Journal of New Media
DOI:10.32604/jnm.2021.018027
Review

Review of GAN-Based Person Re-Identification

Zhiyuan Luo*

School of Computer & Software, Nanjing University of Information Science and Technology, Nanjing, 210044, China
*Corresponding Author: Zhiyuan Luo. Email: luozhiyuan_net@126.com
Received: 22 February 2021; Accepted: 02 March 2021

Abstract: Person re-ID is becoming increasingly popular in the field of modern surveillance. The purpose of person re-ID is to retrieve person of interests in non-overlapping multi-camera surveillance system. Due to the complexity of the surveillance scene, the person images captured by cameras often have problems such as size variation, rotation, occlusion, illumination difference, etc., which brings great challenges to the study of person re-ID. In recent years, studies based on deep learning have achieved great success in person re-ID. The improvement of basic networks and a large number of studies on the influencing factors have greatly improved the accuracy of person re-ID. Recently, some studies utilize GAN to tackle the domain adaptation task by transferring person images of source domain to the style of target domain and have achieved state of the art result in person re-ID.

Keywords: Person re-identification; generative adversarial network

1 Introduction

With the rapid development of modern society, urban population density is getting higher than ever before. In some public places, dense crowds easily lead to the occurrence of public safety incidents. In order to prevent and deal with such incidents in time, a large number of surveillance cameras have been installed and applied in public places. However, to deal with complex monitoring networks and massive monitoring data, it is difficult to quickly analyze and process manually. Thus, the need for computer vision and machine learning technology to assist or even replace human participation in surveillance applications arises. Person re-ID is a key component of video surveillance research, has gradually become a research hotspot in recent years and has attracted extensive attention. The purpose of person re-ID is to identify a person of interest in the view of the surveillance camera accurately and quickly among a large number of people in the view of other cameras in the surveillance network. The application of person re-ID technology can greatly reduce the degree of human participation in video surveillance, analyze the person track precisely and effectively, and play an important role in the prevention of crime and the maintenance of good social security.

At present, person re-ID has been widely used in many applications. In modern surveillance systems, high quality human face pictures cannot be obtained due to camera viewpoint, occlusion caused by environment or other pedestrians, and insufficient resolution. Since it is impossible to use human face to recognize pedestrians, person re-ID, which studies the full body features of pedestrians has become an important alternative technology.

After more than ten years of research by scholars and engineers, person re-ID has achieved remarkable success. Especially in recent years, with the development of deep learning, research on person re-ID based on deep learning has emerged continuously. Generative Adversarial Networks have also been introduced to person re-ID, GAN-based methods have achieved the state of the arts result and become one of the most important approaches in person re-ID.

2 Generative Adversarial Networks in Person Re-ID

Generative Adversarial Networks have achieved great success in the application of computer vision. Recently, with the introduction of image-to-image translation which aims to translate images from the source domain to the target domain, the usage of GAN has largely expanded in image generation. Isola et al. [1] propose a conditional adversarial network to learn the mapping from images of the source domain to the target domain. However, the main limitation of [1] is that it needs paired training samples which are usually impractical in application. To tackle this issue, Liu et al. [2] propose CoGAN to learn join distributions across domains without paired images. Zhu et al. [3] propose CycleGAN which introduced a cycle consistency loss to learn a pair of mapping functions between source domain and target domain. CycleGAN also employed identity mapping to ensure the color identity during transfer, the identity loss is written as:

$$L_{identity} = E_{y\sim p_{data}(y)}[\|G(y) - y\|_1] + E_{x\sim p_{data}(x)}[\|F(x) - x\|_1], \quad (1)$$

where x and y are input images from dataset X and Y, F is the mapping function converts images from X to Y and mapping function G does the opposite. Without the identity loss, the generators are free to change color of input images because the cycle consistency loss only constrains the behavior of the entire paired mapping.

Recently, GAN has been introduced to person re-ID [4–7] and achieved remarkable success. GAN can serve as a domain adaptation approach and data argument approach at the same, existing GAN based studies in person re-ID mainly utilize GAN to tackle the cross-camera adaptation problem or cross-domain adaptation problem. We will introduce the usages in detail in Section 3 and Section 4.

