ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240542

Night-to-Day: Unpaired Image-to-Image Translation for Nighttime Pedestrian Detection

Afnan Althoupety [a,b,*], **Li-Yun Wang** [a,**], **Wu-Chi Feng**[a,***] **and Banafsheh Rekabdar**[a,****]

[a]Portland State University, Portland, USA
[b]Taif University, Taif, KSA

Abstract. In this paper, we show that exploiting Generative Adversarial Networks (GANs) to transform nighttime images into daytime representation increases the robustness of pedestrian detection in low-light conditions. Our work aims at first learning the image translation to transfer the style from daytime images to nighttime images with unpaired GAN training. Second, we use our end-to-end trained GAN model to translate night images as a pre-processing step before feeding them into an object detector that is pre-trained on daytime images only. To demonstrate the effectiveness of our translation approach, we conducted experiments on two real-world pedestrian datasets using both one-stage and two-stage object detectors. Our results outperform the baseline in all experiments and show highly competitive detection performance compared with other GAN-based approaches while holding the most lightweight architecture. We believe that our approach is an effective pre-processing first step that helps in bridging the performance gap between day and night at no expense of re-training object detector networks with more night images.

1 Introduction

Pedestrian detection has been one of the significant research problems in computer vision applications. Accurately detecting pedestrians from scenes with variations of lighting or weather conditions and appearance completeness (e.g., with occluded pedestrians) is the main goal in many real-world applications, including video surveillance, self-driving vehicles, and robotics. Thanks to deep-learning-based object detectors such as Faster R-CNN [26], and FCOS [30], the recent pedestrian detection works achieve optimistic detection performance for occluded and small-size pedestrians [41, 24, 29, 15]; however, most existing pedestrian detection models strictly focus on pedestrian detectability improvement on the images with good illumination and weather conditions. The detectability of those models is degraded when images are captured under low-light and extreme weather (e.g., heavy rain or snow) conditions.

GANs are one type of generative model and are compelling for high-quality image synthesis. Having this, the researchers have exploited GANs as part of a pedestrian detection pipeline to enhance pedestrian detection performance [7, 21, 40, 39]. Guo et al. [7] proposed a domain-adaptive pedestrian detection framework that utilizes the GAN to conduct image domain adaptation between color and thermal images to improve detection performance on thermal images. Liu et al. [21] introduced an attribute-preserving GAN as a novel data augmentation scheme and applied it to increase detection performance for small-size and occluded pedestrians. Zhi et al. [40] also exploited the GAN as a data argumentation scheme that produces synthetic pedestrian images with variations of human poses, and their approach improves the detection performance by increasing the diversity of training data. In addition to pedestrian detection, the researchers apply the GAN to synthesize more diverse training images for classification performance enhancements [39] or leverage the GAN to enlighten dark input images to improve classification performance [12].

In this paper, inspired by the aforementioned work, we focus on leveraging a GAN that performs image-to-image translation operations for image domain adaptation tasks to improve the detection performance of object detectors. We introduce an unpaired GAN-based image translation framework to synthesize daytime-like images given nighttime input images for nighttime pedestrian detection by learning illumination mapping between day and night domains. We then propose an Absolute Mean Brightness Error (AMBE) loss to regularize this learning illumination mapping process by minimizing the absolute difference between the mean brightness of real nighttime and synthetic daytime-like images. The AMBE loss can enforce a generator in our proposed framework to generate non-over-enhanced synthetic images to benefit pedestrian detection. We show that the proposed GAN-based framework and our loss to bolster pedestrian detection performance via qualitative analysis and empirical experimentation. Our proposed framework and loss are generalizable across variations of distinct datasets and object detectors, including one-stage and two-stage detectors.

The main contributions of this paper can be summarized in four-fold:

- We propose an unpaired GAN-based framework that learns the mapping between nighttime and daytime domains to overcome the data bias when detecting pedestrians at night.
- We introduce a novel AMBE loss that serves as a refinement step to improve synthetic nighttime images and therefore enhance detection performance.
- We demonstrate the effectiveness of our framework through several qualitative and quantitative ablation studies.
- We compare our proposed GAN-based framework with the baseline and its GAN-based counterparts in the task of pedestrian detection at nighttime.

* Corresponding Author. Email: afnan2@pdx.edu.
** Corresponding Author. Email: liyuwang@pdx.edu.
*** Corresponding Author. Email: wuchi@pdx.edu.
**** Corresponding Author. Email: rekabdar@pdx.edu.

