UNMMIT: A Unified Framework on Unsupervised Multimodal Multi-domain Image-to-Image Translation

Lei Luo*, Shangxian Wang, and William H. Hsu

***Abstract*—We address the open problem of unsupervised multimodal multi-domain image-to-image (I2I) translation using a generative adversarial network. Previous works, such as MUNIT and DRIT, are able to translate images among multiple domains, but they generate images of inferior quality and less diverse. Moreover, they require training $n(n-1)$ generators and n discriminators for learning to translate images among n domains, which is computationally expensive. In this paper, we propose a simpler yet more effective framework for unsupervised multimodal multi-domain I2I translation. Our approach only consists of a mapping network, a encode-decoder pair (generator), and a discriminator. Our method assumes that the latent space can be decomposed into content and style sub-spaces by the encoder, where content space is deemed domain-invariant and style space is domain-dependent. Unlike MUNIT and DRIT that simply sample style codes from a standard normal distribution when translating, we employ a mapping network to learn the style of different domains, which yields better translation results. Translation is done through the decoder by keeping content codes and exchanging the style codes. To encourage diversity in translated images, we employ style regularizations and inject Gaussian noise into the decoder. Extensive experiments show that our framework is superior or comparable to state-of-the-art baselines.**

***Index Terms*—Unsupervised multimodal multi-domain image-to-image translation, style codes, content codes, mapping network**

I. Introduction

Image-to-image (I2I) translation refers to translating images from one domain to another featuring different styles, which are visually distinctive among different domains. An example is the task of turning images of cartoon sketches into real-life graphs. Many tasks in computer vision can be viewed as I2I translation, such as image translation (MUNIT [1], AMMUNIT [2]), image inpainting [3], style transfer (StyleGANs [4], DRIT [5]), and super-resolution [6]. Supervised I2I translation tasks need paired data sets that are costly to obtain, and such tasks are relatively easier to solve than their unsupervised counterpart. Under paired data supervision, I2I translation can be done by taking a regression approach [7] or using conditional generative models [8]. Our work addresses the more challenging unsupervised I2I translation task without access to paired data sets. Most of works on unsupervised I2I translation draw inspiration from CycleGANs [9] using the cycle consistency constraint and have achieved impressive results. More recent studies have improved upon on CycleGANs and are able to translate images among multiple domains. They, such as MUNIT often assumes latent codes can be decomposed as content codes and style codes. Translation is done by exchanging style codes with different domains while keeping the original content codes. Style codes, however, are simply sampled from a standard normal distribution, which leads to inferior translation results. Moreover, these works require training $n(n-1)$ generators and n discriminators for learning to translate images among n domains, which is costly. In our study we propose a simpler yet more effective approach. Our framework shares the same assumption with style codes being domain-dependent and content codes being domain-invariant. However, our approach only consists of one generator-discriminator pair and a mapping network, which learns the style codes of different domains. We also employ several effective techniques for encouraging translated results being more diverse. Extensive experiments show that our framework is superior or comparable to state-of-the-art (SOTA) baselines. The contrition of our work can be summarized as follows:

- We propose a new unified framework for unsupervised multimodal multi-domain I2I translation that largely simplifies the architecture of existing works and improves the translation performance by a large margin.
- Instead of sampling from a standard normal distribution, we learn the style of domains by employing a mapping network, which yields better translation results.
- We propose two new regularization techniques for learning content and style information of domains.
- Extensive experiments show that our framework is superior or comparable to state-of-the-art (SOTA) baselines.

II. Related Work

A. Generative Adversarial Networks

Ideally, generative models learn how data is distributed, thus allowing data synthesis from the learned distribution. Since the advent of GANs [10], generative models have achieved impressive results in various tasks like data augmentation [11] and style transfer [12]. GANs try to learn the data distribution by approximating the similarity of distributions between the training data and the fake data produced by the learned model. GANs usually comprise a generator and a discriminator. The entire model learns by playing a minimax game: the generator tries to fool the

Manuscript received October 14, 2022; revised July 22, 2022; accepted November 11, 2022.

