ForkGAN: Seeing into the Rainy Night

Ziqiang Zheng[1], Yang Wu[2] ⋆, Xinran Han[3], and Jianbo Shi[3]

[1] UISEE Technology (Beijing) Co., Ltd. zhengziqiang1@gmail.com
[2] Kyoto University wu.yang.8c@kyoto-u.ac.jp
[3] University of Pennsylvania {hxinran,jshi}@seas.upenn.edu

Abstract. We present a ForkGAN for task-agnostic image translation that can boost multiple vision tasks in adverse weather conditions. Three tasks of image localization/retrieval, semantic image segmentation, and object detection are evaluated. The key challenge is achieving high-quality image translation without any explicit supervision, or task awareness. Our innovation is a fork-shape generator with one encoder and two decoders that disentangles the domain-specific and domain-invariant information. We force the cyclic translation between the weather conditions to go through a common encoding space, and make sure the encoding features reveal no information about the domains. Experimental results show our algorithm produces state-of-the-art image synthesis results and boost three vision tasks' performances in adverse weathers.

Keywords: Light illumination · Image-to-image translation · Image synthesis · Generative adversarial networks

1 Introduction

Data bias is a well-known challenge for deep learning methods. An AI algorithm trained on one dataset often has to pay a performance deficit in a different dataset. Take an example of image recognition on a rainy night. An object detector trained on a day time dataset could suffer 30-50 percent accuracy drop on rainy night images. One solution is simply collecting more labeled data in those adverse weather conditions [8, 7, 24, 10, 19]. This is expensive and more fundamentally does not address the data bias issue.

Domain adaptation [28, 13, 23] is a general solution to this data bias problem. Our work is related to a sub-branch of this approach focusing on *image-to-image translation* techniques to explicitly synthesize images in uncommon domains. In the context of day and night domain change, two strategies have been explored recently: one is day-to-night approach [22], which transfers annotated daytime data to nighttime so that the annotations can be reused through data augmentation; the other [1] uses a night-to-day translator to generate images suitable for existing models trained on daytime data. The two strategies both demonstrated that the precise domain translation methods can boost the other vision tasks. In

⋆ *Corresponding author: Yang Wu.*

Fig. 1. Our ForkGAN can boost performance for multiple vision tasks with night to day image translation: localization (by SIT point matching between daytime and nighttime images), semantic segmentation and object detection (by data augmentation) in autonomous driving (results are all shown on the nighttime images).

this paper, we look into a far more challenging case of a rainy night. Our experiments show that existing approaches perform poorly in this case, particularly when we have no supervised data annotation on the rainy night images.

The fundamental challenge is that what makes an image looks good to a human might not improve computer vision algorithms. A computer vision algorithm can handle certain types of lighting change surprisingly well, while minor artifacts invisible to human could be harmful to vision algorithms.

A straightforward method is to introduce task-specific supervision on the new domain to ensure the image translation is task aware. We believe task-aware approaches only shift the data bias problem to a task bias problem. Instead, we ask if we can create a *task agnostic* image-to-image translation algorithm that improves computer vision algorithms without any supervision or task information. Fig. 1 shows our solution can achieve this goal on three untrained tasks: image localization, semantic segmentation, and object detection.

Problem Analysis. Domain translation between adverse conditions (*e.g.* nighttime) and standard conditions (*e.g.* daytime) is inherently a challenging unsupervised or weakly-supervised learning problem, as it is impossible to get precisely aligned ground-truth image pairs captured at a different time for dynamic driving scenes where a lot of moving objects exist. Many objects (*e.g.* the vehicles and street lamps) look totally differently across different weather conditions. There are global scene level texture differences such as raindrops, as well as regional changes such as cars' reflection on the wet road. There is a common semantic and geometrical level similarity between the adverse and normal domain, as well as vast differences. Precisely disentangling the invariant and variant features, without any supervision or task knowledge, is our key objective.

