Old Dominion University

ODU Digital Commons

2020

Generative Adversarial Networks for Visible to Infrared Video Conversion

Mohammad Shahab Uddin
Old Dominion University, muddi003@odu.edu

Jiang Li
Old Dominion University, jli@odu.edu

Chiman Kwan (Ed.)

Follow this and additional works at: https://digitalcommons.odu.edu/ece_fac_pubs

Part of the Artificial Intelligence and Robotics Commons, Electrical and Computer Engineering Commons, and the OS and Networks Commons

Generative Adversarial Networks for Visible to Infrared Video Conversion

Mohammad Shahab Uddin and Jiang Li
Department of ECE, Old Dominion   University, Norfolk, VA

Abstract

Deep learning models are data driven. For example, the most popular convolutional neural network (CNN) model used for image classification or object detection requires large labelled databases for training to achieve competitive performances. This requirement is not difficult to be satisfied in the visible domain since there are lots of labelled video and image databases available nowadays. However, given the less popularity of infrared (IR) camera, the availability of labelled infrared videos or image databases are limited. Therefore, training deep learning models in infrared domain is still challenging. In this chapter, we applied the pix2pix generative adversarial network (Pix2Pix GAN) and cycle-consistent GAN (Cycle GAN) models to convert visible videos to infrared videos. The Pix2Pix GAN model requires visible-infrared image pairs for training while the Cycle GAN relaxes this constraint and requires only unpaired images from both domains. We applied the two models to an open-source database where visible and infrared videos provided by the signal multimedia and telecommunications laboratory at the Federal University of Rio de Janeiro. We evaluated conversion results by performance metrics including Inception Score (IS), Frechet Inception Distance (FID) and Kernel Inception Distance (KID). Our experiments suggest that cycle-consistent GAN is more effective than pix2pix GAN for generating IR images from optical images.

Keywords: Image Conversion, Generative Adversarial Network, Cycle-consistent Loss, IR Image, Pix2Pix, Cycle GAN

1. Introduction

Image-to-image conversion such as data augmentation [1] and style transfer [2] has been applied to recent computer vision applications. Traditional image conversion models had been investigated for specific applications [3-14]. Since the creation of the GAN model [15], it opened a new door to train generative models for image conversion. For example, computer vision researchers have successfully developed GAN models for day-to-night and sketch-to-photograph image conversions [16]. Two recent popular models that can perform image-to-image translations are Pix2Pix GAN [2] and Cycle GAN [16]. Pix2Pix GAN needs paired images for training whereas Cycle GAN relaxes this constraint and can be trained with unpaired images. In practice, paired images from different domains are often

difficult to obtain. Therefore, Cycle GAN is a better choice for image to image translation where paired images are not available.

IR image datasets are not largely available as compared to optical images. As a result, we face the shortage of data when we train models for object detection in IR domain. This problem can be mitigated by using the Cycle GAN model to covert labelled optical images to IR images. In this chapter, we evaluate two models, Pix2Pix GAN and Cycle GAN, for image conversion from optical domain to IR domain. We used four different datasets to perform the conversion and three metrics including Inception Score (IS), Frechet Inception Distance (FID) and Kernel Inception Distance (KID) to assess quality of the converted IR images.

2. Image to Image Conversion Models

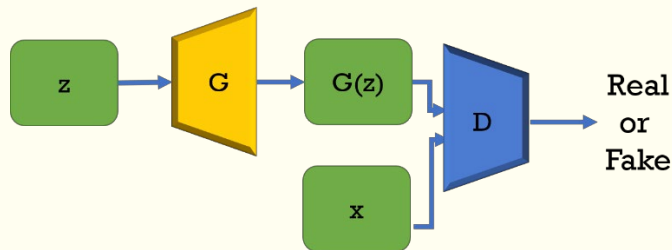### 2.1 Generative Adversarial Network

Figure 1. Structure of Generative Adversarial Network

GAN consists of one generative model and one discriminative model to generate images from noise as shown in Fig. 1. The generator 'G' tries to generate images from the input noise 'z' as realistic as possible to misguide the discriminator 'D' whereas 'D' is trained to discriminate the fake image '$G(z)$' from the real one 'x'. During training, errors at output 'D' are backpropagated to update parameters in 'G' and 'D', and the following loss function is optimized [15]:

