset, Domain A) to human(CelebA Dataset, Domain B) using various network structures. CycleGAN, DRIT, UNIT, MUNIT are all trained to 64×64 resolution using the default settings from the official implementations.

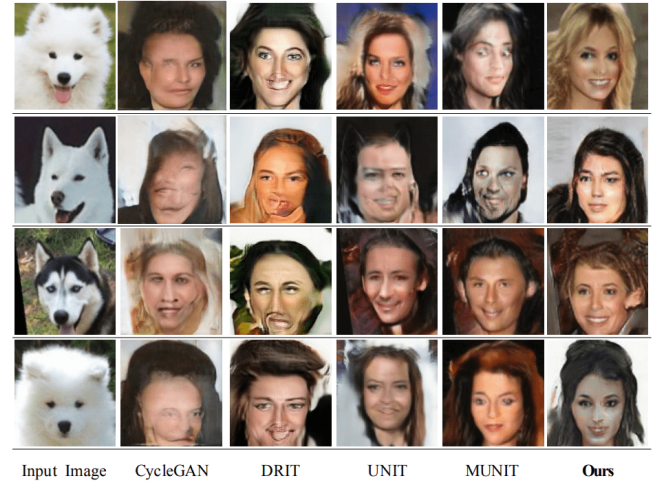


Figure 7. Reverse of Figure 5. Examples of Unsupervised image translation from dog(cat2dog Dataset, Domain A) to human(CelebA Dataset, Domain B) using various network structures. CycleGAN, DRIT, UNIT, MUNIT are all trained to 64×64 resolution using the default settings from the official implementations.

and MUNIT, there was a tendency to leave the shape of human face or be distorted in the translated image (See Figure 4. and Figure 5.).

cat2face and dog2face we also experimented for the translations from cats and dogs to human faces. In this experiments, the proposed method showed much better results than the previous works (See Figure 6. and Figure 7.).

portrait Even at the stage of changing the portrait (domain A) shown in Figure 8. to a face(domain B), CycleGAN has not been able to convert portrait photos to face at all. In the

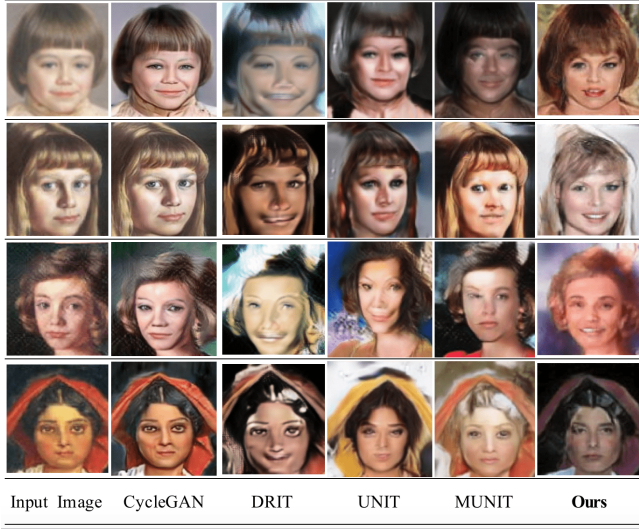


Figure 8. Examples of Unsupervised image translation from portrait (portrait Dataset, Domain A) to human face (portrait Dataset, Domain B) using various network structures. CycleGAN, DRIT, UNIT, MUNIT are all trained to 64×64 resolution using the default settings from the official implementations.



Figure 9. Examples of Unsupervised image translation from edges (edges2shoes Dataset, Domain A) to shoes (edges2shoes Dataset, Domain B) using various network structures. CycleGAN, DRIT, UNIT, MUNIT are all trained to 64×64 resolution using the default settings from the official implementations.