Lei Luo and William H. Hsu are with the Computer Science Department, Kansas State University, Manhattan, KS, 66506, USA.

Shangxian Wang is with the Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, 21210, USA.

*Correspondence: leiluoray@ksu.edu(L.L.)

doi: 10.18178/ijml.2023.13.2.1132

discriminator by gradually generating realistic data samples, and the discriminator, in turn, tries to distinguish real samples from fake ones. GANs have been improved in various ways. To produce more realistic samples, an architecture of stacked GANs has been proposed: the Laplacian pyramid of GANs [13]; layered, recursive GANs [14]; progressive growing GANs [15]; and style based GANs (StyleGANs). Several studies have attempted to solve the instability training of GANs using energy based GANs [16], Wasserstein GANs [17], and boundary equilibrium GANs [18]. In this study, we use GANs with their improved techniques to learn the distribution of data and how to translate among different domains.

B. Unsupervised I2I Translation

Unsupervised I2I translation translates images from one domain to another without paired data supervision. Much success in unsupervised I2I translation is due to the cycle consistency constraint, proposed in three earlier works: CycleGANs [9], DiscoGANs [19], and DualGANs [20]. To translate more than two domains, MUNIT and DRIT are proposed. These methods, however, naively sample style codes from a standard normal distribution, which leads to inferior translation results. Moreover, they require training n(n-1) generators and n discriminators for translating images among n domains, which is computationally expensive and time-consuming. Our method proposes a simpler yet more effective approach that requires only one set of generator-discriminator. Recent systems such as StarGAN2 [21] and ModularGANs [22] are developed to perform multimodal I2I translation to produce images with the same content but different contexts. Inspired by StyleGANs, we employ a mapping network to model style codes of different domains. Furthermore, we add several regularization techniques to encourage the diversity in translated results.

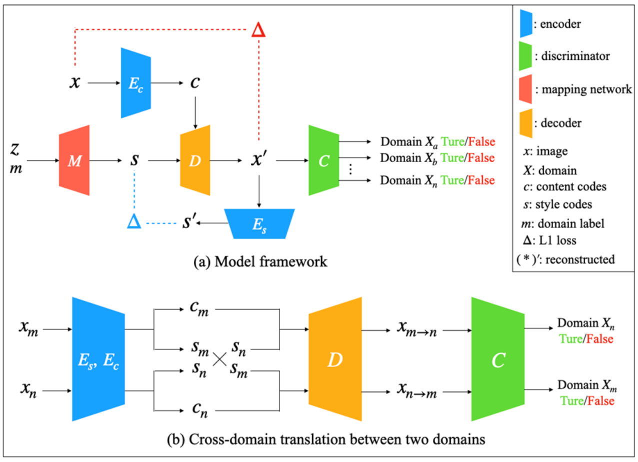
III. Methods

A. Preliminaries

Let x be an image that belongs to one of many domains. The graph (a) in Fig. 1 shows an overview of our model. We start from a latent vector z that is sampled from a standard normal distribution. z goes through a mapping network, which learns style codes s of a specific domain, where m is a domain label and $s = M(z, m)$. Meanwhile, we employ a content encoder E_c to extract content codes c from image inputs. The decoder D takes content and style codes to generate reconstructed images x', which are then used by style encoder E_s to produce reconstructed style codes s'. We compute two L1 losses using the reconstructed images and style codes. Finally, we use a multi-task discriminator to distinguish real images from fake ones. During the translation phase, we keep the same content codes but use the style codes of target domains. The graph (b) of Fig. 1 illustrates an example of image translation within two domains.

B. Model Architecture

We discuss the architectures of different modules in our framework in this section. Even though our framework is closely related to MUNIT and DRIT, we redesign the architecture of neural networks for better performance.



