Applications and Limits of Image-to-Image Translation Models

Pasquale Coscia, Angelo Genovese, Fabio Scotti, Vincenzo Piuri
Department of Computer Science, Università degli Studi di Milano, Italy
{pasquale.coscia, angelo.genovese, fabio.scotti, vincenzo.piuri}@unimi.it

***Abstract*—Image-to-image (I2I) translation models are widely employed in several fields, *e.g.*, computer vision, security or medicine. Their goal is to map images from a source domain to a target domain while preserving content information. Despite their success, these models suffer from multiple weaknesses. For example, many practical scenarios do not consent to collect a sufficient amount of images, leading to imbalanced domains. Furthermore, mode collapse and training instability require a careful design and further discourage their deployment on edge devices. Finally, I2I models need an intensive computation to learn conditional probability distributions and are difficult to adapt to different contexts. These drawbacks mainly limit their large scale applicability. In this work, we want to shed light on the main solutions adopted to overcome the above issues and their impact on the performance. We also investigate several approaches to deploy these models on low-powered devices and weight sharing techniques to reduce the number of parameters and resources used.**

***Index Terms*—Image-to-image translation, GAN, cyclic loss**

I. Introduction

Deep generative modeling [1] represents a fascinating research area due to its numerous practical applications. In this context, image synthesis aims at generating a new image from another image (image-to-image) or a text (text-to-image), for example. In our work, we want to shed light on image-to-image (I2I) translation models, highlighting their main characteristics and limitations for practical scenarios that benefit from their application. There are numerous AI-based services applying I2I techniques for image in-painting, super-resolution, and scene synthesis. Nevertheless, in some contexts (*e.g.*, industrial [2] or medical [3]) privacy concerns and real-time constraints need to prioritize the deployment of in-loco solutions since data manipulation cannot be delegated to external services. Furthermore, neural networks trained on limited sets of images typically lead to poor quality outputs, requiring the use of ad-hoc solutions for improving their results. In this paper, we discuss several limitations that prevent I2I models to be applied to practical scenarios, along with commonly adopted solutions to overcome such limitations. The remainder of this paper is organized as follows. Section II describes the main characteristics of I2I models. Section III summarizes main I2I models. Section IV presents the main drawbacks and solutions adopted by major I2I methods. Section V illustrates several applications of I2I models to multiple scenarios. Section VI proposes research directions to be investigated. Finally, Section VII concludes the paper.

This work was supported in part by the EC under project EdgeAI (101097300), by the Italian MUR under PON project GLEAN, and by project SERICS (PE00000014) under the MUR NRRP funded by the EU - NextGenerationEU. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the Italian MUR. Neither the European Union nor Italian MUR can be held responsible for them.

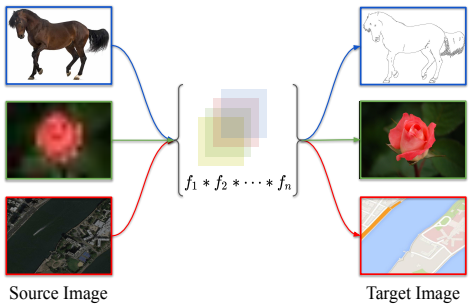

Fig. 1: Image-to-image (I2I) translation tasks can be regarded as an application of multiple filters to translate a source image to a target image. For example, multiple filters can be adopted for contours extraction, super-resolution or domain translation.

II. I2I translation

I2I models transfer the content information from a source to a target domain. This task can be supervised or unsupervised. Supervised I2I can be formulated as learning the conditional probability of samples drawn from a joint probability distribution. As shown in Fig. 1, for each image of the target domain, there exists its counterpart in the target domain. For example, for a satellite image, there is the corresponding map representation, or, for a low-resolution image, it is provided its high-resolution version. In this case, the model can leverage these paired images for learning a supervised translation. Nevertheless, collecting paired datasets is a time-consuming process and such data may only be used for specific contexts. For unsupervised I2I, conditional mappings must be learned from samples drawn from marginal distributions. In this case, images belong to two (or more) different domains that are weakly related. For example, to learn a translation between two seasons, images of different landscapes can be collected during winter and summer in order to learn specific elements of each context (*e.g.*, snow covered mountains or green grasslands). In this case, datasets can be easily collected yet several constraints must be adopted to solve this highly ill-posed problem. For example, to restrict the mapping space to reasonable high-