$$\min_{G} \max_{D} V(D; G) = E_{\mathrm{x} \sim p_{data}(\mathrm{x})}[\log D(x)] + E_{\mathrm{z} \sim p_z(\mathrm{z})}[\log(1 - D(G(z)))] \tag{1}$$

where x and z represent training data and input noise, respectively. $p_{data}(x)$ and $p_z(z)$ are distributions of training data and input noise. The discriminator 'D' is trained to minimize the probability of the generated fake image to be real so that it can correctly assign labels to '$G(z)$' and 'x' in Fig. 1. The generator 'G' is trained to maximize $D(G(z))$ or equivalently to minimize $\log(1 - D(G(z)))]$ in equ (1), generating realistic images. Essentially, the generator learns to generate real data's distribution given by the training dataset. Once the goal is achieved, the generator can be used to generate realistic images by sampling from the learned probability distribution.

2.2 Conditional GAN

GAN can be converted into a conditional model with auxiliary information that is used to impose condition on generator and discriminator [17]. In the conditional GAN model, additional data are fed into the generator and discriminator so that data generation can be controlled. The loss function in conditional GAN becomes [17]:

$$\min_{G} \max_{D} V(D;G) = E_{y \sim p_{data}(y)}[\log D(y|x)] \; + \; E_{z \sim p_z(z)}[\log(1 - D(G(z|x)))] \qquad (2)$$

where y and z are training data and input noise, respectively. The input noise z combined with extra information x generate the output $G(z|x)$. Fig. 2 shows the diagram of conditional GAN.
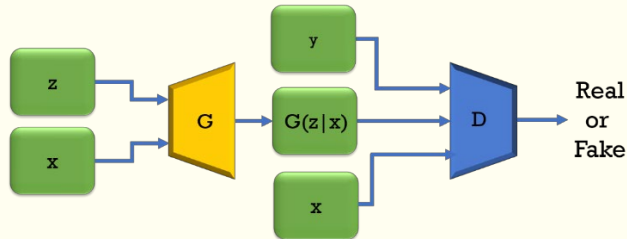

Figure 2. Architecture of Conditional GAN. Extra information x is given to both G and D. The discriminator trains itself to distinguish between real and fake image. The generator trains itself to fool discriminator by generating images similar to real images. Here both G and D get x as input.
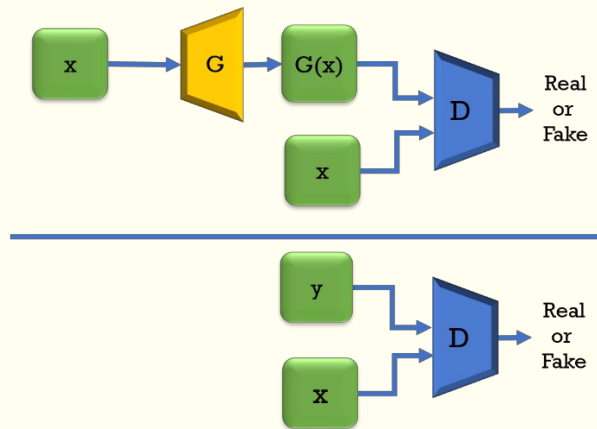
2.3 Pix2Pix GAN

Figure 3. Block Diagram of Pix2Pix GAN.

The Pix2Pix GAN model is built upon the concept of conditional GAN and it has been a common platform for various image conversion tasks. The diagram of Pix2Pix GAN model is given in Fig. 3. Pix2Pix GAN consists of a "U-Net" [18] based generator and a "PatchGAN" discriminator [2]. The "U-Net" generator passes low level information of input image to output image, and the "PatchGAN" discriminator helps capture statistics of local styles. The loss function of pix2pix GAN is:

$$\min_{G} \max_{D} V(D;G) = E_{x,y}[\log D(x,y)] \; + \; E_{x,z}[\log(1 - D(x, G(x,z)))] + E_{x,y,z}[||y - G(x,z)||_1]$$
$$(3)$$

Pix2Pix GAN learns to map input image x and random noise z to output image y. The generator tries to minimize the loss function while the discriminator tries to maximize the loss function. The L_1 loss between real image and fake one is included to achieve pixel level matching. Pix2Pix GAN had been applied to many applications including edges-to-photo conversion, sketch-to-photo conversion, map-to-aerial photo conversion etc. The main drawback of Pix2Pix GAN is that it needs paired images in both domains for training, which is not always possible in practice.