(a) Model framework

(b) Cross-domain translation between two domains

Fig. 1. The structure of our framework. (a) shows how our framework learns, and (b) shows cross-domain translation within two domains.

Encoder Our encoder has two sub-encoders: the style encoder and the content encoder. Both start with a convolution layer. The content encoder consists of six residual blocks [23]. All the layers are downsampled by average pooling operation (except for the last two layers) and are followed by an instance normalization (IN) [24]. The style encoder also comprises six residual blocks but without any activation function expect for the last residual block. Lastly, the style encoder consists of a convolution layer with leaky ReLU and a reshape operation before outputting style codes by a linear layer.

Mapping Network Style codes of domains are modelled by a mapping network, which consists of eight linear layers with ReLU activation function expect for the last layer.

Decoder The decoder maps latent codes, which consist of style codes and content codes, to the original image space. To apply style to images of different domain, the style codes are injected into the decoder by AdaIN [25] coupled with residual blocks. Inspired by StyleGANs, we also introduce stochastic variation into our model by injecting noise into the decoder. The decoder consists of six residual blocks with AdaIN, and the last layer is a convolution layer whose outputs are generated images.

Discriminator The architecture of discriminator is similar to that of the style encoder except that it has one more convolutional layer to predict domains.

C. Training Objectives

In this section, we discuss the loss functions for learning our framework.

Image Reconstruction Loss After images are encoded to style and content codes, the decoder maps the latent space back to the image space and reconstructs the image. Image reconstruction loss is formulated as:

$$L^{x}_{recon} = |D\big(E_c(x), M(z,m)\big) - x|_1 \tag{1}$$

where m is the domain, to which image x belongs.

Style Code Reconstruction Loss After encoding reconstructed images using the style encoder, we can obtain reconstructed style codes. We construct the style code reconstruction loss as follows:

$$L^{s}_{recon} = |s - E_s(x')|_1, \tag{2}$$

where $x' = D\big(E_c(x), M(z,m)\big)$ and $x \in X_m$.

Regularization on Style and Content Codes To further encourage style codes being domain-variant and content codes being domain-invariant, we add regularization on style and content encoders. The style regularizer forces style codes of different domains to be different by minimizing L_{regu}^{s}, which is calculated as:

$$L_{regu}^{s} = -| D(c_m, s_m) - D(c_m, s_n)|_1 \\ - | D(c_n, s_m) - D(c_n, s_n)|_1, \tag{3}$$

where $(c_m, s_m) = \left(E_c(x_m), E_s(x_m)\right)$ and $(c_n, s_n) = \left(E_c(x_n), E_s(x_n)\right)$. c_m and s_m are content and style codes of image $x_m \in X_m$. c_n and s_n are content and style codes of image $x_n \in X_n$.

The content regularizer encourages content codes of different domains to be similar by minimizing L_{regu}^{c}, which is formulated as:

$$L_{regu}^{c} = | D(c_m, s_m) - D(c_n, s_m)|_1 \\ + | D(c_m, s_n) - D(c_n, s_n)|_1. \tag{4}$$

Inspired by StarGAN2 [21], we calculate style diversity as:

$$L_{ds} = | E_s(x_1) - E_s(x_2)|_1, \tag{5}$$

where z_1 and z_2 are two random latent vectors; $x_1 = D\left(E_c(x), M(z_1, m)\right)$, and $x_2 = D\left(E_c(x), M(z_2, m)\right)$.

