SinForkGAN: ForkGAN with Single Rainy Night Images

Seri Lee[1 [2021−26978]]

sally20921@snu.ac.kr
Seoul National University,
Seoul, Republic of Korea

Abstract. *ForkGAN* has been proposed as a task-agnostic image translation method that can boost the performance of multiple vision tasks in adverse weather conditions. Although ForkGAN achieved remarkable image translation quality without any downstream task-awareness in an unsupervised way, the two complicated 'Night to Day' and 'Day to Night' image translation module requires some kind of division between the training images to night image and day image. In this paper, we show that we can actually do away with the 'Day to Night' translation module and only train ForkGAN with modified 'Night to Day' translation module using nighttime images only. We accomplish this by incorporating the recently proposed *RumiGAN framework* into the ForkGAN architecture without compromising its performance. Extensive experimental results on nighttime datasets show that our algorithm produces on par or sometimes even better results on image localization/retrieval, semantic segmentation, and object detection tasks compared to ForkGAN and other state-of-the-art methods. Code will be available at `https://github.com/sally20921/SinForkGAN`.

Keywords: Unsupervised Learning, Image-to-Image Translation, Low-Light Image Enhancement, Generative Adversarial Networks

1 Introduction

Rainy night represents one of the most challenging yet very probable case of data bias in real life scenario. Many vision approaches perform poorly in this case. Take object detection for example. An object detector trained on a day time dataset suffers 30-50 percent accuracy drop on rainy night images [16]. Considering the fact that huge domain change is quite common for computer vision tasks, such as day and night change or change in weather conditions, it is quite clear that we need algorithms that can work well on any situations.

In this paper, we opt for an approach explicitly designed to translate the adverse weather conditions (i.e. symbolized as 'rainy night') to a standard weather condition (i.e. bright daytime). Of course each existing algorithm can be optimized in a task-specific way, but we ask ourselves, "If we could design a module that

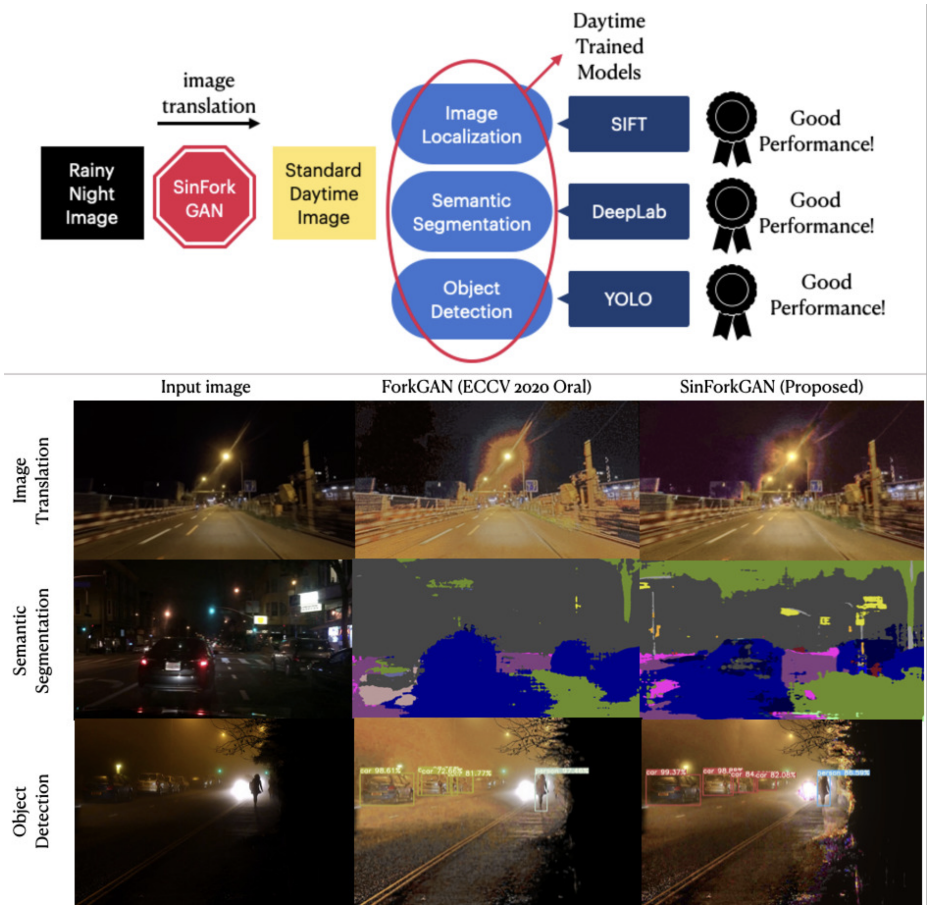

Fig. 1. SinForkGAN improves upon ForkGAN [16] model to boost the performance for multiple vision tasks.