2.4 Cycle GAN

In many cases, it is difficult to get paired images from different domains. Cycle GAN [16] addressed this challenge by introducing the cycle-consistent loss function as shown in Fig 4. There are two generator G and F in Cycle GAN along with two adversarial discriminator D_x and D_y. X and Y are input domain and target domain, respectively. While D_x helps G to generate images from X domain to Y domain, F is trained to generate images from Y domain to X domain. G: X→Y and F:Y→X are two mappings that are trained in Cycle GAN and these are kept consistent by two cycle-consistency losses. The total loss function of Cycle GAN is given by:

$$\min_{G,F} \max_{D_x,D_y} L(G,F,D_x,D_y) = L_{GAN}(G,D_Y,X,Y) + L_{GAN}(F,D_x,Y,X) + \lambda L_{cyc}(G,F) \quad (4)$$

where

$$L_{GAN}(G,D_Y,X,Y) = E_{y\sim pdata(y)}[\log D_Y(y)] + E_{x\sim pdata(x)}[\log(1-D_Y(G(x)))]$$

$$L_{GAN}(F,D_x,Y,X) = E_{x\sim pdata(x)}[\log D_X(x)] + E_{y\sim pdata(y)}[\log(1-D_X(G(y)))]$$

$$L_{cyc}(G,F) = E_{x\sim pdata(x)}[[||G(F(x)) - x||_1] + E_{y\sim pdata(y)}[||G(F(y)) - y||_1]$$

There are two terms in the loss function of Cycle GAN: adversarial losses and cycle-consistency losses. The adversarial losses $L_{GAN}(G,D_Y,X,Y) + L_{GAN}(F,D_x,Y,X)$ for G: X → Y and F: Y → X mapping, respectively, ensure that target images' distribution and generated images' distribution are close. The cycle-consistency loss, $L_{cyc}(G,F)$, ensures that the two mappings have no contradictions. λ is a weight controlling balance between the two categories of losses.

Cycle GAN has been used in different applications including season transfer, style transfer, etc [16]. In addition, Cycle GAN has resolved the mode collapse problem in training if only the adversarial loss is used [19]. Mode collapse happens when the generator outputs the same image for different inputs. Though other methods [2, 8, 10-2, 20-24] can also offer image-to-image translation with unpaired images, Cycle GAN has become a common platform for many image translation related tasks.
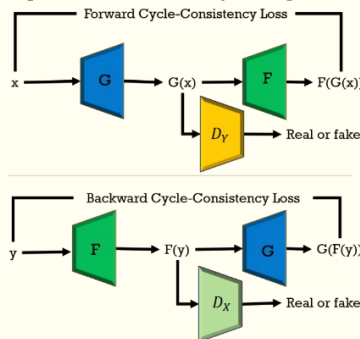

Figure 4. Overall Architecture of Cycle GAN

3. Experimental Setups

3.1 Datasets

For training Pix2Pix GAN and Cycle GAN, we have used images pairs from the open-source visible and infrared video database from the signal multimedia and telecommunications laboratory at the Federal University of Rio de Janeiro [25]. IR

and visible-light video pairs in the database are synchronized and registered. We utilized 80% of frames in the "Guanabara Bay_take_1" video pair for training and the remaining 20% frames for testing. In addition, we evaluated the trained model on other three image pairs named "Guanabara Bay_take_2", "Camouflage_take_1" and "Camouflage_take_2". Detailed information of the four video pairs are listed in Table 1 and some example pairs are shown in Fig. 5.

Table 1. Detailed Information of Video Pairs Used in Our Experiments.