Adversarial Loss GANs are used to match the distribution of translated results to real image samples, so the discriminator finds real and fake samples indistinguishable. We use two adversarial losses with one for learning latent-guided translation and the other for reference-guided translation. Latent-guided translation refers to using the mapping network to obtain target style codes, and reference-guided translation uses the style encoder to extract style codes of target domains. The adversarial loss for learning the discriminator C_m with latent-guided translation is formulated as:

$$L_{adv}^{l} = \mathop{\mathbb{E}}_{z\sim N(0,I), x_n\sim p(X_n)} \left[log C_m \left(D\left(E_c(x_n), M(z, m)\right)\right)\right] \\ + \mathop{\mathbb{E}}_{x_m\sim p(X_m)} \left[log\left(1 - C_m(x_m)\right)\right], \tag{6}$$

and the adversarial loss for learning the discriminator $C_m$with reference-guided translation is constructed as:

$$L_{adv}^{r} = \mathop{\mathbb{E}}_{x_m\sim p(X_m), x_n\sim p(X_n)} \left[log C_m \left(D\left(E_c(x_n), E_s(x_m)\right)\right)\right] \\ + \mathop{\mathbb{E}}_{x_m\sim p(X_m)} \left[log\left(1 - C_m(x_m)\right)\right], \tag{7}$$

where the discriminator C_m tries to tell if images are from the domain m.

Full Objective Our full objective is formulated as follows:

$$\min_{M,E,D} \max_{C} \lambda_1 L_{recon}^{x} + \lambda_2 L_{recon}^{s} \\ + \lambda_3\left(L_{regu}^{s} + L_{regu}^{c}\right) \\ + \lambda_4\left(L_{adv}^{l} + L_{adv}^{r}\right) - \lambda_5 L_{ds}, \tag{8}$$

where λ_1 to λ_5 are hyperparameters for each loss term.

IV. Experiments

In this section we talk about data sets, baselines, evaluation metrics, and implementation details of our framework.

A. Baselines

We compare our framework against three baseline models developed in recent years. Our framework is closely related to MUNIT and DRIT, which we use as baseline models. More recent works on unsupervised I2I translation include StarGAN2, TransGaGa [26], ContrastiveGAN [27], and FUNIT [28], which achieved impressive results. Only StarGAN2, however, is used as another baseline as other works try to focus on different aspects of I2I. TransGaGa studies a different aspect of I2I than our work, which is how to preserve geometry information before and after image translation. FUNIT also focuses on a different problem, which is few-shot I2I. Our approach is not directly comparable to that of ContrastiveGAN, which crops input images into small patches and learns by increasing the mutual information between patches from the same location, while we treat the entire image as input.

B. Data Sets

We evaluate our framework on the CelebA-HQ and AFHQ data sets. Similar to StarGAN2, we also separate CelebA-HQ as domains of male and female, and AFHQ as domains of cat, dog, and wild. For fair comparison purposes, all images are trained with size 256 × 256, which is the largest resolution supported by the baselines.

C. Evaluation Metrics

We evaluate the visual quality using Fréchet inception distance (FID) [29] and the diversity of translated images with learned perceptual image patch similarity (LPIPS) [30]. Images generated by our framework are compared with the testing data set to calculate FID and LPIPS. Lower FID values indicate that the two sets of images have more similar distributions. Higher values of LPIPS indicate higher diversity of generated images.

D. Experiment Settings

We use a NVIDIA RTX 3090 GPU to conduct all our experiments. Adam optimizer is used for all the experiments with $\beta_1 = 0$, $\beta_2 = 0.999$, and initial learning rate of $1e^{-4}$ with weight decay of $1e^{-4}$. Batch size is set to 8 for all the experiments. For the CelebA-HQ date set we set hyperparameters λ_1 to λ_5 to 1, and when training on the AFHQ data set we set λ_5 to 2 and the rest to 1. We train all models for 100, 000 iterations, which take about 2.5 days. Model training using MUNIT and DRIT, however, take more than 6 days.

V. Results

In this section, we show the qualitative and quantitative results of the experiments. Ablation study is also carried out to evaluate the effectiveness of several key design choices.

A. Quantitative Results

Similar to StarGAN2, we perform reference-guided and latent-guided translation, examples of which are shown in Figs. 2 and 3. We use FID to evaluate the similarity of distributions and LPIPS to evaluate the diversity of generated images. As Table I and Table II show, performance by our method and StarGAN2 are close, both outperforming MUNIT and DRIT by a great margin except for latent-guided

LPIPS results of MUNIT on AFHQ. StarGAN2 achieves the lowest FID on both data sets, and our method achieves highest LPIPS among all models.