we could plug in before any vision task, wouldn't that be more convenient?" We decide that this module we fix upon has to meet the following three criteria. First, since it is oftentimes impossible to get precisely aligned ground-truth image pairs captured at different times, (because of dynamic scene changes), the algorithm has to be learned in an unsupervised way without any explicit supervision. Second, we translate the image not in a way that might look good to humans, but in a way that could improve the subsequent computer vision task in adverse weather conditions. (In this case, we don't know what the task is going to be.) In other words, we create a task agnostic module that we could plug in before any vision task. Third, the model has work well on real-world datasets. Considering the fact that no amount of adequate data exists for adverse weather conditions, some algorithms exist that is trained on and works well on synthetic

dataset. However, to truly achieve its goal, the model would have to operate well on real-world circumstances.

We find few previous works that meets our objective. The recently proposed *ForkGAN* [16] share many similarities with what we are trying to bring about. Zheng aet al. has successfully decoupled domain-invariant content and domain-specific style by utilizing a fork-shaped cyclic generative module and therefore showed remarkable performance on various vision tasks. To the best of our knowledge, we are the first work that directly improves upon ForkGAN. We realize this in 4 different ways:

- ForkGAN uses unpaired training samples to learn the ampping between different image domains. Although this is a huge gain compared to deep networks requiring aligned image pairs, using 2 images at a time still requires some sort of division between nighttime and daytime images. We train with only single nighttime image at a time.
- We propose a unique adversarial loss L_{adv} inspired by [5]. The modified *RumiGAN framework* allows the network to differentiate between positive and negative examples.
- While using only nighttime images at training and testing time, the method does not compromise its performance. We prove this by proceeding extensive experiments on various vision tasks including image localization/retrieval, semantic segmentation and object detection.
- SinForkGAN differs from single image dehazing methods in that it is trained and tested on real-world datasets. Unsupervised single image dehazing methods tend to fail under real-world circumstances where noises are different from synthetic dataset. Our problem setting (i.e. rainy night) is much more challenging than simple image denoising or dehazing.

2 Related Works

2.1 Deep Low-Light Image Enhancement: Paired to Unpaired

Low light image restoration models aim to improve the visual quality of under-exposed photos by manipulating color, brightness and contrast of the image. In the past, deep learning solutions mostly relied on paired training, where most low-light images are synthesized from normal images. RetinexNet [14] is a famous example.

Inspired by Cycle-GAN [17], EnlightenGAN [8] was proposed to enhance low-light images without paired training data. However, prior knowledge on whether the input image is *too dark* or *too bright* is required. Furthermore, it mainly focuses on improving subjective visual quality, rather than facilitating subsequent high-level computer vision task. Recent works like ForkGAN [16] explicitly translate the whole image to daytime and introduce multiple deocders.

2.2 Unsupervised Single Image Dehazing

Low illumination can be viewed as a common kind of visual distortion. In this way, our method can be related to unsupervised single image dehazing methods. Self-supervised denoising models, including Noise2Void [10], Noise2self [6] were proposed to train the network only with one noisy observation per scene. However, the relatively low accuracy and heavy computation greatly limit the application. Methods with noise model assumptions degrade sharply when dealing with real-world noisy images where the noise distribution remains unknown and are mainly designed for human vision rather than machine vision.

2.3 *RumiGAN*: GAN Inspired by Contrastive Self-Supervised Learning

Self-supervised representation learning involves splitting the data into positive and negative samples for discriminative learning. The relative distances between samples are used to train a neural network. The contrastive loss compares pairs of samples, and assigns positive weights to similar pairs and negative weights to dissimilar ones. Relying on the visual information presented on the training data, such methods can be applied in label-free tasks.

Inspired by the recent success on contrastive self-supervised learning, *RumiGAN* [5] provides GAN not only positive data it must learn to model but also negative samples it must learn to avoid. This formulation allows the discriminator to represent the underlying target distribution better and accelerates the learning process of the generator.