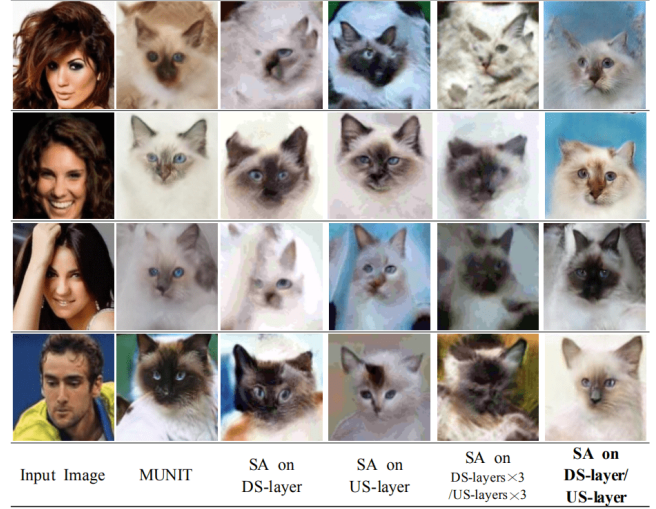


Figure 10. Ablation-study result of four self attention techniques.

case of DRIT, conversion is not performed by generating irrelevant images. In the case of UNIT and MUNIT, there is a problem that the image is distorted although it reflects the shape.

edges2shoes In the process of translating from the edges image (domain A) to a shoes (domain B) image, our model generated more realistic results keeping the pose and style of A domain than the results from other models (See Figure 9).

D. Ablation Study

In this section, the other experiments are conducted to evaluate the effectiveness of the self-attention (SA) networks in our unsupervised image-to-image translation model. In Figure 10, self attention unsupervised image-to-image translation models "SA on downsampling layer (DS-layer)", "SA on upsampling layer (US-layer)" and "SA on DS-layers $\times 3$ / US-layers $\times 3$ " are compared with our "SA on DS-layer / US-layer" model.

In case of "SA on DS-layer", "SA on US-layer" and "SA on DS-layers $\times 3$ / US-layers $\times 3$ ", we could not obtain the well-translated results. However, "SA on DS-layer / US-layer" model generated more realistic images than other methods. Based on this experiment, we applied "SA on DS-layer / US-layer" to our model.

E. User Study

For the qualitative evaluation, we also conducted a user study on 80 participants. The results of this study are summarized as follows. First, 192 images were selected randomly in the questionnaires, and the questionnaires were used to select the best image that reflects the pose of input Image and the appearance of target domain well. Figure 11 shows that our method yields quantitatively much more superior results than the existing GAN models.

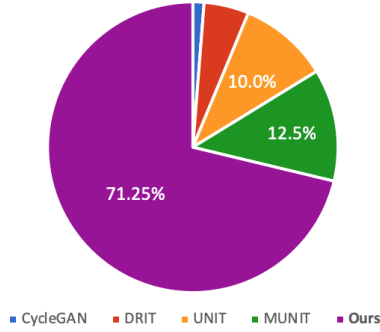


Figure 11. User-study result of five image-to-image translation algorithms.

Dataset	CycleGAN	DRIT	UNIT	MUNIT	Ours
cat2dog	133.21	148.87	101.41	122.04	96.34
face2cat	274.61	117.05	85.78	104.09	79.95
face2dog	279.74	108.29	82.12	133.96	90.70
cat2face	454.99	242.78	359.62	269.44	208.33
dog2face	366.33	229.21	229.06	228.06	217.58
portrait	233.34	282.29	263.28	269.56	256.04
edges2shoes	269.18	273.93	250.99	274.11	238.57

Table I
QUANTITATIVE EVALUATION ON 7 IMAGE TRANSLATION EXAMPLES.
WE USED FRECHET INCEPTION DISTANCE(FID) TO MEASURE THE
PERFORMANCE OF VARIOUS NETWORK STRUCTURES.

F. Quantitative Evaluation Analysis

We used Frchet Inception Distance (FID) [14] to measure the distance between the data distributions of the source and target domains using the features extracted by the inception networks [28]. The lower FID score indicates that the data distribution of two domains are similar. Table I. shows the results of the FID score analysis, and we can see that our model translated more similar images than other image-to-image translation methods.