Dataset Name	Description [25]
Guanabara Bay_take_1	• Contains scenes of "the Guanabara Bay and the Rio de Janeiro-Niteroi bridge". • Taken during Nighttime. • Contains 1 scene plane at approximately 500m distance.
Guanabara Bay_take_2	• Contains scenes of "the Guanabara Bay and the Rio de Janeiro-Niteroi bridge". • Taken during nighttime. • Contains 1 scene plane at approximately 500m distance.
Camouflage_take_1	• Contains outdoor scenes. • Taken during bright sunlight. • Contains 2 scene planes at approximately 10m and 300m distances. • Contains people who are hiding behind vegetation.
Camouflage_take_2	• Contains outdoor scenes. • Taken during bright sunlight. • Contains 2 scene planes at approximately 10m and 300m distances. • Contains people who are hiding behind vegetation.

a)   Visible Images   b) IR Images

Figure 5. Visible-IR Images from Guanabara Bay_take_1 Video Pair used for Training Pix2Pix GAN and Cycle GAN Models.

3.2 Performance Metrics

3.2.1 Inception Score

Inception score (IS) is widely used for evaluating GANs [26]. IS considers quality and diversity of generated images by evaluating the entropy of probability distribution created by the pre-trained 'Inception v3' model on the generated data [27]. A large inception score represents high quality of the generated images. One drawback of the inception score is that it does not consider information in the real images used for training the GAN model. Therefore, it is not clear how the generated images compare to the real training images.

3.2.2 Frechet Inception Distance

Frechet Inception Distance (FID) indicates the similarity between two sets of datasets and is often used for evaluating GANs [28-29]. FID is the Wasserstein-2 distance between feature representations of real and fake images computed by the Inception v3 model [27]. We used the coding layer of the Inception model to obtain feature representation of each image. FID is consistent with the human-judgement of image quality and it can also detect intra-class mode collapse. A lower FID score indicates that the two groups of images are similar so that the generated fake images are of high quality.

3.2.3 Kernel Inception Distance

Kernel Inception Distance (KID) is another metric often used to assess quality of GAN generated images relative to real images [30]. KID first uses the Inception v3 model to obtain representations of generated images. It then calculates the squared maximum mean discrepancy (MMD) between the representations of real training images and generated images. KID score is also consistent with human judgement of image quality. A small KID value indicates high quality of the generated images.

4 Results

4.1 Testing Results on "Guanabara Bay_take_1"

We trained the Pix2Pix GAN and Cycle GAN on 80% of the frames in "Guanabara Bay_take_1" video pair and tested the trained models on the remaining 20% frames. Some visible and IR images that we have used for training are shown in Fig. 5. After training, we applied both models to the testing frames and Fig. 6 shows some generated IR images. By visual inspection, Cycle GAN can generate better results than Pix2Pix GAN does. In addition, we observe that IR images generated by Cycle GAN are similar to the real IR images. Table 2 lists the quantitative performance metrics of the generated images by the two models. Cycle GAN outperforms Pix2Pix GAN in terms of all the metrics including IS, FID and KID on this dataset.
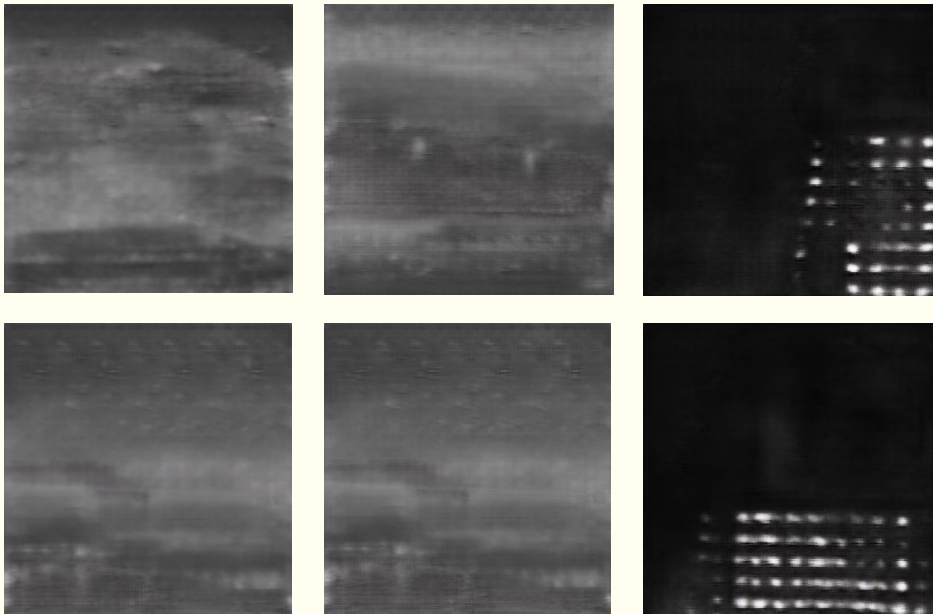