3 Cross-Camera Adaptation Approaches Based on GAN

Style variation is an essential challenge to person re-ID because person images captured by different cameras represent different styles due to the disparity of different camera settings and different surveillance locations. GAN has proved to be effective to narrowing the domain gap between images captured by different cameras and smoothing the disparities within the same camera [4,7].

Zhong et al. [4] propose CamStyle to generate supplementary training images with CycleGAN. Because the widely used person re-ID datasets are collected by multiple different cameras, and the images captured by different cameras are different in style. Given N cameras, to narrow the domain gap between images captured by these cameras, C_N^2 CycleGAN is trained to transfer images in each camera to the style of other cameras. The style transferred images were used as supplementary training samples, the original images and transferred images form the new training set.

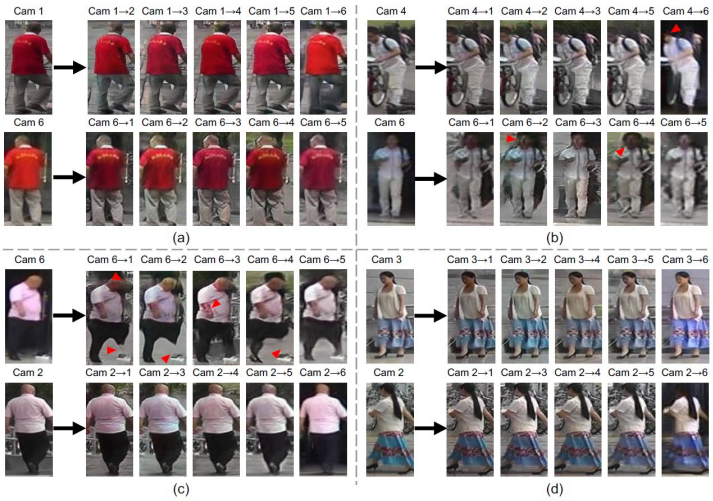

Figure 1: Examples of transferred images in Market-1501

There are six cameras in Market-1501, as shown in Fig. 1, for each training sample, CamStyle transfers it to the style of other five cameras and generates five additional training images. The generated images are similar to the images of the target domain in style. However, due to the inaccuracy of the mapping function learned by CycleGAN, there are some errors in the transferred images which will introduce noise to the system.

Fig. 2 shows the overall architecture of CamStyle. N is the number of cameras in the dataset. First, for each training sample, $\mathrm{N}-1$ transferred images are generated corresponding to other $\mathrm{N}-1$ cameras. Then, the real images and all style transferred images are mixed to form the new training set for supervised training. However, to deal with the noise introduced by the transferred images, the real images and the transferred images need to be treated differently. The cross-entropy loss is applied to the real images and label smooth regularization (LSR) is applied to the transferred images. LSR loss is written as follows:

$$L_{LSR} = -(1-\epsilon)\log p(y) - \frac{\epsilon}{C}\sum_{c=1}^{C}\log\ p(c), \tag{2}$$

where C is the number of classes, $p(c)$ is the predicted probability that the input image belonging to label c, and $p(y)$ is the predicted probability that the person image with identity y is belonging to label y.

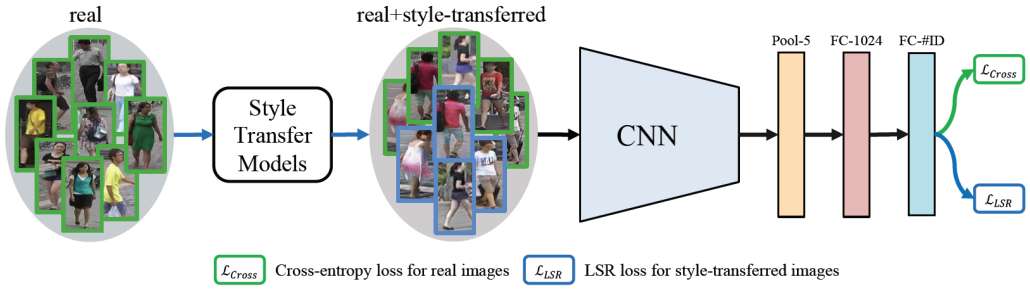

Figure 2: The overall architecture of CamStyle

By introducing transferred images to the training set, CamStyle not only narrows the domain bias between images captured by different cameras, it also expands the training data set to provide more training samples for re-ID model.