2 Related Work

Inspired by game theory, Goodfellow *et al.* introduced the first Generative Adversarial Networks (GAN) framework, where the two game players, a generative model and a discriminative model, outsmart one another as in a minimax two-player game [6]. The competitive results of GAN training has led to a widespread adaptation in numerous applications [25, 18, 3, 36, 5, 37, 28]. One of the vision-based applications that leverage GAN is image-to-image translation.

Pix2pix [11] is a well-known framework that brings GAN to image-to-image translation tasks using conditional GAN (cGAN) [23], a design that provides prior information, like a class label, for better generation process and training stability. Nonetheless, pix2pix training acquires paired samples, which is a main barrier to real-world adaptations. To remove the restriction of paired training, CycleGAN [42] was proposed as the first unpaired GAN framework based on a cycle-consistency loss. To elaborate, CycleGAN utilizes two generators and two discriminators and introduces a cycle-consistency loss such that one generator learns the mapping from source to target domains and the other generator learns the inverse mapping to ensure geometrical and structural preservation.The benefit of unpaired training has made CycleGAN a classical base model for many image-to-image translation frameworks [34, 33].

Concerning unpaired night-to-day translation, ToDayGAN [2] was developed based on the ComboGAN architecture [1], an extension of CycleGAN framework to allow *n* domain translation, for retrieval-based localization problem. Specifically, ToDayGAN alters nighttime images to daytime representation to boost the degraded performance in visual localization when query images come from different illumination conditions compared to reference images. Another recent work EnlightenGAN [12] was proposed to enhance low-light images. EnlightenGAN uses a generator and two discriminators, global and local, to perform low-light to normal-light image translation while preserving texture and structural similarities. Moreover, EnlightenGAN applies a simple yet effective attention map, the inverse of the luminance channel, to assist the generator in producing better-quality images.

Similar to the aforementioned work, our approach aims to close the gap between night and day domains utilizing GAN without the need for paired supervision. Motivated by EnlightenGAN [12], we adopt the one-path architecture and adjust it over the course of our experiments to suit our exact problem, minimizing the data bias in pedestrian detection. Concretely, our proposed framework is more lightweight and only consists of one generator and one discriminator, allowing fast training.

3 Proposed Approach

3.1 Overall Framework

We perform image-to-image translation using a lightweight GAN architecture and to allow for unpaired training; in other words, we have no requirements for aligned night/day training image pairs. As demonstrated in Fig. 1, the proposed GAN-based approach is composed of one generator, one discriminator, and three losses, namely adversarial loss, perceptual loss, and AMBE loss.

The generator takes a nighttime image and alters it so that it emulates the distribution of real daytime images. The discriminator takes a translated night image and a real day image to distinguish realness. The training procedure is performed in an alternating adversarial manner. On the one hand, the generator is trained to fool the discriminator by synthesizing realistic translated images, close to the target domain distribution. On the other hand, the discriminator is trained to differentiate between real and synthetically generated images.

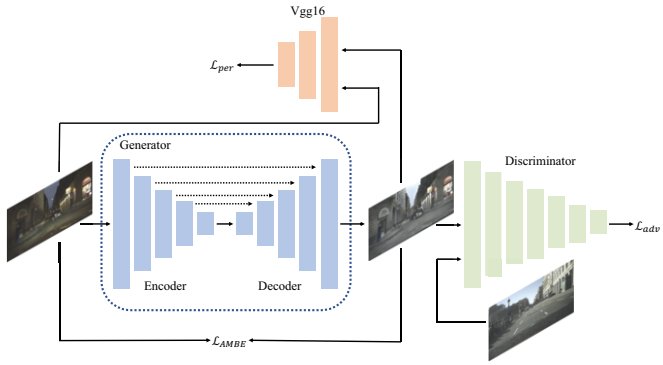

Figure 1. Overall proposed framework, consisting of a U-Net generator, a discriminator, and three loss functions: adversarial loss ($\mathcal{L}_{adv}$), perceptual loss ($\mathcal{L}_{per}$), and AMBE loss ($\mathcal{L}_{AMBE}$).The generator translates a night image to a fake day-like image and the discriminator differentiates between real and fake images.

In the following subsections, we describe the proposed framework, including the generator and the discriminator in terms of network architecture and loss functions.