In the standard GAN formulation [7], the generator transforms the input noise $z \sim p_Z$ to the output $G(z)$ with distribution $p_g(x)/$ The target data is sampled from an underlying distribution $p_d(x)$. The discriminator $D(x)$ predicts the probability of its input coming from p_d. The RumiGAN loss comprises of three terms: the expected cross-entropy between $[1,0]^T$ and $[D(x), 1-D(x)]^T$ for the positive data; $[0,1]^T$ and $[D(x), 1-D(x)]^T$ for the negative generator samples; and $[0,1]^T$ and $[D(x), 1-D(x)]^T$ for samples drawn from the negative class. The discriminator loss is given as

$$L = -(\alpha^+ \mathbb{E}_{x\sim p_d^+}[\log D(x)] + \mathbb{E}_{x\sim p_g}[\log(1-D(x))] + \alpha^- \mathbb{E}_{x\sim p_d^-}[\log(1-D(x))]) \quad (1)$$

where α^+ and α^- are weights attached to the losses, $\alpha^- \geq \alpha^+ - 1$ and $\alpha^+ \in [0,1]$ for optimization.

3 Proposed Method

In this section, we briefly explain the previous approach, the ForkGAN model [16]. Then, overview and the details of the proposed method are described.

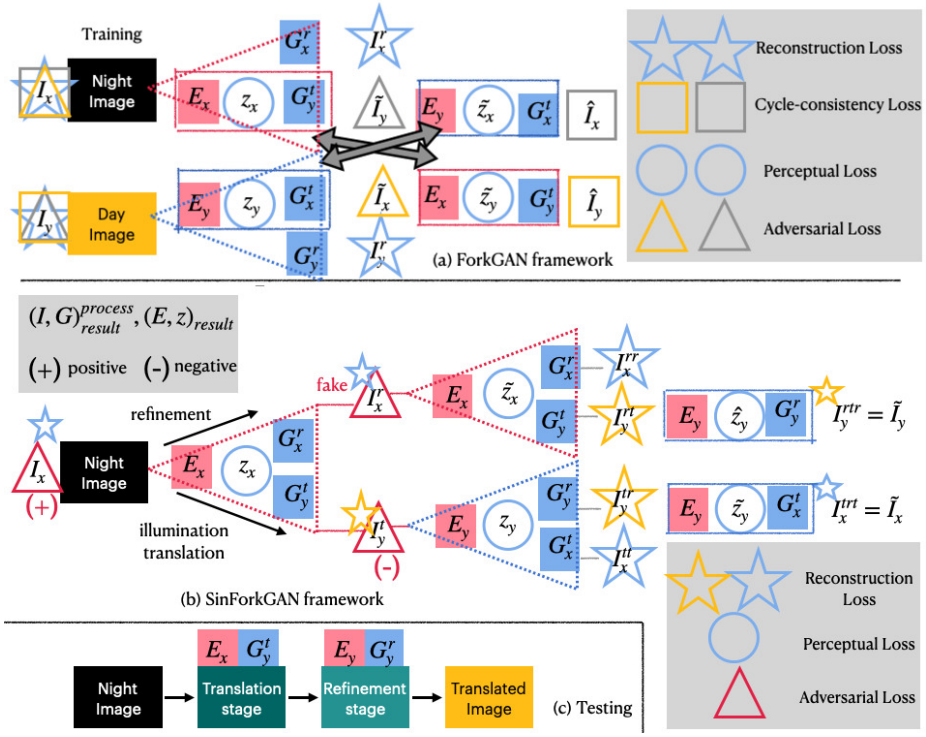

Fig. 2. The framework of ForkGAN and SinForkGAN. I_x and I_y denotes random image from night domain $\mathbb{X}$ and day domain $\mathbb{Y}$. E is the encoder, G_x^t and G_y^t are responsible for achieving domain translation. G_x^r and G_y^r aim to reconstruct the input images.

3.1 Revisiting the ForkGAN Model

There are two separate modules, night-to-day and day-to-night translation module in ForkGAN, each containing one encoder and two decoders. Images are first fed into each encoder and then domain-invariant representation z is obtained. The two decoders (reconstruction decoder and translation decoder) both have the same z as input.

A pixel-level L_1-norm based reconstruction loss is calculated between the original image and the reconstructed image. The adversarial loss, same as the one in Cycle-GAN [17] is computed which aims to distinguish the random real image and the translated image.