V. CONCLUSIONS

In this paper, we proposed a method about unsupervised image-to-image translation with self-attention networks, in which long range dependency helps to not only capture strong geometric change but also generate details using cues from all feature locations. In experiments, we showed superiority of the proposed method compared to existing state-of-the-art unsupervised image-to-image translation methods.

A. Acknowledgements

We would like to thank our sponsors, especially, Sejong Academy of Science and Arts(SASA) in Korea. We also thanks to Artificial Intelligence Research Institute(AIRI) in Korea. This research was funded by the Korea Foundation for the Advancement of Science & Creativity(KOFAC) Science High School Student R&E support program.

REFERENCES

- [1] Asha Anoosheh, Eirikur Agustsson, Radu Timofte, and Luc Van Gool. Combogan: Unrestrained scalability for image domain translation. *arXiv preprint arXiv:1712.06909*, 2017.
- [2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [3] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. In *ACM Transactions on Graphics (ToG)*, volume 28, page 24. ACM, 2009.
- [4] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 60–65. IEEE, 2005.
- [5] Harold Christopher Burger, Christian J Schuler, and Stefan Harmeling. Image denoising with multi-layer perceptrons, part 1: comparison with existing algorithms and with bounds. *arXiv preprint arXiv:1211.1544*, 2012.
- [6] Harold Christopher Burger, Christian J Schuler, and Stefan Harmeling. Image denoising with multi-layer perceptrons, part 2: training trade-offs and analysis of their mechanisms. *arXiv preprint arXiv:1211.1552*, 2012.
- [7] Jianpeng Cheng, Li Dong, and Mirella Lapata. Long short-term memory-networks for machine reading. *arXiv preprint arXiv:1601.06733*, 2016.
- [8] Yunje Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *arXiv preprint arXiv:1711.09020*, 2017.
- [9] K Dabov, A Foi, V Katkovnik, and K Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *image processing, ieee transactions on* 16 (8), pp. 2080-2095. 2007.
- [10] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2016.
- [11] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 2414–2423. IEEE, 2016.
- [12] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *2009 IEEE 12th International Conference on Computer Vision (ICCV)*, pages 349–356. IEEE, 2009.
- [13] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015.

- [14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.
- [15] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. *CoRR*, abs/1703.06868, 2017.
- [16] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. *arXiv preprint arXiv:1804.04732*, 2018.
- [17] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (TOG)*, 36(4):107, 2017.
- [18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.
- [19] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1646–1654, 2016.
- [20] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jungkwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017.
- [21] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–51, 2018.
- [22] Chunyuan Li, Hao Liu, Changyou Chen, Yuchen Pu, Liqun Chen, Ricardo Henao, and Lawrence Carin. Alice: Towards understanding adversarial learning for joint distribution matching. In *Advances in Neural Information Processing Systems*, pages 5501–5509, 2017.
- [23] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pages 700–708, 2017.
- [24] Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*, 2016.
- [25] Dae Young Park and Kwang Hee Lee. Arbitrary style transfer with style-attentional networks. *arXiv preprint arXiv:1812.02342*, 2018.
- [26] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016.
- [27] Amélie Royer, Konstantinos Bousmalis, Stephan Gouws, Fred Bertsch, Inbar Moressi, Forrester Cole, and Kevin Murphy. Xgan: Unsupervised image-to-image translation for many-to-many mappings. *arXiv preprint arXiv:1711.05139*, 2017.
- [28] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
- [29] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016.
- [30] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. *arXiv preprint arXiv:1711.11585*, 2017.
- [31] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.
- [32] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.
- [33] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016.
- [34] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. *arXiv preprint*, 2017.
- [35] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.
- [36] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European Conference on Computer Vision*, pages 649–666. Springer, 2016.
- [37] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S Lin, Tianhe Yu, and Alexei A Efros. Real-time user-guided image colorization with learned deep priors. *arXiv preprint arXiv:1705.02999*, 2017.
- [38] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017.