3.2 Network Architecture

U-Net [27], a successful architecture on the task of semantic segmentation, is widely adapted in GANs as it enables extracting rich features from different depth layers and incorporates skip connections for better global coherence of translated images. Therefore, adapting U-Net helps the generator to produce high-quality synthesized images with semantic preservation.

Similar to [12], we use a U-Net generator that has 8 convolutional blocks. Each block consists of two 3 × 3 convolutional layers followed by a LeakyReLU [32] and a batch normalization [10]. Downsampling is achieved by MaxPooling while upsampling is implemented through bilinear interpolation to avoid checkerboard artifacts. Unlike [12], we consider only one discriminator and it consists of seven 4 × 4 convolutional layers followed by a LeakyReLU [32]. Please refer to the supplementary material in [12] for in-depth network architecture details and note that we do not include the local discriminator in our implementation.

3.3 Loss Function

Our proposed framework involves three loss functions as follows:

Adversarial Loss. The standard minimax adversarial loss, introduced by Goodfellow *et al.* in [6], is designed in an absolute manner to discriminate real from fake. As an optimization, Jolicoeur [14] replaced the standard discriminator with the Relativistic average Discriminator, denoted as D_{Ra}. The core idea is that the discriminator determines if an input is more real compared to all fake data in the mini-batch on average and similarly an input is fake compared to all real data in the mini-batch on average. Indeed, it is evident that considering relative determination has a positive impact on generated image quality and training stability [31, 17, 38]. The standard relativistic discriminator function is formulated as:

$$D_{Ra}(x_r, x_f) = \sigma(C(x_r) - \mathbb{E}_{x_f}[C(x_f)]) \quad (1)$$

$$D_{Ra}(x_f, x_r) = \sigma(C(x_f) - \mathbb{E}_{x_r}[C(x_r)]) \quad (2)$$

where x_r and x_f represent samples from real and fake distributions, respectively. C denotes the discriminator network, $\mathbb{E}_{x_f}[.]$ and $\mathbb{E}_{x_r}[.]$ represent the average for all fake, and real data in a mini-batch, respectively. Finally, σ is the sigmoid function.

Similar to [12], we utilize a modified version of the standard relativistic discriminator function, where the sigmoid function is replaced with Least Square GAN (LGAN) [22] to overcome the vanishing gradient problem. Considering both optimizations, the final loss functions for the discriminator D and the generator G are expressed as:

$$\mathcal{L}_D = \mathbb{E}_{x_r}[(D_{Ra}(x_r, x_f) - 1)^2] + \mathbb{E}_{x_f}[(D_{Ra}(x_f, x_r) - 0)^2] \quad (3)$$

$$\mathcal{L}_G = \mathbb{E}_{x_f}[(D_{Ra}(x_f, x_r) - 1)^2] + \mathbb{E}_{x_r}[(D_{Ra}(x_r, x_f) - 0)^2] \quad (4)$$

Having labels 0 and 1 for fake and real, respectively, both Eq. 3 and 4 are loss functions to be minimized. For example, the discriminator is trained to minimize the square difference between real data and label 1, and fake data and label 0. And for the generator, the target is to minimize the square difference between fake data and label 1, and real data and label 0.

Perceptual Loss. Johnson *et al.* [13] suggested *perceptual loss* for image transformation tasks to constrain structural similarity based on high-level features extracted from a pre-trained VGG network. Our framework employs this loss using a VGG-16 model pre-trained on ImageNet to measure the distance between original and translated night images in feature space. We use the activations of `relu5_1` layer from VGG-16. Perceptual loss is defined as:

$$\mathcal{L}_{per} = \|\phi(I) - \phi(G(I))\|_2^2 \quad (5)$$

where ϕ denotes the feature map, *I* denotes the original night image, and *G(I)* denotes the translated night image, by generator *G*.

AMBE Loss. To account for over brightness and noise amplification that unpaired training might cause, we leverage the image quality metric AMBE, which was introduced in [9] for mean brightness preservation. Indeed, incorporating AMBE loss can be a double-edged sword as it might prevent night-to-day translation. Thus, we use an empirically fine-tuned importance factor α in optimizing the total loss as we will show later in our ablation studies (Section 5.1). AMBE loss is formulated as:

$$\mathcal{L}_{AMBE} = \left|\mu_I - \mu_{G(I)}\right| \quad (6)$$

where μ_I and $\mu_{G(I)}$ are mean brightness for original and translated night images, by generator *G*, respectively.