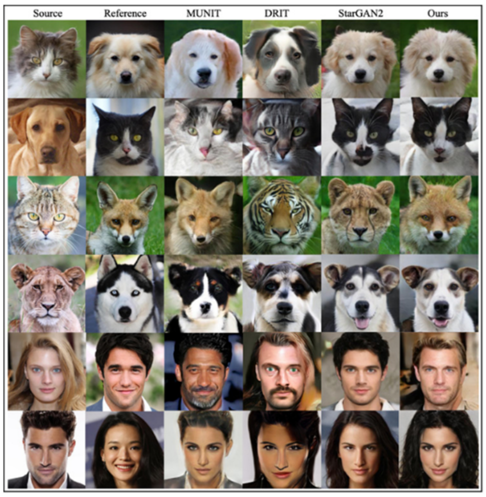


Fig. 2. Examples of reference-guided translation.

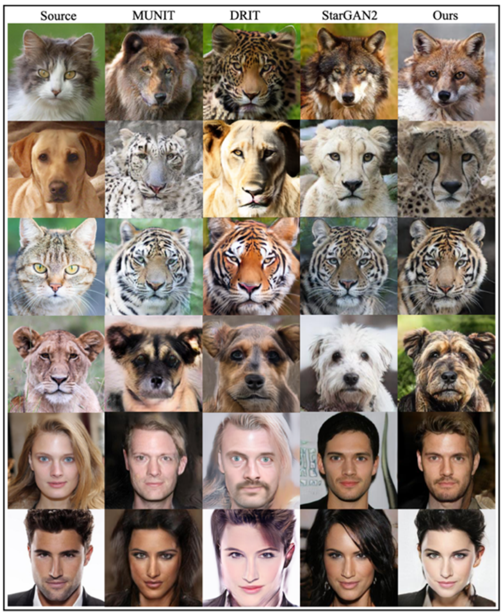


Fig. 3. Examples of latent-guided translation.

TABLE I: QUANTITATIVE RESULTS ON LATENT-GUIDED TRANSLATION

Models	CelebA-HQ		AFHQ	
	FID (↓)	LPIPS (↑)	FID (↓)	LPIPS (↑)
MUNIT	31.4	0.363	41.5	**0.511**
DRIT	52.1	0.178	95.6	0.326
StarGAN2	**13.7**	0.452	**16.2**	0.450
Ours	17.5	**0.459**	19.9	0.476
Test data	14.8	--	12.9	--

TABLE II: QUANTITATIVE RESULTS ON REFERENCE-GUIDED TRANSLATION

Models	CelebA-HQ		AFHQ	
	FID (↓)	LPIPS (↑)	FID (↓)	LPIPS (↑)
MUNIT	107.1	0.176	223.9	0.199
DRIT	53.3	0.311	114.8	0.156
StarGAN2	**23.8**	0.388	**19.8**	0.432
Ours	25.3	**0.391**	22.3	**0.439**
Test data	14.8	--	12.9	--

B. Qualitative Results

We utilize the Amazon Mechanical Turk (AMT) to compare our results against the baselines based on user preferences. Given a source image and a reference image, we instruct AMT workers to select the best transfer result among all models. We ask 60 questions for all ten workers. As shown in Table III, our method slightly outperforms StarGAN2 and exceed MUNIT and DRIT for a large margin.

TABLE III: VOTES FROM ATM WORKERS FOR MOST PREFERRED STYLE TRANSFER RESULTS

Models	Performance
MUNIT	2.820%
DRIT	9.050%
StarGAN2	43.50%
Ours	**44.63%**

C. Ablation Studies

To further validate effects of key design choices in our framework, we carry out ablation studies on the AFHQ data set, whose results are shown in Fig. 4 and Table IV. Let the model without style and content regularizer, and noise injection be the vanilla model. We can see that style regularizer is effective in increasing diversity in generated images.