Then the encoder once again extracts the domain-invariant feature $\tilde{z}$ from the translated image (this time the encoder is exchanged between the two domains). A perceptual loss between $\tilde{z}$ and z is performed.

Finally, we obtain the original-like translated-translated image using the translation decoder (this time the decoder is exchanged). The cycle-consistency loss is computed between the original and final translated-translated image.

3.2 SinForkGAN with Only Single Rainy Night Images

There is a distinct limitation when training ForkGAN in real-life scenarios. ForkGAN uses unpaired training samples to learn mapping between two different image domains. This requires prior knowledge on the training dataset to tell whether the input image is dark or bright. Not only is this a cumbersome process but also it drifts apart from the true meaning of 'unsupervised'. To avoid this kind of division, we adjust the designs of the ForkGAN model. The main idea is getting rid of 'Day to night' translation module and introducing a new adversarial loss L_{adv} modeling the *RumiGAN* [5] framework.

First, we experiment with a reduced baseline version of SinForkGAN framework which contains only one fork-shape generator. This reduced baseline model results in a still dark image. We speculate that the reason why it shows poor performance is the lack of constraints erased by getting rid of the 'Day to night' module. Consequently, we introduce a new adversarial loss, and also modify the perceptual loss and reconstruction loss to act as stronger constraints that could make up for the eliminated part. The improved loss function can be expressed as:

$$L(E, G^r, G^t) = L_{adv} + \beta L_{per} + \gamma L_{rec} \tag{2}$$

where β and γ are loss weights. With the total loss, the three components E, G^r, G^t are optimized together and we set $\beta = \gamma = 1$ in our experiments. During inference time, SinForkGAN takes an night image as input and translates it to a more refined output by going through translation and refinement stage. Now we explain the details of each part of the loss function.

Adversarial Loss

We apply the *Rumi framework* to the adversarial loss. The adversarial loss L_{adv} aims to distinguish between the real night image I_x and the reconstructed night image while penalizing the translated day image.

$$L_{adv} = -(\alpha^+ \mathbb{E}[\log D(I_x)] + \mathbb{E}[\log(1 - D(I_x^r)] + \alpha^- \mathbb{E}[\log(1 - D(I_y^t))]) \tag{3}$$

We set $alpha^+ = 0.2$ and $\alpha^- = -0.8$ for experiment.

Perceptual Loss

We impose a perceptual loss as

$$L_{per} = \mathbb{E}_{x,y\sim U}(\sum_{n=1}^{N} \|\Phi_n(x) - \Phi_n(y)\|_1) \tag{4}$$

, where $U = \{z_x, \tilde{z_x}, z_y, \tilde{z_y}, \hat{z_y}\}$. which makes elements in U perceptually similar to each other. Here, Φ_n denotes the feature extractor at the n_{th} level of the pretrained VGG-19 network on ImageNet. Different from the way perceptual loss is typically used (feeding image data to the VGG network), we rearrange the feature maps of elements in U through bilinear interpolation to fit into only the last three layers of VGG as stated in [16].

Reconstruction Loss

We perform a pixel-level l_1-norm based reconstruction loss L_{rec} between elements in W and V.

$$L_{rec} = (\sum_{w \in W} \sum_{v \in V} \|w - v\|_1) \tag{5}$$

where $W = \{I_x, I_x^r, I_x^{rr}, I_x^{tt}, \tilde{I}_x\}$ and $V = \{I_y^t, I_y^{rt}, I_y^{tr}, \tilde{I}_y\}$. In the preliminary experiments, we tried replacing the l_1-norm based reconstruction loss by creating another adversarial loss with multiple negatives, but did not observe improved performance.

3.3 Implementation and Experiments

3.4 Training details

We follow most of the training details proposed in [16]. The encoder E contains 3 Conv-Ins-ReLU modules and 4 dilated residual blocks, while both reconstructed decoder G^r and the translated decoder G^t have 4 dilated residual blocks and 3 Deconv-Ins-ReLU modules followed by a Tanh activity function. All the domain-specific discriminators adopt the multi-scale discriminaotr archietcture and we set the number of scales as 2. We adopt Adam optimizer and set learning rate to 0.0002.