The generator total loss function is defined as a weighted sum over these three loss terms:

$$\mathcal{L}_{total} = \mathcal{L}_{per} + \alpha\mathcal{L}_{AMBE} + \mathcal{L}_G \quad (7)$$

4 Experimental Setup

In this section, we describe implementation details of our experiments, including datasets, GAN training configurations, object detectors, and evaluation protocol.

4.1 Dataset Preparation

For clarity, we split our dataset preparation into two parts as follows:

For GAN training. We used two datasets separately: Berkeley DeepDrive (BDD100K) [35] and EuroCity Persons (ECP) [4]. BDD100K consists of 70k training and 10k validation images captured at different times of the day (with a resolution of 1280 × 720). ECP consists of 28k training and 5k validation images captured at different times of the day (with a resolution of 1920 × 1024). We sample 6032 night and 7372 day images from the BDD100K training set. Additionally, we sample 4221 night and 5220 day images from the ECP training set. It is noteworthy that we train for each dataset separately and sampling size varies due to the limited night images in ECP.

For object detection inference. On BDD100K, we utilize all night images containing at least a pedestrian from the validation set, resulting in 910 images with 3013 persons. Similarly for ECP, we use all night images containing at least a pedestrian from the validation set, resulting in 770 images with 4073 persons. Rather than running inference on testing real night images, we first translate them using the corresponding trained GAN model to emulate target daytime data distribution. Next, we feed translated night images into a pre-trained object detector that has seen only clear and day images.

4.2 Implementation Notes

To train our GAN-based framework, we adopt the Adam optimizer [16] with β_1 = 0.9 and β_2 = 0.999; the batch size is set to 16. We use a learning rate of 1e-4 for the first 100 epochs, followed by another 100 epochs with the learning rate linearly decayed to 0. We use crop size of 320 x 320 and no resizing. All training is performed using 4 Nvidia A40 GPUs. We trained other GAN-based methods using the same data and configurations for a fair comparison.

To evaluate pedestrian detection performance, we use PyTorch[1] pre-trained Faster R-CNN [26] and FCOS [30] object detectors, with default weights on MS-COCO dataset [19] and leveraging both ResNet-50 [8] and FPN [20]. We obtained evaluation metrics using COCO API[2], a publicly available evaluation protocol. It is worth noting that we threshold both confidence score and Intersection over Union (IoU) to be above 50%.

5 Experimentation and Discussion

We have conducted several experiments to fine-tune and validate the proposed night-to-day translation to elevate pedestrian detection robustness in dark settings. In this section, we first examine our design choices and the contribution of loss functions in our proposed framework through multiple ablation studies both quantitatively and qualitatively. Second, we fairly compare our approach with its GAN-based counterparts and the baseline (unprocessed real night images).

5.1 Ablation Studies

As stated earlier, we control imposing our AMBE loss by an importance factor. Here, we show how we empirically set α for training our framework on BDD100K and ECP datasets, respectively. As Table 1 shows, we examine three α values [0.05, 0.01, 0.15] and evaluate their performance for pedestrian detection on each dataset separately. Our empirical experiments indicate that the highest performance on BDD100k is when α = 0.10 while setting α = 0.15 gives the best performance for ECP. Thus, we empirically set α to be 0.10

[1] https://pytorch.org/vision/stable/models.html
[2] https://github.com/cocodataset/cocoapi

for BDD100K and 0.15 for ECP. Note that all results in the following sections are based on that α value selection.

Table 1. AMBE loss importance factor, α value, fine-tuning on BDD100K and ECP datasets by Faster R-CNN. Used performance metrics are average precision (AP) and average recall (AR).

Dataset	α value	AP (%)	AR (%)
BDD100K	0.05	41.10	23.97
	0.10	**41.27**	**24.10**
	0.15	41.12	23.73
ECP	0.05	56.81	36.37
	0.10	56.25	36.20
	0.15	**56.93**	**36.37**

In regards to loss functions, we have conducted an ablation study to measure the contribution of the two loss functions, perceptual loss and AMBE loss. Specifically, we test our framework performance in an incremental fashion: (i) without perceptual loss nor AMBE loss, (ii) with perceptual loss only, and (iii) with perceptual loss + AMBE loss.