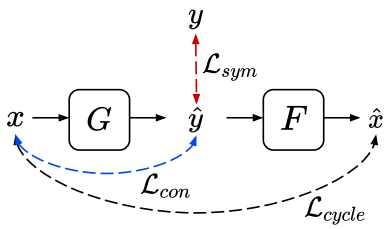

Fig. 2: Commonly adopted objective functions used for I2I tasks. The cyclic loss uses the translated image mapped back to the source domain as a form of content preservation while the contrastive loss typically compares source and translated patches.

quality images, several works employ cycle consistency or symmetry losses, weight-sharing [4] or shared latent spaces [5] (see Fig. 2). In this regard, an adversarial loss typically enforces target appearance while cycle consistency enables content preservation [6]. Contrastive learning is also widely employed for unsupervised learning. Its aim is to maximize the mutual information between source and translated images instead of minimizing their point-wise absolute deviation [6]–[8].

III. I2I APPROACHES

Pix2Pix [9] represents the first successful proposed conditional image synthesis method employing a conditional GAN (cGAN) for paired I2I translation mapping from input pixels to output pixels. This method combines an adversarial loss and an L_1 loss to force low-frequency correctness. A patch-based Markovian discriminator (PatchGAN), instead, is used for high-frequencies evaluation. Since paired data are difficult to collect, CycleGAN [10] introduces a cycle consistency loss to limit the mapping space and obtain an adversarial autoencoder. It also reduces mode collapse, a typical phenomenon of adversarial training. As Pix2Pix, CycleGAN ignores input noise, both in terms of random vectors or dropout, resulting in lack of diversity.

To reduce both the number of parameters and mappings when multiple domains are considered, StarGAN [11] proposes a conditional generator to the target domain, represented as a vector, and uses this label vector to perform a translation to multiple domains. To enforce the output to be consistent with the target domain, it employs a classification loss in an adversarial fashion. Nevertheless, this multitask approach does not present any stochastic variations. To include diversity, DRIT [4] disentangles domain content and domain attribute spaces using weight sharing among the considered domains. To enforce cyclic reconstruction, it proposes a cross-cycle constraint using cross-translations between input images: more simply, attribute representations are preserved for each domain, while content representations are swapped and then restored. This form of cyclic reconstruction appears more robust than cyclic loss in CycleGAN [10], [12] and a proper interpolation into the latent space allows the model to explore the target space. A different assumption is proposed in UNIT [13] for the unsupervised setting. This work proposes a weight-sharing and a shared latent space constraint to perform a uni-modal translation. CUT [6] operates at patch-level rather than image-level, using contrastive learning to replace cycle-consistency. This method increases efficiency providing a one-sided translation and reduces the number of training samples to be used for maximizing the mutual information between source and translated patches. In this way, it implicitly enforces a shared latent space for patches related to similar areas and benefits from intra-relationships within the image.

Scalability still remains an important issue for I2I models due to the inherent difficulty to fully capture variations among multiple domains. StarGANv2 [14] disentangles image generation and style encoding to create a scalable approach able to generate diverse images across multiple domains. Different domains are considered with a multi-head mapping network to transform a latent code into a domain-specific style code. In this way, style reference is better defined and injected into the generator using adaptive instance normalization (AdaIN). NEGCUT [8] uncovers an important limitation of contrastive learning-based I2I methods, since their efficacy heavily relies on negative examples able to efficiently push closer positive to query patches. This method generates hard negative patches through contrastive adversarial training. However, set-labels may not be available to be associated to domains of interest. This limitation is addressed in TUNIT [15], which investigates a *truly* unsupervised setting, where pseudo-domain labels are obtained maximizing the mutual information between pairs of samples while style features are defined by means of a contrastive loss.

Previous methods consider source and target images as a whole and tend to fail with images containing multiple instances, *e.g.*, different cars or objects, showing limited diversity and generation capability to translate specific instances. Multi-instance transfiguration problems, for example, deal with performing this type of translation [16]–[18].

IV. LIMITATIONS AND SOLUTIONS

I2I models suffer from several drawbacks that could limit their applicability only to specific contexts. In the following, we list their main limitations and proposed solutions.