3.5 Datasets

- **Dark Zurich Dataset** [13] contains 2,416 nighttime images along with the respective coordinates of the camera for each image used to construct cross-time-of-day correspondences. Since it contains nighttime images and its corresponding daytime images, we use it to test SIFT feature matching.
- **RaidaR** [9] is a rich annonated dataset of rainy street scenes, and it provides 5,000 images and corresponding color-coded labels for semantic segmentation task. We use this dataset to test if our model works on rainy images as well as nighttime images.
- **BDD100K** [15] is a large scale high-resolution autonomous driving dataset, which collected 100,000 video clips in multiple cities and under various conditions. It provides 27,971 night images and we use it for training the SinForkGAN model. It also provides 137 night images and corresponding segmentation ground truth for evaluation.

– **ExDark** [11] provides 7,363 low-light images from very low-light environments to twilight with 12 object classes annotated on local bounding boxes. We use it for object detection task.

4 Results

4.1 Localization by SIFT Point Matching

We follow the basic feature matching pipeline as stated in the OpenCV document [1]:

1. Detect keypoints using the SIFT detector, compute the descriptors.
2. Match descriptor vectors with a BF-based matcher.
3. Filter matches using the Lowe's ratio test [12]($ratio\_threshold = 0.7$).
4. Draw matches.

Figure 3 shows the image translation result compared to ForkGAN. As you can see, SinForkGAN produces more precise night-to-day image translation. There is a huge amount of difference in the number of SIFT point matching points before and after applying SinForkGAN to nighttime images.

4.2 Semantic Segmentation

Figure 4 and Figure 5 presents the translated results and corresponding segmentation of various methods including our SinForkGAN. We adopt state-of-the art method ToDayGAN [4], our baseline model ForkGAN and compare them with SinForkGAN to better understand the performance of SinForkGAN. We use a pretrained DeepLab-V3 [2] model on Cityscapes dataset which does not contain any nighttime image. Therefore, according to our experiment, when the network is fed with nighttime images, the performance of semantic segmentation task suffers considerably. Again, Night-to-day image translation models prove to be a powerful tool in improving the segmentation performance. SinForkGAN preserves better detailed information than ToDayGAN or ForkGAN such as small traffic signs.

4.3 Object Detection

For object detection task, we once again compare our SinForkGAN with the most related ForkGAN and state-of-the-art method ToDayGAN on ExDark dataset. We use a pretrained YOLOV3-tiny [3] model on PASCAL VOC 2007 and PASCAL VOC 2012 dataset. The author of YOLOV3 [3] notes that you can easily trade off between speed and accuracy by changing the size of the model. Therefore, for our purposes, we choose the YOLOV3-tiny model, whose performance relatively suffers compared to other standard models, when fed with nighttime input images. Figure 6 shows that our method can boost the performance of object detection by preserving and enhancing detailed information such as pedestrians.

4.4 Ablation Study

Several experiments were conducted before finalizing our SinForkGAN design. Firstly, we created a baseline model where we only kept what we think is the absolute necessity to to constrain the SinForkGAN model. We visually saw that the results generated by the baseline model does not perform image translation adequately enough. Then, we tried to put more constraints by adding another fork-shape generator and leveraging all the loss proposed in the original ForkGAN paper. As shown in Fig 7, this architecture improves upon the baseline but still does not work well enough for subsequent downstream tasks. Finally, we created 3 fork-shape generators and reached the final SinForkGAN model. We also evaluated a twisted verison of SinForkGAN by leveraging multiple negatives. This did not lead to improved performance, according to our experiment.

5 Conclusion and Future Work

We propose a novel framework SinForkGAN to achieve night-to-day image translation without any supervision using real-world single nighttime images. Visually, we have shown that SinForkGAN works well on real-world nighttime rainy images. However, we did not cover any quantitative result in our experiment, which could be pointed out as limitation of this paper. We plan to continue our work and conduct quantitative study as well. But for now, we leave this as future work and conclude this paper. The significance of SinForkGAN model is that we can leverage existing daytime computer models. Possible future work include a multi-task learning network to share the backbone of different vision tasks.