Quantitatively speaking, Table 2 demonstrates that perceptual loss is a substantial component for detection performance across all metrics by both Faster R-CNN and FCOS. The improvement margin is at least 50% when incorporating perceptual loss. We attribute this to the one-path GAN design where no cycle consistency is implied. On top of that, it is challenging to guide the translation training in an unpaired setting, thus perceptual loss plays a key role in assisting the generator to produce structurally similar translated images compared to the original ones. To provide more guidance and refinement for the generator, AMBE loss supports unpaired training and balances the generation process with respect to brightness. Numerically, AMBE loss alleviates AP and AR for both Faster R-CNN and FCOS.

Table 2. Quantitative ablation study of proposed framework loss functions: without perceptual loss (w/o $\mathcal{L}_{per}$), with perceptual loss (w/ $\mathcal{L}_{per}$) and with perceptual and AMBE losses (w/ $\mathcal{L}_{AMBE}$), on BDD100K dataset by Faster R-CNN and FCOS. Used performance metrics are average precision (AP) and average recall (AR).

Detector	Metric (%)	w/o $\mathcal{L}_{per}$	w/ $\mathcal{L}_{per}$	w/ $\mathcal{L}_{AMBE}$
Faster R-CNN	AP	18.47	41.21	**41.27**
	AR	12.52	23.95	**24.10**
FCOS	AP	15.30	36.01	**36.08**
	AR	10.22	21.32	**21.37**

Qualitatively speaking, the first column in Fig. 2 affirms the importance of perceptual loss in maintaining texture and structure similarity. All images generated without perceptual loss suffer from distortion and unrealism. Looking at the second and third columns in Fig. 2, red dotted boxes illustrate the visual impact of AMBE loss. Although the good-quality image translation using perceptual loss only, we observe that the generator tends to alter night images with a haze-like effect and sometimes over-brightness if there are many streetlights, as shown in the first example in Fig. 2 (first row from bottom). AMBE loss helps in those regions and improves the translation quality.

5.2 Comparison with Other Approaches

In this section, we showcase the effectiveness of our proposed approach by evaluating it against the baseline and other GAN-based approaches in the context of nighttime pedestrian detection. The evaluation includes two real-world datasets, BDD100K and ECP, and two state-of-the-art object detectors, Faster R-CNN and FCOS.

Table 3. The table presents a detection performance comparison of the baseline, our approach and other GAN-based approaches on BDD100K dataset by Faster R-CNN and FCOS. Used performance metrics are average precision (AP) and average recall (AR). The **bold blue** denotes best performance and **bold black** denotes second best per detector.

Detector	Metric (%)	Baseline	CycleGAN	ToDayGAN	EnlightenGAN	Proposed
Faster R-CNN	AP	40.43	30.08	35.81	**40.75**	**41.27**
	AR	23.45	18.51	20.43	**23.64**	**24.10**
FCOS	AP	34.59	23.22	29.58	**35.44**	**36.08**
	AR	20.74	14.56	17.68	**21.10**	**21.37**

As Table 3 and 5 depict, our results outperform the baseline among all metrics on both BDD100K and ECP. In particular, AP is higher by 0.84 and 1.49% for Faster R-CNN and FCOS, respectively on BDD100K. Likewise on ECP, our approach elevates AP by 2.05 and 0.72% for Faster R-CNN and FCOS, respectively. Compared to other GAN-based approaches, we achieve the best performance on BDD100K in terms of AP and AR. Moreover, our performance on ECP is the second best by FCOS. However, we outperform EnlightenGAN AP by an improvement margin of 0.62% for Faster R-CNN on ECP with our more lightweight version. Additionally, Table 4 affirms the simplicity of our architecture compared to other GAN-based methods. In other words, the proposed approach achieves better detection performance with fewer parameters. Having a smaller model size is beneficial in deployment, avoiding over-fitting, extra computation, and memory requirements.

Table 4. The table presents model size comparison in terms of number of parameters between our approach and other GAN-based approaches.

	CycleGAN	ToDayGAN	EnlightenGAN	Proposed
Parameters	28.298 M	56.726 M	26.750 M	19.791 M

Table 5. The table presents a detection performance comparison of the baseline, our approach and other GAN-based approaches on ECP dataset by Faster R-CNN and FCOS. Used performance metrics are average precision (AP) and average recall (AR). The **bold blue** denotes best performance and **bold black** denotes second best per detector.