A. Mode collapse and training instability

Adversarial training is commonly adopted to ensure high-quality outputs yet suffers from instability due to the high-dimensional non-convex space and may not converge limiting the diversity requirement required by this task. In this regard, LSGAN loss [19], or, for non-overlapping distributions, Wasserstain distance metric [20], is typically adopted. MSGAN [21] proposes regularization terms to avoid mode collapse. Jung *et al.* [22] employ a decoupled contrastive loss to avoid the vanishing gradient problem caused by *easy* negative samples.

B. Imbalanced or limited data

Class-imbalance or unbalanced domains negatively impact on the learned representations. I2I methods, in fact, implicitly

consider a symmetry between translated domains from both styles and data quantity perspectives. This assumption may not be met in practical scenarios. To this end, Pizzati *et al*. [23] employ additional web-crawled data sharing similar annotations to original data to solve domain shift or weather condition translation. Nevertheless, these annotations may be misleading and not properly reflect image content and style being highly subjective. Convex interpolation at features level is proposed in ReMix [24] for data augmentation. Similar to contrastive-based approaches, this method involves a similarity measure between pairs of features of real and interpolated images and does not require an unsupervised reconstruction for unseen targets. IrwGAN [25] proposes to reweigh aligned samples more than unaligned ones through a neural network. In this way, the authors favor paired data in large datasets with unpaired images. CACOLIT [26] adopts a similar cyclic concept proposed in CycleGAN [10] to perform a translation with data-poor domains. Specifically, it employs a data-rich domain that a teacher model uses for learning robust feature statistics; this knowledge is co-learned by two students with an additional constraint on instance-level outputs. One main drawback of this method stands in finding an appropriate auxiliary domain with similar characteristics to data-poor domains, limiting its practical applicability to industrial scenarios. To cope with small domains, Wang *et al*. [27] propose a data-free knowledge distillation procedure which takes advantage of a pre-trained I2I teacher to generate a large amount of style-mixed triplets, then used for training a student network via feature-based alignment.

Data limitation is also addressed by few-shot approaches which overcome mapping ambiguities using attributes disentanglement [28], additional inputs or multiscale processing [29]. ManiFest [30], for example, proposes an additional set of images and enforces features alignment while SEMIT [31] trains a pseudo-labeling classifier to annotate train unlabeled data.

As an example, we employ a paired dataset used for supervised translation in an unsupervised setting, and compare two methods, Cycle-GAN [10] and NEGCUT [8], to evaluate their ability to synthesize high-quality satellite images from maps in both normal (1096 images) and limited (548 and 110 images) data training regimes. Table I reports our results while Fig. 3 shows the qualitative outputs. Both models are trained for 200 epochs using default configurations. Our results demonstrate that a cyclic loss compared to a contrastive adversarial training appears better suited for generating aerial images in the case of scarcity of data, which has a detrimental impact on the generated images. On the contrary, for a full data training regime, generated images benefit from instance-wise hard negative samples.

C. Metrics

One main challenge of this task consists in using metrics able to really capture the quality of synthesized images. Frèchet Inception Distance (FID) [32] for measuring fidelity and Learned Perceptual Image Patch Similarity (LPIPS) [33]

TABLE I: Comparison between Cycle-GAN [10] and NEGCUT [8] models for $Maps \rightarrow Aerial\ photos$ translation. Three data training regimes are considered.

Model	Training regime	FID metric
Cycle-GAN	100%	69.48
NEGCUT		61.93
Cycle-GAN	50%	66.67
NEGCUT		62.46
Cycle-GAN	10%	99.51
NEGCUT		166.35

Fig. 3: Qualitative comparison of two I2I translation models for $Maps \rightarrow Aerial\ photos$ transformation. (a) Input, (b) Ground-truth, (c) Cycle-GAN (100%), (d) Cycle-GAN (50%), (e) Cycle-GAN (10%), (f) NEGCUT (100%), (g) NEGCUT (50%) and (h) NEGCUT (10%).

for assessing diversity appear unable to fully capture effective quality of synthesized samples, and it is rather common to complement quantitative analysis with studies involving human subjects. Bashkirova *et al*. [34] investigate this aspect, evaluating the quality of the disentanglement of domain-invariant and domain-specific attributes in labeled datasets.