CamStyle [4] reduce the domain bias across different cameras, however, there are still style disparities within the images captured by the same camera due to the difference of capture time and the variation of pedestrian n distance. Liu et al. [7] consider the with-in camera disparites and propose UnityStyle to further reduce the domain bias between different images.

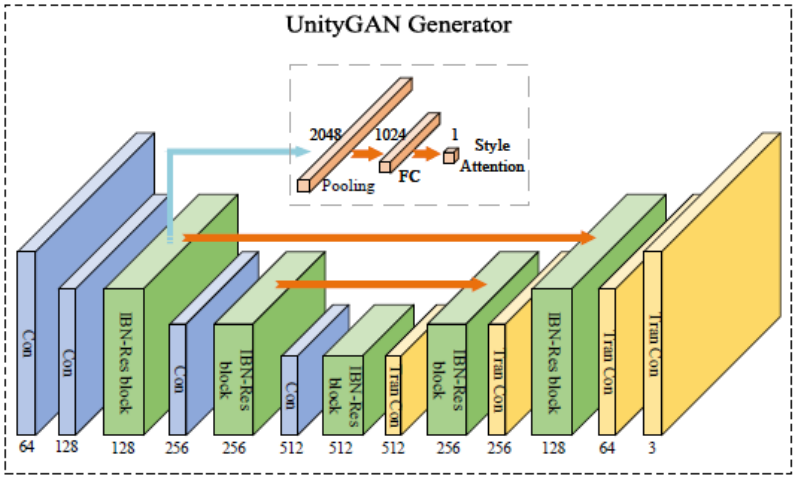

Figure 3: The architecture of UnityGAN generator

UnityStyle utilizes UnityGAN to generate UnityStyle images which can narrow the domain gap of images captured by the same camera and across different cameras. Different from CycleGAN aims to learn mapping functions between each pair of styles, UnityGAN aims to generate a uniform style for all images. UnityGAN takes advantage of both DiscoGAN and CycleGAN, then the combined network is improved by introducing residual blocks and skip connections. Moreover, IBN-Res block based on [8] is adopted to increase the robustness of style changes. The architecture of UnityGAN generator is shown in Fig. 3.

The identity loss of UnityStyle is the same one as we discussed in Section 2, which can be written as:

$$L_{ID} = E_{x\sim l_x}(\|F(x) - x\|_1) + E_{y\sim l_y}(\|G(y) - y\|_1), \quad (3)$$

where F is the mapping function used to transfer images from domain X to Y and mapping function G does the opposite. Moreover, to guarantee the UnityGAN can generate UnityStyle images, Style Attention module is introduced to the UnityGAN generator, it is defined as:

$$A(x) = Sigmoid(A_{style}(G_1(x))), \quad (4)$$

where G_1 is the first IBN-Res block output. As shown in Fig. 4, we can see that there are multiple errors in the images generated by CycleGAN. Compared to CycleGAN, UnityGAN overcome the wrong structure problem and can generate more structurally stable images.