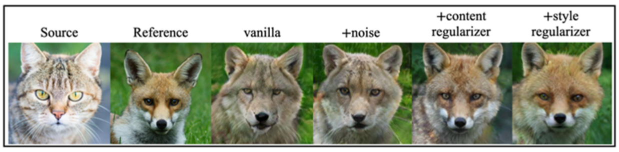


Fig. 4. Examples of reference-guided translation by adding modules.

TABLE IV: FID AND LPIPS RESULTS OF INCREMENTALLY ADDING MODULES TO OUR FRAMEWORK FOR REFERENCE-GUIDED TRANSLATION ON THE AFHQ DATA SET. THE VANILLA MODEL DOES NOT REPORT LPIPS RESULT AS IT IS A DETERMINISTIC MODEL

Models	FID (↓)	LPIPS (↑)
vanilla model	29.1	--
+ noise injection	27.6	0.407
+ content regularizer	23.8	0.414
+ style regularizer	**22.3**	**0.439**

VI. CONCLUSIONS

In this research, we present a simpler yet more effective framework for unsupervised multimodal multi-domain I2I translation. Our model only consists of a mapping network and a generator-discriminator pair. Unlike MUNIT and DRIT that simply sample style codes from a standard normal distribution when translating, we employ a mapping network to learn the style of different domains, which yields better translation results. To further encourage diversity in translated images, we employ style regularizations and inject Gaussian noise into the decoder. The qualitative and quantitative results show that our framework is superior or comparable to the SOTA baselines in unsupervised multimodal multi-domain I2I translation.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Lei Luo and Shangxian Wang conducted the research, performed the experiments, and wrote the first draft; William

Hsu edited the draft; all authors had approved the final version.

References

[1] X. Huang, M.-Y. Liu, S. J. Belongie, and J. Kautz, "Multimodal unsupervised I2I translation," in *Proc. 15th European Conference Computer Vision*, Munich, Germany, 2018, pp. 179–196.

[2] L. Luo and W.H. Hsu, "AMMUNIT: An attention-based multimodal multi-domain unsupervised image-to-image translation framework," in *Proc. ICANN 2022: 31st International Conference on Artificial Neural Networks*, Bristol, UK, 2022, pp 358–370.

[3] Y. Wang, X. Tao, X. J. Qi, X. Y. Shen, and J.-A. Jia, "Image inpainting via generative multi-column convolutional neural networks," in *Proc. Annual Conference on Neural Information Processing Systems*, Montreal, Canada, 2018, pp. 329–338.

[4] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 4401–4410.

[5] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse I2I translation via disentangled representations," in *Proc. 15th European Conference Computer Vision*, Munich, Germany, 2018, pp. 36–52.

[6] Z. H. Wang, J. Chen, and S. C. H. Ho, "Deep learning for image super-resolution: A survey," in *Proc. CoRRi*, 2019, vol. abs/1902.06068.

[7] Q. F. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," in *Proc. IEEE International Conference on Computer Vision*, Venice, Italy, 2017, pp. 1520–1529.

[8] P. Isola, J.-Y. Zhu, T. H. Zhou, and A. A. Efros, "I2I translation with conditional adversarial networks," in *Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition*, Hon-olulu, HI, USA, 2017, pp. 5967–5976.

[9] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired I2I translation using cycle-consistent adversarial networks," in *Proc. IEEE International Conference on Computer Vision*, Venice, Italy, 2017, pp. 2242–2251.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza *et al.*, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, Montreal, Canada, pp. 2672–2680 2014.

[11] L. Luo, W.H. Hsu, and S. Wang, "Data augmentation using generative adversarial networks for electrical insulator anomaly detection," in *Proc. 2nd Int. Conf. Manage. Sci. Ind. Eng*., Apr. 2020, pp. 231-236.