a) Generated IR Images by Pix2Pix GAN
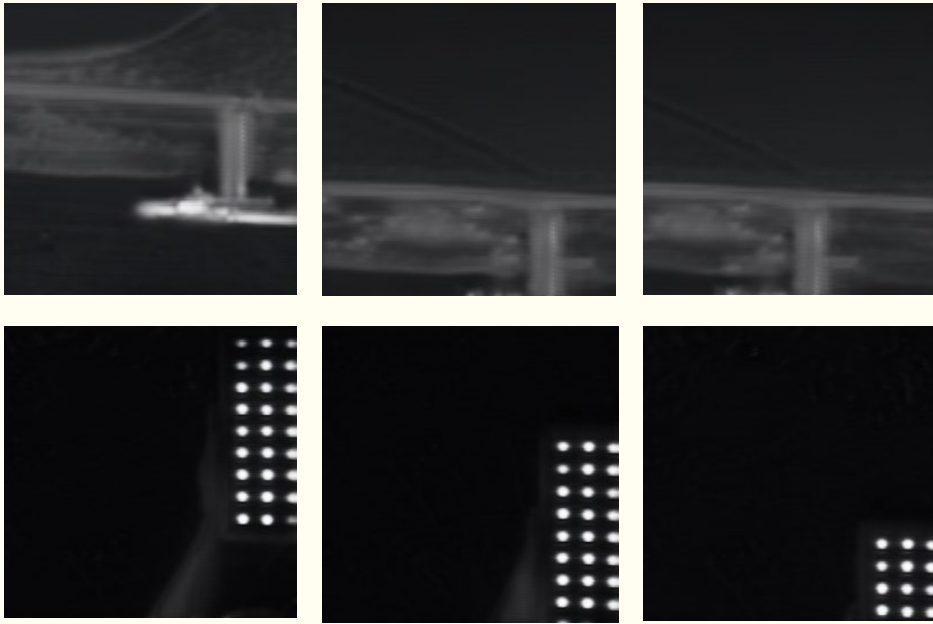
b) Generated IR Images by Cycle GAN

Figure 6. Fake IR Images generated by Pix2Pix GAN and Cycle GAN from the visible images in the Guanabara Bay_take_1 dataset.

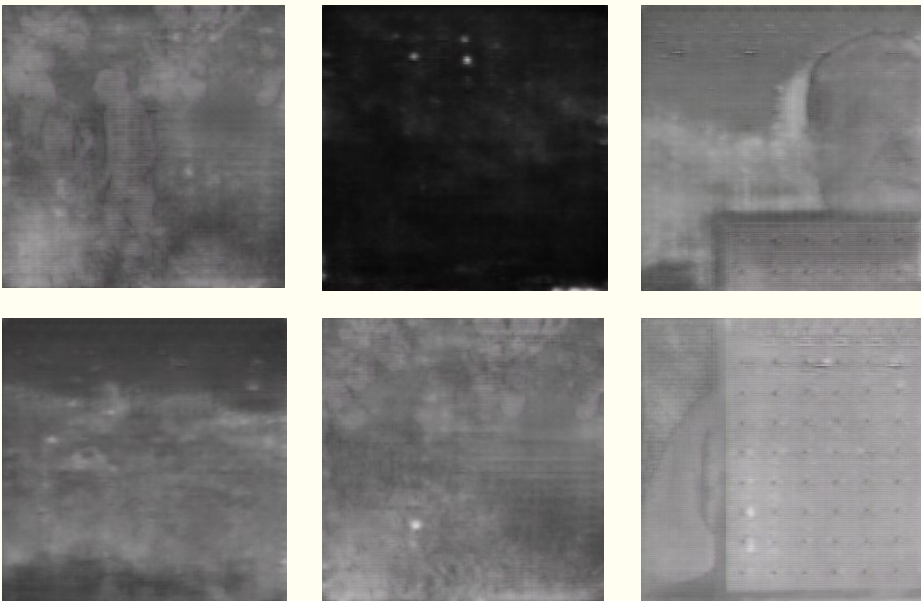a)   Generated IR Images by Pix2Pix GAN Cycle GAN