Detector	Metric (%)	Baseline	CycleGAN	ToDayGAN	EnlightenGAN	Proposed
Faster R-CNN	AP	54.88	54.15	51.79	**56.31**	**56.93**
	AR	35.24	34.98	32.95	**36.65**	**36.37**
FCOS	AP	45.40	43.19	43.61	**46.13**	**46.12**
	AR	31.02	29.60	29.30	**31.61**	**31.36**

In terms of visual inspection, Fig. 3 shows two examples to compare the detection performance of our framework with other approaches on BDD100K. CycleGAN suffers from noise and content distortion and ToDayGAN generally fails to enhance low-light conditions. EnlightenGAN tends to overly enhance brightness which either causes a wrong detection, false positive, or a miss detection, false negative. Conversely, our approach moderately emulates the daytime domain such that both Faster R-CNN and FCOS can detect night pedestrians robustly. Similarly, Fig. 4 shows two examples to compare detection performance on ECP. We successfully translate night images so that both detectors can distinguish pedestrians from the dark background. CycleGAN and ToDayGAN introduce false positives resulting in poor detection performance which indicates that one-path GAN methods, EnlightenGAN and our proposed approach, are more effective than cycle-based GAN methods, CycleGAN and ToDayGAN.

Overall, the two-stage object detector, Faster R-CNN, detects dark pedestrians more robustly compared to the one-stage detector, FCOS.

Figure 2. Qualitative ablation study of loss functions in our framework. Each row represents an example, from BDD100K dataset, and each column represents: without perceptual loss (w/o $\mathcal{L}_{per}$), with perceptual loss (w/ $\mathcal{L}_{per}$), with perceptural and AMBE losses (w/ $\mathcal{L}_{AMBE}$), respectively. The dotted red boxes denote refined regions by AMBE loss.

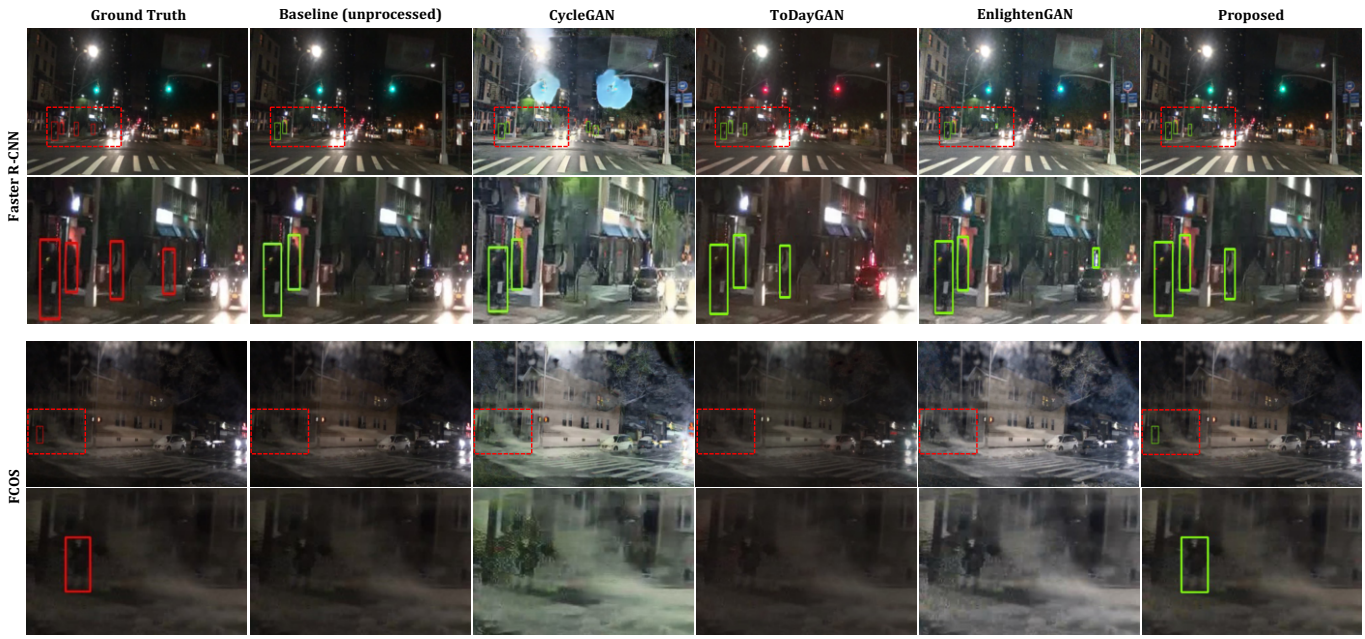

Figure 3. Visual comparison of detection performance, including the ground truth, the baseline, our approach, and other GAN-based approaches on BDD100K dataset by Faster R-CNN and FCOS. For each detector: top row represents detections on full images where the red boxes are ground truth, green boxes are detected boxes, and dotted red boxes denote the region of interest, whereas bottom row represents the zoom-in region of interest to illustrate detection performance.