[12] L. Luo, W.H. Hsu, and S. Wang, "Shape-aware generative adversarial networks for attribute transfer," in *13th International Conference on Machine Vision. Vol. 11605.* SPIE, 2021.

[13] J.-Y. Zhu, P. Krahenbuhl, E. Shechtman, and A. A. Efros, "Generative visual manipulation on the natural image manifold," in *Proc. 14th European Conference Computer Vision,* Amsterdam, The Netherlands, 2016, pp. 597–613.

[14] R. Abdal, Y. P. Qin, and P. Wonka, " Image2stylegan: How to embed images into the stylegan latent space?" in *Proc. 2019 IEEE/CVF International Conference on Computer Vision*, Seoul, Korea (South) , 2019, pp. 4431–4440.

[15] E. L. Denton, S. Chintala, A. Szlam, and R. Fergus, "Deep generative image models using a laplacian pyramid of adversarial networks," *Advances in Neural Information Processing Systems*, Montreal, Canada, pp. 1486–1494, 2015.

[16] J. W. Yang, A. Kannan, D. Batra, and D. Parikh, 2017, "LR-GAN: Layered recursive generative adversarial networks for image generation," in *Proc. 5th International Conference on Learning Representations*, Toulon, France.

[17] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *Proc. 6th International Conference on Learning Representations*, Vancouver, Canada, 2018.

[18] J. J. Zhao, Mi. Mathieu, and Y. LeCun, "Energy-based generative adversarial networks," in *Proc. 5th International Conference on Learning Representations*, Toulon, France, 2017.

[19] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. the 34th International Conference on Machine Learning*, 2017, pp. 214–223.

[20] D. Rthelot, T. Schumm, and L. Metz, *BE-GAN: Boundary Equilibrium Generative Adversarial Networks*, CoRR, 2017.

[21] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proc. the 34th International Conference on Machine Learning*, Sydney, NSW, 2017, pp. 1857–1865.

[22] Z. L. Yi, H. Zhang, P. Tan, and M. L. Gong, "Dualgan: Unsupervised dual learning for I2I translation," in *Proc. IEEE International Conference on Computer Vision*, Venice, Italy, 2017, pp. 2868–2876.

[23] Y. Choi, Y. J. Uh, J. Yoo, and J.-W. Ha, "Stargan v2: Diverse image synthesis for multiple domains," in *Proc. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 8185–8194.

[24] B. Zhao, B. Chang, Z. Q. Jie, and L. Sigal, "Modular generative adversarial networks," in *Proc. 15th European Conference Computer Vision*, Munich, Germany, 2018, pp. 157–173.

[25] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016, pp. 770–778.

[26] D. Ulyanov, A. Vedaldi, and V. S. Lem-Pitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," in *Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, 2017, pp. 4105–4113.

[27] X. Huang and S. J. Belongie, "Arbitrary style trans-fer in real-time with adaptive instance normalization," in *Proc. IEEE International Conference on Computer Vision*, 2017, Venice, Italy, pp. 1510–1519.

[28] W. Wu, K. Cao, C. Li, C. Qian, and C. C. Loy, "Transgaga: Geometry-aware unsupervised I2I translation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, 2019, pp. 8012–8021.

[29] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired I2I translation," in *Proc.* 16th *European Conference Computer Vision*, Glasgow, UK, 2020, pp. 319–345.

[30] M.-Y. Liu, X. Huang *et al.*, "Few-shot unsupervised I2I translation," in *Proc. 2019 IEEE/CVF International Conference on Computer Vision*, Seoul, Korea (South), 2019, pp. 10550–10559.

[31] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two-time scale update rule converge to a local Nash equilibrium," in *Proc. Annual Conference on Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 6626–6637.

[32] R. Zhang, P. Isola, A. A. Efros, EliShechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. 2018 IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 586–595.