D. Large models

To devise I2I models to low-powered edge devices, several techniques have been proposed. DMAD [35] conceives a differentiable mask to approximate a step function, removes unnecessary weights, and encourages sparsity in residual-based networks. To construct a reliable high-dimensional mapping, a co-attention distillation scheme transfers intermediate attention maps to a student network, improving both multiply-accumulate operations (MACs) and FID metric. Some attempts to propose a light-weighted I2I model is presented in Grassucci *et al*. [36] where correlations among RGB input channels are preserved using quaternion- and hypercomplex-based generative models. Benefiting from shared weight sub-matrices, this approach involves less parameters than traditional multi-modal approaches.

Since discriminators typically learn more expressive features than generators, CWT-GAN [37], for example, employs a cross-model weights transfer technique for sharing weights

between discriminators and generators. Another approach consists in using subnetworks for different tasks. Since input encoding is similarly performed for both the generation and classification phase, NICE-GAN [38] proposes to reuse part of the discriminator for the generation process. A decoupled training strategy is devised for the adversarial loss and avoids training conflicts.

To properly deploy models on edge devices, multiple factors should be taken into account, *e.g.*, memory, latency and number of operations. As compression method, Li *et al*. [39] propose to unify both paired and unpaired translations and distill knowledge to a student network. Subnetworks are then extracted and trained using a once-for-all (OFA) approach. This method drastically reduces the computation needed to train multiple networks for different devices. SoloGAN [40] also proposes weight sharing and cross distillation losses for increasing memory efficiency.

E. Validation

Lack of validation studies, mainly for industrial and medical fields [3], due to privacy concerns, shortages of pre-trained architectures or difficulties in reproducing similar operational settings, slow down the immediate applicability of I2I models.

V. I2I APPLICATIONS

I2I models are widely used in several real-world applications. In the following, we describe some adaptations of previous presented I2I models for industrial, remote sensing and medical scenarios.

a) Industrial scenarios: Liu *et al*. [41] propose a cGAN-based multi-discriminator framework to perform collocating clothes. This task consists in matching pairs of lower and upper clothes in a generative fashion. Co-supervised by attributes and categories, their models improve visual quality using multiple multitasks discriminators. Increasing scene perception at night represents another important goal for many industrial scenarios. Contours and textures are, in fact, lost by infrared (IR) cameras and require multiple steps for being reliable processed. Liu *et al*. [42] propose to translate thermal IR images into RGB images using two steps: a Texture-Net architecture, which adds textures and focuses on details with a ROI pooling layer, and an image colorization network for translating processed images into colored ones.

Defect generation is an important I2I task involving defect-free and defective images. Niu *et al*. [43] focus on boundary defect samples, *i.e.*, defects that involve small regions or do not manifest relevant features. Since an inverse mapping is redundant, and normal samples can be easily collected, this method proposes an encoder-decoder network in whose latent space a detection hyper-surface divides normal and defective samples. An interpolation in the latent space creates defects with different strengths representing different stages of a production line, while an input segmentation mask can control both size and location of the defects. Differently from the previous method, Defect-GAN [2] controls both location and type of the defect using an additional map introduced via SPADE normalization [44]. This method allows output images to separate the background from the foreground and generate realistic samples. To constrain the feature space, this method uses a reconstruction loss for spatial distribution maps generated during the forward (normal $\rightarrow$ defect) and backward (defect $\rightarrow$ normal) steps. Likewise, Wang *et al*. [45] build upon StarGANv2 [14] a model to separate background and foreground elements, and generate multiple defects by capturing both style and content information.

Supervised approaches typically show superior performance than unsupervised ones. To overcome the lack of paired data, Wang *et al*. [46] define a pair of generation methods based on radiometry for fabric smoothness assessment of decolorized fabrics. This approach provides a good approximation of winkle patterns for creating a paired dataset.

b) Remote sensing scenarios: Images from satellites, drones or thermal sensors are also widely employed for I2I tasks since they are well-suited for measuring the effects of climate change, monitoring the state of lands or oceans, tracking geomorphological features or for short-term forecasting. For example, Vandal *et al*. [47] propose to consider a shared spectral reconstruction loss to preserve spectral information and generate unobserved spectral bands from multiple satellites. Other approaches [48], [49] combine generation and segmentation methods for synthesizing high-quality digital maps or devising panoptic-aware strategies to reduce the content loss during translation.

c) Medical scenarios: I2I models are also successfully applied to several medical domains [3], [50]. In this case, images typically regard computed tomography (CT), magnetic resonance imaging (MRI), x-ray or ultrasound. Main applications include denoising, segmentation and cross-modality translation.