References

1. https://www.docs.opencv.org/master/dc/dc3/tutorial_py_matcher.html
2. https://github.com/fregu856/deeplabv3
3. https://github.com/Lornatang/YOLOv3-PyTorch
4. Anoosheh, A., Sattler, T., Timofte, R., Pollefeys, M., Gool, L.V.: Night-to-day image translation for retrieval-based localization (2019)
5. Asokan, S., Seelamantula, C.S.: Teaching a gan what not to learn. arXiv preprint arXiv:2010.15639 (2020)
6. Batson, J., Royer, L.: Noise2self: Blind denoising by self-supervision. In: International Conference on Machine Learning. pp. 524–533. PMLR (2019)
7. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. arXiv preprint arXiv:1406.2661 (2014)
8. Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., Yang, J., Zhou, P., Wang, Z.: Enlightengan: Deep light enhancement without paired supervision. IEEE Transactions on Image Processing (2021)
9. Jin, J., Fatemi, A., Lira, W., Yu, F., Leng, B., Ma, R., Mahdavi-Amiri, A., Zhang, H.: Raidar: A rich annotated image dataset of rainy street scenes (2021)

10. Krull, A., Buchholz, T.O., Jug, F.: Noise2void-learning denoising from single noisy images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2129–2137 (2019)
11. Loh, Y.P., Chan, C.S.: Getting to know low-light images with the exclusively dark dataset (2018)
12. LoweDavid, G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision (2004)
13. Sakaridis, C., Dai, D., Gool, L.V.: Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation (2019)
14. Wei, C., Wang, W., Yang, W., Liu, J.: Deep retinex decomposition for low-light enhancement. arXiv preprint arXiv:1808.04560 (2018)
15. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning (2020)
16. Zheng, Z., Wu, Y., Han, X., Shi, J.: Forkgan: Seeing into the rainy night (2020)
17. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)

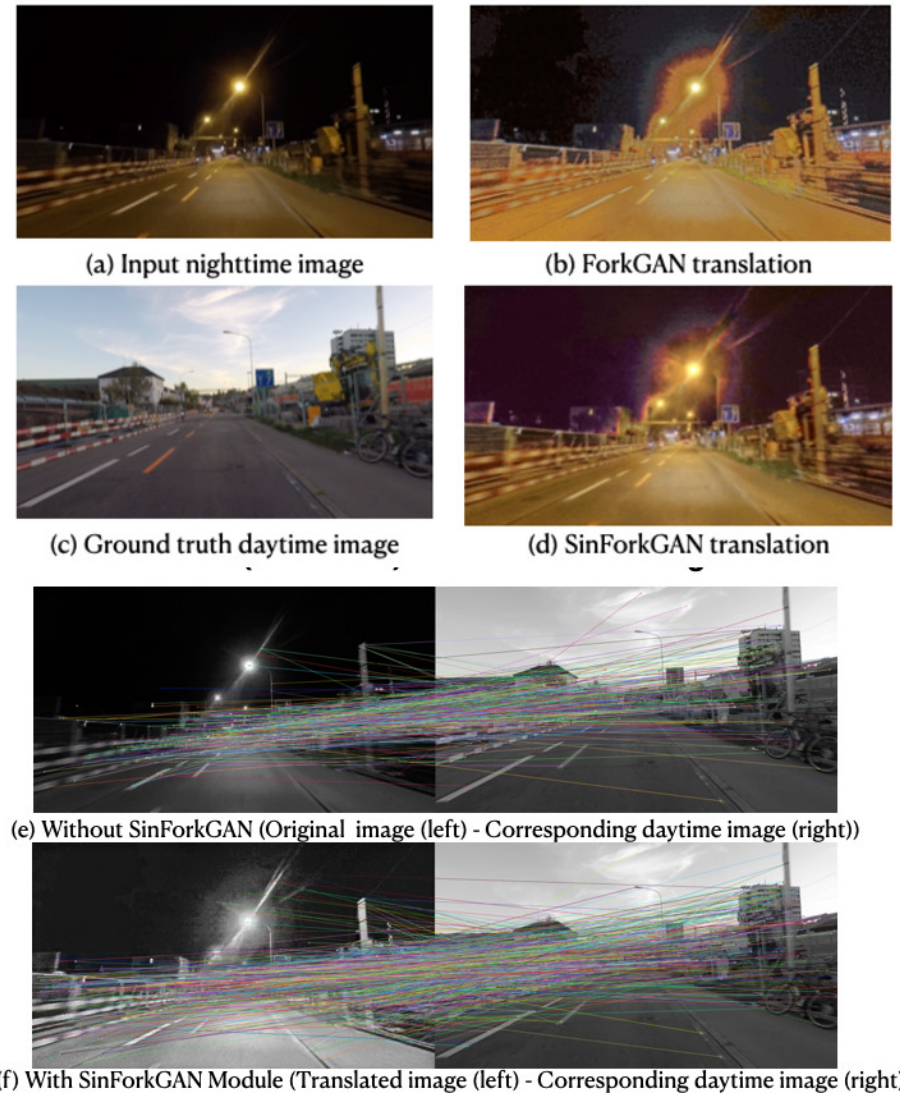

Fig. 3. SinForkGAN translation result compared to ForkGAN, and the last two pictures are the visualization SIFT interest point matching before and after applying SinForkGAN.