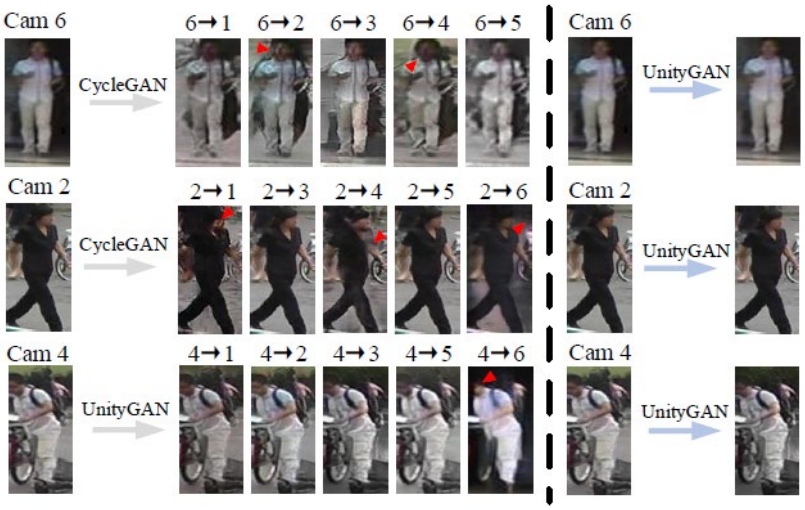

Figure 4: Images generated by CycleGAN and UnityGAN

To train person re-ID model, real images and UnityStyle images are combined as an enhanced training set for supervised training. Due to the high quality of the generated images, LSR adopted in CamStyle is no longer needed in UnityStyle, generated images can be treated as the same as the original images. In testing, UnityStyle images are used instead of the original images directly to ensure the query and gallery images are uniform styled.

4 Cross-Domain Adaptation Approaches Based on GAN

Although existing approaches have achieved remarkable success while training and testing models on the same dataset, their performance drops significantly while training and testing on different datasets. As shown in Fig. 5, person images of different datasets present different styles, different styles can be seen as different domains and re-ID method is easily affected by the existence of domain bias. In real-world applications, due to the expensive cost of annotating samples, it is usually impractical to collect enough training data. To alleviate the lack of training samples, some studies utilize GAN to transfer existing annotated data to the style of target dataset, transferred images used as supplementary training samples.

Figure 5: Person images of different datasets present different styles

Wei et al. [2] propose a data augment approach named Person Transfer GAN (PTGAN) to transfer annotated images of dataset A to the style of dataset B. As there are no paired image samples in the datasets, the cross-dataset transfer can be seen as unpaired image-to-image translation task. Due to the effectiveness of CycleGAN in unpaired image-to-image translation task, PTGAN utilizes CycleGAN as backbone network. Moreover, an additional constraint is adopted to maintain the color consistency during transfer, it can be written as:

$$L_{ID} = E_{a\sim p_{data}(a)}[\|(G(a)-a)\odot M(a)\|_2] + E_{b\sim p_{data}(b)}[\|(\overline{G}(b)-b)\odot M(b)\|_2], \quad (5)$$

where $a\sim p_{data}(a)$ is the data distribution of data set A, and $b\sim p_{data}(b)$ is the data distribution of data set B. G denotes the transformation function from dataset A to dataset B, and $\bar{G}$ denotes the transformation function from data set B to A. M(a) and M(a) represent the masks of input images a and b. Note that the calculation of L_{ID} is different from the typical identity loss described in Section 2. L_{ID} utilizes human segmentation extract the person mask to force the GAN to maintain the color consistency within the foreground area while the consistency of background area is ignored. The overall loss function of PTGAN is written as:

$$\mathrm{L} = L_{GAN}(G, D_B, A, B) + L_{GAN}(\bar{G}, D_A, B, A) + \lambda_1 L_{cyc} + \lambda_2 L_{ID}, \quad (6)$$

where D_A and D_B is the corresponding discriminator of dataset A and B.

As shown in Fig. 6, the quality of images transferred by PTGAN is significantly higher than that of CycleGAN.

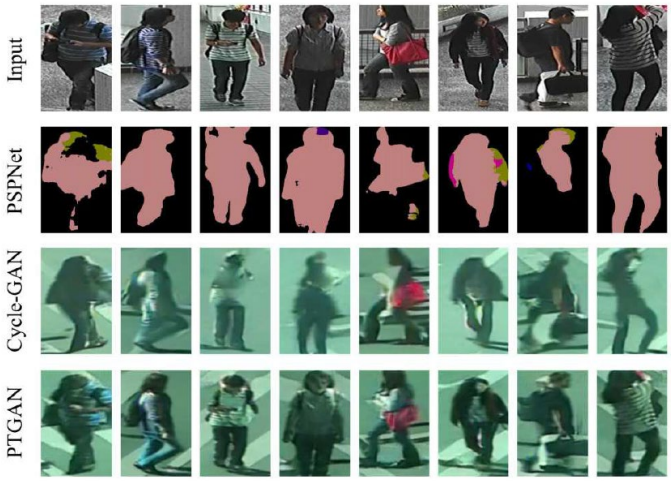

Figure 6: Images transferred from CUHK03 to PRID-cam1

While PTGAN tackles the cross-domain transfer task as a whole, Liu et al. [6] consider there are three main factors that affect the overall transfer quality, e.g., illumination, resolution and viewpoints. Liu et al. [6] propose to divide the cross-domain transfer-task into three sub tasks, each sub task trains a GAN network to tackle one of the three main factors. Then, an ensemble strategy is adopted to combine the encoded features of three networks together for person re-ID.