VI. RESEARCH DIRECTIONS

Despite several datasets for I2I translation tasks are available, they find limited applicability to real-world scenarios. In this regard, new approaches should focus on reusing pre-trained architectures to transfer low-level features more easily adaptable to different contexts, or investigating advanced optimization strategies, especially on edge devices.

VII. CONCLUSION

Image-to-image (I2I) translation models represent a valuable aid for solving industrial, medical and optical tasks, but several obstacles still limit their large scale diffusion. In this paper, we illustrated the problem and provided an analysis of some solutions to overcome the main drawbacks of I2I methods. Our analysis shows the need of publicly available data and models for applications external to academia for independent validation, and proposes new research directions to be investigated.

REFERENCES

[1] S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks, "Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022.

[2] G. Zhang, K. Cui, T.-Y. Hung, and S. Lu, "Defect-gan: High-fidelity defect synthesis for automated defect inspection," in *Proc. of WACV*, 2021.

[3] J. Chen, S. Chen, L. Wee, A. Dekker, and I. Bermejo, "Deep learning based unpaired image-to-image translation applications for medical physics: a systematic review," *Physics in Medicine & Biology*, 2023.

[4] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proc. of ECCV*, 2018.

[5] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. of ECCV*, 2018.

[6] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *Proc. of ECCV*, 2020.

[7] F. Zhan, J. Zhang, Y. Yu, R. Wu, and S. Lu, "Modulated contrast for versatile image synthesis," in *Proc. of CVPR*, 2022.

[8] W. Wang, W. Zhou, J. Bao, D. Chen, and H. Li, "Instance-wise hard negative example generation for contrastive learning in unpaired image-to-image translation," in *Proc. of ICCV*, 2021.

[9] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. of CVPR*, 2017.

[10] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. of ICCV*, 2017.

[11] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. of CVPR*, 2018.

[12] A. Genovese, V. Piuri, and F. Scotti, "Towards explainable face aging with generative adversarial networks," in *Proc. of ICIP*, 2019.

[13] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. of NeurIPS*, 2017.

[14] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "Stargan v2: Diverse image synthesis for multiple domains," in *Proc. of CVPR*, 2020.

[15] K. Baek, Y. Choi, Y. Uh, J. Yoo, and H. Shim, "Rethinking the truly unsupervised image-to-image translation," in *Proc. of ICCV*, 2021.

[16] Z. Shen, M. Huang, J. Shi, X. Xue, and T. S. Huang, "Towards instance-level image-to-image translation," in *Proc. of CVPR*, 2019.

[17] S. Mo, M. Cho, and J. Shin, "Instagan: Instance-aware image-to-image translation," in *Proc. of ICLR*, 2019.

[18] S. Kim, J. Baek, J. Park, G. Kim, and S. Kim, "Instaformer: Instance-aware image-to-image translation with transformer," in *Proc. of CVPR*, 2022.

[19] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proc. of ICCV*, 2017.

[20] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. of ICML*, 2017.

[21] Q. Mao, H.-Y. Lee, H.-Y. Tseng, S. Ma, and M.-H. Yang, "Mode seeking generative adversarial networks for diverse image synthesis," in *Proc. of CVPR*, 2019.

[22] C. Jung, G. Kwon, and J. C. Ye, "Exploring patch-wise semantic relation for contrastive learning in image-to-image translation tasks," in *Proc. of CVPR*, 2022.

[23] F. Pizzati, R. d. Charette, M. Zaccaria, and P. Cerri, "Domain bridge for unpaired image-to-image translation and unsupervised domain adaptation," in *Proc. of WACV*, 2020.

[24] J. Cao, L. Hou, M.-H. Yang, R. He, and Z. Sun, "Remix: Towards image-to-image translation with limited data," in *Proc. of CVPR*, 2021.

[25] S. Xie, M. Gong, Y. Xu, and K. Zhang, "Unaligned image-to-image translation by learning to reweight," in *Proc. of ICCV*, 2021.

[26] Y. Wang, T. Liang, and J. Lin, "Cacolit: Cross-domain adaptive co-learning for imbalanced image-to-image translation," in *Proc. of ACM Int. Conf. on Multimedia*, 2022.

[27] Y. Wang, J. van de Weijer, L. Yu, and S. Jui, "Distilling gans with style-mixed triplets for X2I translation with limited data," in *Proc. of ICLR*, 2022.