b) Generated IR Images by cycle GAN

Figure 7. Fake IR Images generated by Pix2Pix GAN and Cycle GAN from the visible images of Guanabara Bay_take_2 dataset.

4.3 Testing Results on "Camouflage_take_1" and "Camouflage_take_2"

We have applied the trained models to "Camouflage_take_1" and "Camouflage_take_2" datasets and results are shown in Figs. 8 and 9. Both models did not generate good quality IR images though the quantitative metrics as shown in Table 2. Cycle GAN is slightly better than Pix2Pix GAN. One possible reason is that the data in the two sets have different distributions as those in the training data, making both models failed.
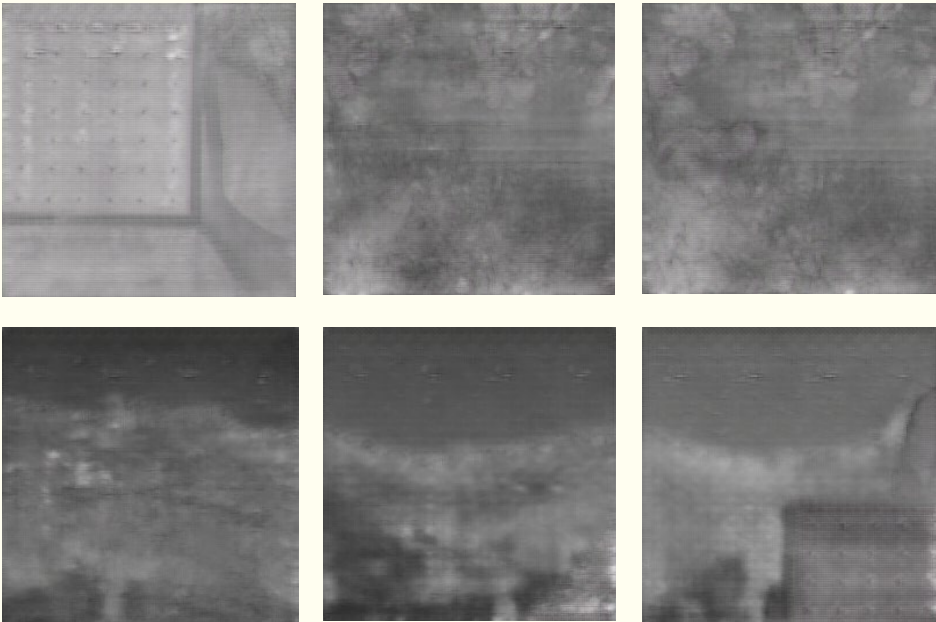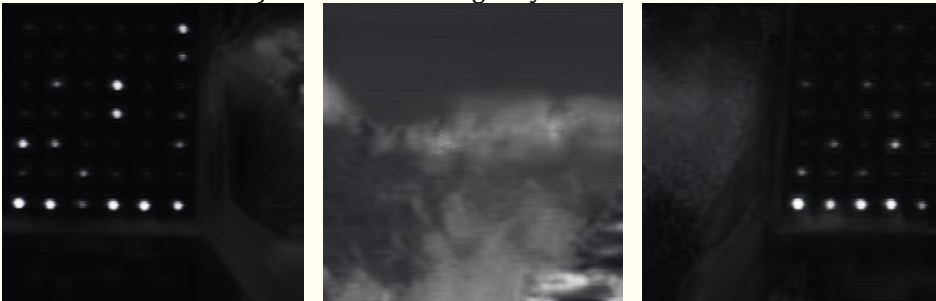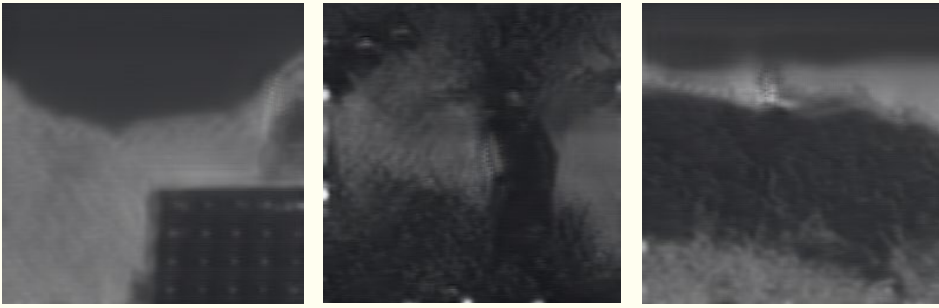