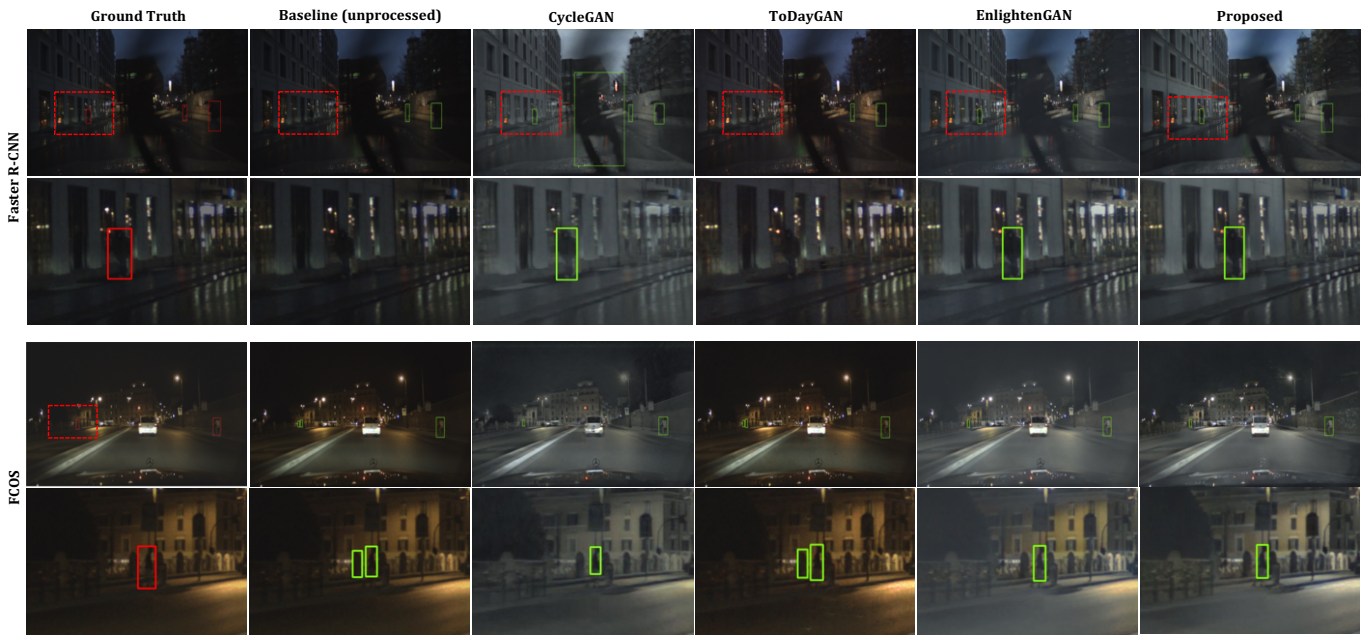

Figure 4. Visual comparison of detection performance, including the ground truth, the baseline, our approach, and other GAN-based approaches on ECP dataset by Faster R-CNN and FCOS. For each detector: top row represents detections on full images where the red boxes are ground truth, green boxes are detected boxes, and dotted red boxes denote the region of interest, whereas bottom row represents the zoom-in region of interest to illustrate detection performance.

Additionally, one-path GAN architectures benefiting from perceptual loss to constrain structural similarity in feature space give remarkably higher detection performance in comparision with cycle-based GAN architectures that ensure structural similarity by learning the bidirectional mapping between source and target domains. Moreover, all experimental results assure the effectiveness of our approach in addressing the dark pedestrian detection issue with a simple yet effective GAN architecture.