[28] Y. Chen, X. Yu, S. Liu, W. Gao, and G. Li, "Zero-shot unsupervised image-to-image translation via exploiting semantic attributes," *Image and Vision Computing*, 2022.

[29] J. Lin, Y. Pang, Y. Xia, Z. Chen, and J. Luo, "Tuigan: Learning versatile image-to-image translation with two unpaired images," in *Proc. of ECCV*, 2020.

[30] F. Pizzati, J.-F. Lalonde, and R. de Charette, "Manifest: Manifold deformation for few-shot image translation," in *Proc. of ECCV*, 2022.

[31] Y. Wang, S. Khan, A. Gonzalez-Garcia, J. van de Weijer, and F. S. Khan, "Semi-supervised learning for few-shot image-to-image translation," in *Proc. of CVPR*, 2020.

[32] 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. of NeurIPS*, 2017.

[33] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. of CVPR*, 2018.

[34] D. Bashkirova, B. Usman, and K. Saenko, "Evaluation of correctness in unsupervised many-to-many image translation," in *Proc. of WACV*, 2022.

[35] S. Li, M. Lin, Y. Wang, C. Fei, L. Shao, and R. Ji, "Learning efficient gans for image translation via differentiable masks and co-attention distillation," *IEEE Trans. Multimed.*, 2022.

[36] E. Grassucci, L. Sigillo, A. Uncini, and D. Comminiello, "Hypercomplex image- to- image translation," in *Proc. of IJCNN*, 2022.

[37] X. Lai, X. Bai, and Y. Hao, "Unsupervised generative adversarial networks with cross-model weight transfer mechanism for image-to-image translation," in *Proc. of ICCVW*, 2021.

[38] R. Chen, W. Huang, B. Huang, F. Sun, and B. Fang, "Reusing discriminators for encoding: Towards unsupervised image-to-image translation," in *Proc. of CVPR*, 2020.

[39] M. Li, J. Lin, Y. Ding, Z. Liu, J.-Y. Zhu, and S. Han, "Gan compression: Efficient architectures for interactive conditional gans," in *Proc. of CVPR*, 2020.

[40] S. Huang, C. He, and R. Cheng, "Sologan: Multi-domain multimodal unpaired image-to-image translation via a single generative adversarial network," *IEEE Trans. on Artificial Intelligence*, 2022.

[41] L. Liu, H. Zhang, X. Xu, Z. Zhang, and S. Yan, "Collocating clothes with generative adversarial networks cosupervised by categories and attributes: A multidiscriminator framework," *IEEE Trans. on Neural Networks and Learning Systems*, 2020.

[42] S. Liu, M. Gao, V. John, Z. Liu, and E. Blasch, "Deep learning thermal image translation for night vision perception," *ACM Trans. Intell. Syst. Technol.*, 2020.

[43] S. Niu, B. Li, X. Wang, and Y. Peng, "Region- and strength-controllable gan for defect generation and segmentation in industrial images," *IEEE Trans. on Industrial Informatics*, 2022.

[44] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. of CVPR*, 2019.

[45] R. Wang, S. Hoppe, E. Monari, and M. Huber, "Defect transfer gan: Diverse defect synthesis for data augmentation," in *Proc. of BMVC*, 2022.

[46] J. Wang, K. Shi, L. Wang, Z. Li, R. Pan, and W. Gao, "Decoloration of multi-color fabric images for fabric appearance smoothness evaluation by supervised image-to-image translation," *IEEE Access*, 2019.

[47] T. J. Vandal, D. McDuff, W. Wang, K. Duffy, A. Michaelis, and R. R. Nemani, "Spectral synthesis for geostationary satellite-to-satellite translation," *IEEE Trans. on Geoscience and Remote Sensing*, 2022.

[48] Y. Fu, S. Liang, D. Chen, and Z. Chen, "Translation of aerial image into digital map via discriminative segmentation and creative generation," *IEEE Trans. on Geoscience and Remote Sensing*, 2022.

[49] T. Zhang, F. Gao, J. Dong, and Q. Du, "Remote sensing image translation via style-based recalibration module and improved style discriminator," *IEEE Geoscience and Remote Sensing Letters*, 2022.

[50] L. Kong, C. Lian, D. Huang, z. li, Y. Hu, and Q. Zhou, "Breaking the dilemma of medical image-to-image translation," in *Proc. of NeurIPS*, 2021.