a)   Generated IR Images by Pix2Pix GAN
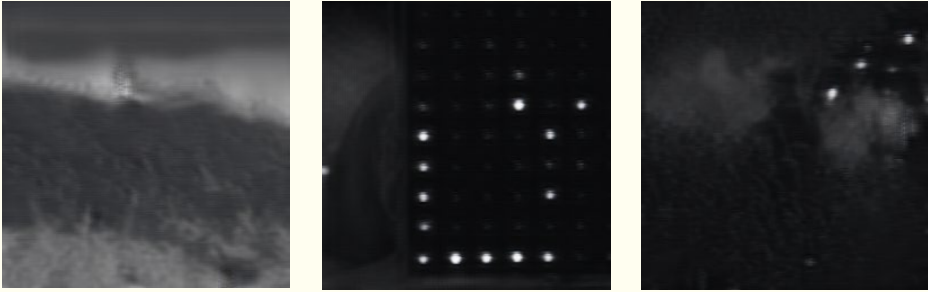
b) Generated IR Images by Cycle GAN

Figure 8. Fake IR Images generated by Pix2Pix GAN and Cycle GAN from the visible images of Camouflage_take_1 dataset.

a) Generated IR Images by Pix2Pix GAN

b) Generated IR Images by Cycle GAN

Figure 9. Fake IR Images generated by Pix2Pix GAN and Cycle GAN from the visible images of Camouflage_take_2 dataset

Table 2. *Evaluation Metrics on Generated IR Images of Different Datasets using Pix2Pix GAN and Cycle GAN.*

Metrics	Datasets							
	Guanabara Bay_take_1		Guanabara Bay_take_2		Camouflage take_1		Camouflage take_2	
IS Score	PixPix GAN	Cycle GAN	PixPix GAN	Cycle GAN	PixPix GAN	Cycle GAN	PixPix GAN	Cycle GAN
	2.70	2.88	1.85	3.61	1.02	2.72	1.02	2.66
FID	0.90	0.84	2.33	1.12	3.64	1.51	3.35	1.52
KID	4.24	2.42	24.00	7.10	48.61	9.13	43.55	9.15

4. Conclusion

In this chapter, we have investigated visible-to-IR image conversion using Pix2Pix GAN and Cycle GAN. Cycle GAN is a better model than Pix2Pix GAN and both can generate good visual quality IR images based on visible images, if training data and test data are similar. Overall, IR images generated by Cycle GAN have sharper appearances and better quantitative performance metrics than those by Pix2Pix GAN. However, if testing data have significant distribution shift as compared to training data, both models cannot generate quality IR images. Therefore, our recommendations are 1). Cycle GAN appears to be a better tool to convert optical images to IR images if training and testing datasets have similar distributions and 2) Both models are sensitive to distribution shift and additional techniques are needed to address the challenge.

References

[1] Fahimi F, Dosen S, Ang KK, Mrachacz-Kersting N, Guan C. Generative Adversarial Networks-Based Data Augmentation for Brain–Computer Interface. IEEE transactions on neural networks and learning systems. 2020.

[2] Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition 2017 (pp. 1125-1134).

[3] Zhao H, Yang H, Su H, Zheng S. Natural Image Deblurring Based on Ringing Artifacts Removal via Knowledge-Driven Gradient Distribution Priors. IEEE Access. 2020 Jul 8;8:129975-91.