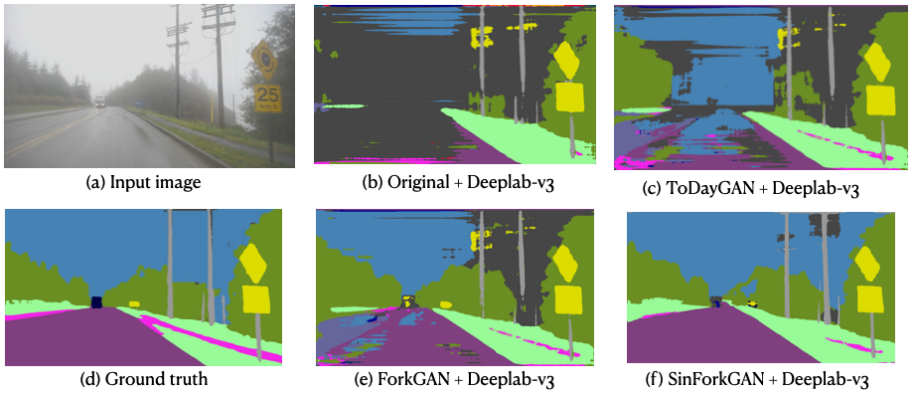

Fig. 4. The visualization of semantic segmentation performance of different methods, tested on RaidaR dataset.

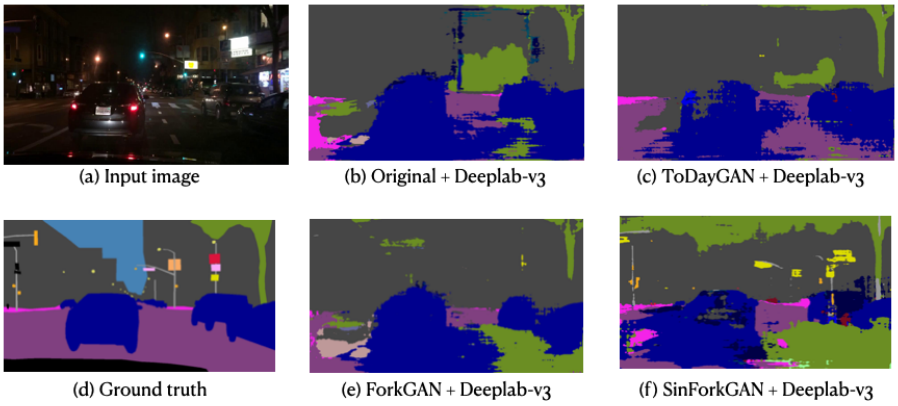

Fig. 5. The visualization of semantic segmentation tested on BDD100K dataset, comparing the performance of different methods.

Fig. 6. Visual comparison of detection results on ExDark dataset. ForkGAN++ improves the detection of small objects.

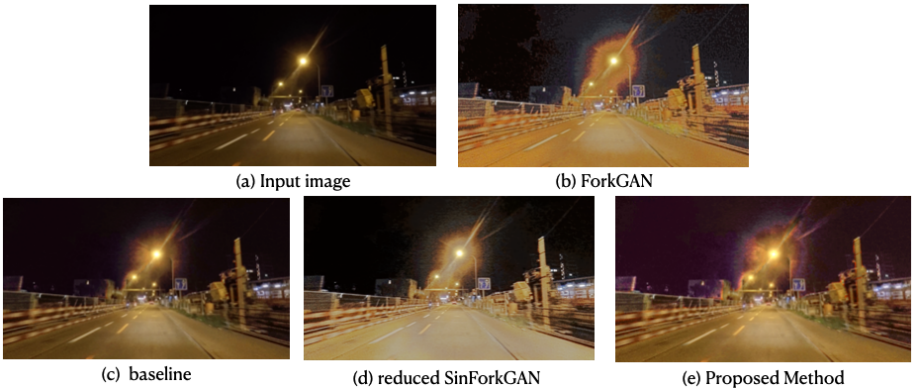

Fig. 7. Visual results for ablation studies on BDD100K dataset.