6 Conclusion

This paper presents a GAN-based approach to discover the underlying relationship between night and day domains and resolve the domain shift problem in pedestrian detection. In other words, we benefit from GAN's ability to learn the transformation between night and day images to mitigate the degraded performance of nighttime pedestrian detection. The evaluation results show that our approach outperforms the baseline on two real-world nighttime datasets and generalizes well to one-stage and two-stage object detectors. Plus, our comparison results with other GAN-based methods reveal that our approach is competitive while being the most lightweight. We believe that night-to-day image translation through GAN is a practical domain adaptation solution for pedestrian detection task.

References

[1] A. Anoosheh, E. Agustsson, R. Timofte, and L. Van Gool. Combogan: Unrestrained scalability for image domain translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 783–790, 2018.

[2] A. Anoosheh, T. Sattler, R. Timofte, M. Pollefeys, and L. Van Gool. Night-to-day image translation for retrieval-based localization. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5958–5964. IEEE, 2019.

[3] G. Antipov, M. Baccouche, and J.-L. Dugelay. Face aging with conditional generative adversarial networks. In *2017 IEEE international conference on image processing (ICIP)*, pages 2089–2093. IEEE, 2017.

[4] M. Braun, S. Krebs, F. Flohr, and D. M. Gavrila. Eurocity persons: A novel benchmark for person detection in traffic scenes. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1844–1861, 2019.

[5] Y. Chen, Y.-K. Lai, and Y.-J. Liu. Cartoongan: Generative adversarial networks for photo cartoonization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9465–9474, 2018.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

[7] T. Guo, C. P. Huynh, and M. Solh. Domain-adaptive pedestrian detection in thermal images. In *2019 IEEE international conference on image processing (ICIP)*, pages 1660–1664. IEEE, 2019.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] S.-C. Huang and C.-H. Yeh. Image contrast enhancement for preserving mean brightness without losing image features. *Engineering Applications of Artificial Intelligence*, 26(5-6):1487–1492, 2013.

[10] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

[11] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[12] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE transactions on image processing*, 30:2340–2349, 2021.

[13] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016.

[14] A. Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018.

[15] J. U. Kim, S. Park, and Y. M. Ro. Robust small-scale pedestrian detection with cued recall via memory learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3050–3059, 2021.

[16] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[17] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8878–8887, 2019.

[18] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[20] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[21] S. Liu, H. Guo, J.-G. Hu, X. Zhao, C. Zhao, T. Wang, Y. Zhu, J. Wang, and M. Tang. A novel data augmentation scheme for pedestrian detection with attribute preserving gan. *Neurocomputing*, 401:123–132, 2020.

[22] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

[23] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[24] Y. Pang, J. Xie, M. H. Khan, R. M. Anwer, F. S. Khan, and L. Shao. Mask-guided attention network for occluded pedestrian detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4967–4975, 2019.

[25] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016.

[26] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[27] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[28] K. K. Singh, U. Ojha, and Y. J. Lee. Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6490–6499, 2019.

[29] T. Song, L. Sun, D. Xie, H. Sun, and S. Pu. Small-scale pedestrian detection based on topological line localization and temporal feature aggregation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 536–551, 2018.

[30] Z. Tian, C. Shen, H. Chen, and T. He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019.

[31] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018.

[32] B. Xu, N. Wang, T. Chen, and M. Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

[33] S. Yang, M. Sun, X. Lou, H. Yang, and H. Zhou. An unpaired thermal infrared image translation method using gma-cyclegan. *Remote Sensing*, 15(3):663, 2023.

[34] Z. Yi, H. Zhang, P. Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017.

[35] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020.

[36] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017.

[37] H. Zhang, V. Sindagi, and V. M. Patel. Image de-raining using a conditional generative adversarial network. *IEEE transactions on circuits and systems for video technology*, 30(11):3943–3956, 2019.

[38] K. Zhang, W. Luo, Y. Zhong, L. Ma, B. Stenger, W. Liu, and H. Li. Deblurring by realistic blurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2737–2746, 2020.

[39] X. Zhang, Z. Wang, D. Liu, Q. Lin, and Q. Ling. Deep adversarial data augmentation for extremely low data regimes. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(1):15–28, 2020.

[40] R. Zhi, Z. Guo, W. Zhang, B. Wang, V. Kaiser, J. Wiederer, and F. B. Flohr. Pose-guided person image synthesis for data augmentation in pedestrian detection. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, pages 1493–1500. IEEE, 2021.

[41] C. Zhou and J. Yuan. Bi-box regression for pedestrian detection and occlusion estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 135–151, 2018.

[42] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.