[4]

[5]   Su JW, Chu HK, Huang JB. Instance-aware Image Colorization. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020 (pp. 7968-7977).

[6]   Park B, Yu S, Jeong J. Densely connected hierarchical network for image denoising. InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops 2019 (pp. 0-0).

[7]   Chen T, Cheng MM, Tan P, Shamir A, Hu SM. Sketch2photo: Internet image montage. ACM transactions on graphics (TOG). 2009 Dec 1;28(5):1-0.

[8]   Shi W, Qiao Y. Fast Texture Synthesis via Pseudo Optimizer. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020 (pp. 5498-5507).

[9]   Anwar S, Barnes N. Real image denoising with feature attention. InProceedings of the IEEE International Conference on Computer Vision 2019 (pp. 3155-3164).

[10]   Pan L, Dai Y, Liu M. Single image deblurring and camera motion estimation with depth map. In2019 IEEE Winter Conference on Applications of Computer Vision (WACV) 2019 Jan 7 (pp. 2116-2125). IEEE.

[11]   Shih Y, Paris S, Durand F, Freeman WT. Data-driven hallucination of different times of day from a single outdoor photo. ACM Transactions on Graphics (TOG). 2013 Nov 1;32(6):1-1.

[12]   Laffont PY, Ren Z, Tao X, Qian C, Hays J. Transient attributes for high-level understanding and editing of outdoor scenes. ACM Transactions on graphics (TOG). 2014 Jul 27;33(4):1-1.

[13]   Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition 2015 (pp. 3431-3440).

[14]   Eigen D, Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In Proceedings of the IEEE international conference on computer vision 2015 (pp. 2650-2658).

[15]   Fergus R, Singh B, Hertzmann A, Roweis ST, Freeman WT. Removing camera shake from a single photograph. In ACM SIGGRAPH 2006 Papers 2006 Jul 1 (pp. 787-794).

[16]   Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In Advances in neural information processing systems 2014 (pp. 2672-2680).

[17]   Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision 2017 (pp. 2223-2232).

[18]   Mirza M, Osindero S. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784. 2014 Nov 6.

[19]   Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention 2015 Oct 5 (pp. 234-241). Springer, Cham.

[20]   Goodfellow I. NIPS 2016 tutorial: Generative adversarial networks. arXiv preprint arXiv:1701.00160. 2016 Dec 31.

[21]   Eigen D, Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In Proceedings of the IEEE international conference on computer vision 2015 (pp. 2650-2658).

[22]   Johnson J, Alahi A, Fei-Fei L. Perceptual losses for real-time style transfer and super-resolution. In European conference on computer vision 2016 Oct 8 (pp. 694-711). Springer, Cham.

[23]   Wang X, Gupta A. Generative image modeling using style and structure adversarial networks. In European conference on computer vision 2016 Oct 8 (pp. 318-335). Springer, Cham.

[24]   Xie S, Tu Z. Holistically-nested edge detection. In Proceedings of the IEEE international conference on computer vision 2015 (pp. 1395-1403).

[25]   Zhang R, Isola P, Efros AA. Colorful image colorization. In European conference on computer vision 2016 Oct 8 (pp. 649-666). Springer, Cham.

[26]   Ellmauthaler A, Pagliari CL, da Silva EA, Gois JN, Neves SR. A visible-light and infrared video database for performance evaluation of video/image fusion methods. Multidimensional Systems and Signal Processing. 2019 Jan 15;30(1):119-43.

[27]   Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. Improved techniques for training gans. In Advances in neural information processing systems 2016 (pp. 2234-2242).

[28]   Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 2818-2826).

[29]   Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. InAdvances in neural information processing systems 2017 (pp. 6626-6637).

[30]   Fréchet M. Sur la distance de deux lois de probabilité. COMPTES RENDUS HEBDOMADAIRES DES SEANCES DE L ACADEMIE DES SCIENCES. 1957 Jan 1;244(6):689-92.

[31]   Bińkowski M, Sutherland DJ, Arbel M, Gretton A. Demystifying mmd gans. arXiv preprint arXiv:1801.01401. 2018 Jan 4.