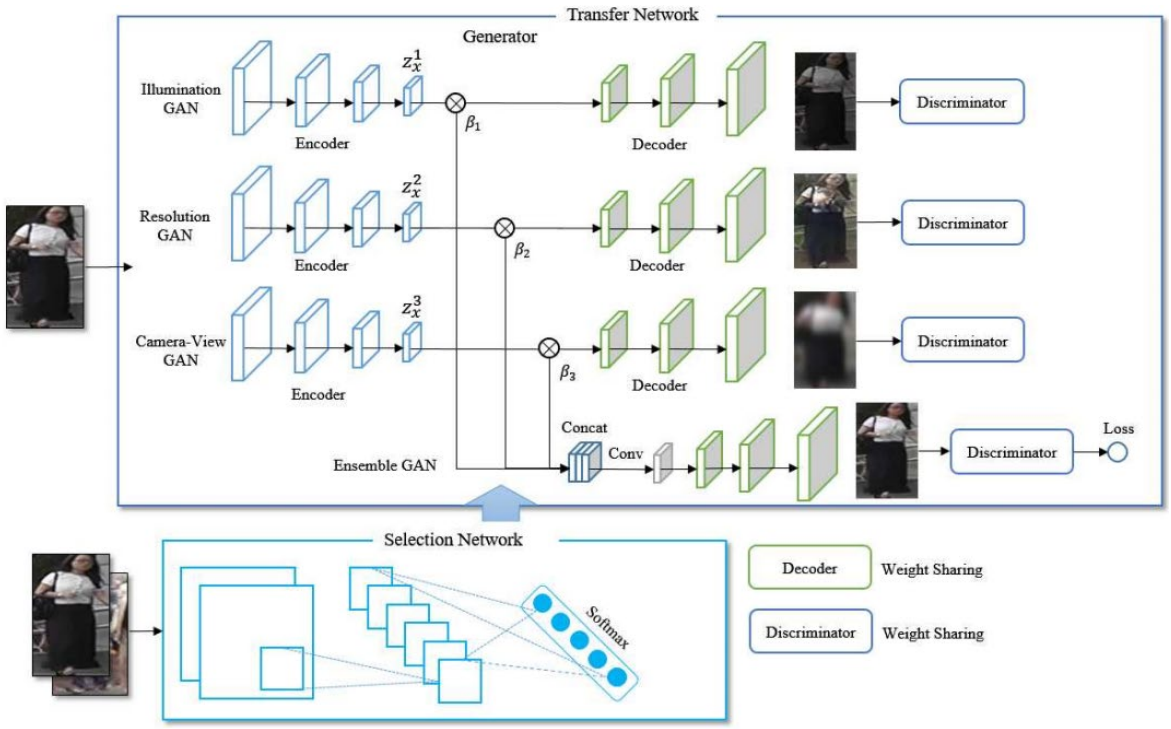

Figure 7: The overall architecture of ATNet

As shown in Fig. 7, there are four sub networks in ATNet: The Illumination GAN, the Resolution GAN, the Camera-View GAN and the Selection Network. The three transfer networks are based on CycleGAN, the loss function can be written as:

$$L_{gan} = L_{adv} + \lambda_1 L_{cyc} + \lambda_2 L_{ide}, \tag{7}$$

where L_{cyc} denotes the cycle loss function, and L_{ide} ensure the transfer function keeps the color consistency after image translation. The Illumination GAN is designed to transfer images from the illumination scale of source domain to the scale of target domain. It is trained on the simulated dataset containing images with different illumination scales. To further force the Illumination GAN focus on the illumination variation, an illumination constraint is adopted and shown as follows:

$$L_{ill} = E_{x\sim p(x)}[\left\|H\big(G(x)\big) - H(x)\right\|_1], \tag{8}$$

where H(·) denotes abstracting illumination insensitive features [9], the overall loss function of the Illumination GAN is written as $L_{gan} + \eta_1 L_{ill}$. The Resolution GAN aims to transfer images from the resolution of source domain to the resolution of target domain. It is trained on the simulated dataset containing images of different resolutions. Moreover, to guarantee the Resolution GAN focus on the resolution variation, the resolution constraint is adopted and shown as follows:

$$L_{res} = E_{x\sim p(x)}[\left\|I\big(G(x)\big) - I(x)\right\|_2^2], \tag{9}$$

where I(·) denotes extracting resolution-insensitive features [10], the overall loss function of the Resolution GAN is written as $L_{gan} + \eta_2 L_{res}$. The Camera-View GAN is trained on the simulated dataset

containing images captured by different cameras, no additional constraint is added to the loss function of the Camera-View GAN. Finally, the Emsemble GAN calculates the combined feature as:

$$z_x = [\beta_1 \cdot z_x^1; \beta_2 \cdot z_x^2; \beta_3 \cdot z_x^3], \tag{10}$$

where $\beta = (\beta_1, \beta_2, \beta_3)$ denote the weight scores of the encoded features of the three transfer functions z_x^1, z_x^2 and z_x^3. A Jensen-Shannon divergence is adopted to guarantee the learned features processing different semantic information:

$$L_{js}(z_x^1, z_x^2, z_x^3) = f\left(\overline{z_x^1}, \overline{z_x^2}\right) + f\left(\overline{z_x^1}, \overline{z_x^3}\right) + f\left(\overline{z_x^2}, \overline{z_x^2}\right), \tag{11}$$

where f(·) denotes the reciprocal of Jensen-Shannon divergence. The overall function of the Emsemble GAN is written as $L_{gan} + \eta_3 L_{js}$. The selection network is designed to infer the weight score β during testing.

5 Conclusion

Person re-ID is a cross-camera retrieval task that aims to retrieve a person of interest captured by one camera from images captured by all cameras in the surveillance system. Style disparities exist due to the difference camera settings and GAN is effective to narrow the domain gap caused by style by transferring images from source domain to the style of target domain. Moreover, GAN can also be utilized to tackle the cross-dataset transfer task. The transferred images from annotated datasets can be used as supplementary training samples and alleviate the lack of annotated training data. GAN-based approaches have achieved state of the art results in person re-ID.

Funding Statement: The author did not receive any specific funding for this study.

Conflicts of Interest: There is no conflict of interest in reporting on this study.

References

[1] P. Isola, J. Y. Zhu, T. H. Zhou and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1125–1134. 2017.

[2] M. Y. Liu and O. Tuzel, "Coupled generative adversarial networks," arXiv preprint arXiv:1606.07536 (2016).

[3] J. Y. Zhu, T. Park, P. Isola and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 2223–2232. 2017.

[4] Z. Zhong, L. Zheng, Z. D. Zheng, S. Z. Li and Y. Yang, "Camera style adaptation for person re-identification," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 5157–5166. 2018.

[5] L. H. Wei, S. L. Zhang, W. Gao and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 79–88. 2018.

[6] J. W. Liu, Z. J. Zha, D. Chen, R. C. Hong and M. Wang, "Adaptive transfer network for cross-domain person re-identification," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 7202–7211, 2019.

[7] C. Liu, X. J. Chang and Y. D. Shen, "Unity style transfer for person re-identification," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 6887–6896. 2020.

[8] X. G. Pan, P. Luo, J. P. Shi and X. O. Tang, "Two at once: Enhancing learning and generalization capacities via ibn-net," in *Proc. of the European Conf. on Computer Vision*, pp. 464–479. 2018.

[9] T. Zhang, Y. Y. Tang, B. Fang, Z. W. Shang and X. Liu, "Face recognition under varying illumination using gradientfaces," *IEEE Transactions on Image Processing,* vol. 18, no. 11, pp. 2599–2606, 2009.

[10] L. Christian, T. Lucas, H. Ferenc, C. Jose, C. Andrew *et al,* "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 4681–4690, 2017.