Proposed Solution. An ideal task-agnostic image translation preserves the image contents at all scale levels: scene level layout to object details such as letters on a traffic sign, while automatically adjusting to the illumination and weather conditions. For CycleGAN-based models that mainly rely on cycle-consistency losses, altering the global conditions can be done effectively, but faithfully maintaining the informative content details is not guaranteed. We first 'tie' the two encoding space of the CycleGAN together, to make sure we have kept only the necessary invariant information in both domains. We further explicitly check this encoding is domain agnostic: by looking at the encoded features, we cannot tell the domain they come from. This step can potentially remove much invariant information. We add a 'Fork' branch to check if we have encoded sufficient information to reconstruct the original image data in both domains. The model is called ForkGAN. It has the following main contributions.

- We propose a Fork-shaped Cyclic generative module that can decouple domain-invariant content and domain-specific style during domain translation. We force both encoders to go through a common encoding space and explicitly use an anti-contrastic loss to ensure necessary invariant information is produced in the disentanglement.
- We introduce a Fork-branch on each generator stage, to ensure sufficient information is kept for image recognition tasks in both domains.
- We boost the performance of localization, semantic segmentation and object detection in adverse conditions using our ForkGAN.

2 Related Work

2.1 Unpaired Image-to-image Translation

Many models have been proposed for unpaired image-to-image translation task, which aims to translate images from source domain to the corresponding desired images in target domain without corresponding image pairs for training. Introduced by Zhu et al. [28], CycleGAN is a classical and elegant solution for unpaired image-to-image translation. The cycle-consistency loss provides a natural and nice way to regularize the image translation, and it has become a widely used base. However, it does not enforce the translated image to share the same semantic space as the source image, and therefore its disentanglement ability is rather weak. UNIT [17] added a shared latent space assumption and enforced weight sharing between the two generators. However, weight sharing does not always guarantee that the network will learn to disentangle the images from different domains. To improve the diversity of generated results, models such as MUNIT [12], DRIT [16] were proposed to better decompose visual information into domain-invariant content and domain-specific style. To handle the translation among multiple domains, StarGAN [5] was developed by combining an additional classification loss. One drawback of those models is that the user will need to specify a Etyle codeEor label to sample from. For application in autonomous driving, we want the model to translate the image in adverse weathers to an appropriate condition without any human guidance during inference time.

2.2 Low-light Image Enhancement

Besides translating images from adverse conditions (*e.g.* night domain) to standard conditions, another possible approach to tackle the lack of visibility at night is to use low-light image enhancement models. Those models aim to improve the visual quality of underexposed photos by manipulating the color, brightness and contrast of the image. Recently, more deep learning based models have been proposed to solve underexposure problem. EnlightenGAN [14] can perform low-light enhancement without paired training data. The model increases the luminosity of an image while preserving the texture and structure of objects. However, without emphasis on foreground objects, EnlightenGAN provides limited details that are helpful for driving purposes. Different from these low-light image enhancement methods, we target to translate the whole image to day time and enhance the weak object signals in the dark.

2.3 Bad Weather Vision Tasks

Adverse weathers and undesirable illumination conditions pose challenges to common vision tasks such as localization, semantic segmentation and object detection. Visual localization and navigation allow the vehicle or robot to estimate its location and orientation in the real world. One efficient approach for this task is to use image retrieval techniques [1] and feature matching methods [18, 2]. However, these methods suffer from performance degradation when the query image is sampled from different illumination and weather conditions as compared with the labelled database. ToDayGAN [1] modified the image translation model to improve image retrieval performance for localisation task. Porav et al. [21] proposed a system that translates input images to a desired domain to optimize feature-matching results.

For semantic segmentation, Porav et al.[20] proposed a system that uses lightweight adapters to transform images of different weather and lighting conditions to an ideal condition for training off-the-shelf computer vision models. To train the adapters, they chose a sequence of reference images under ideal condition, and use CycleGAN [28] to synthesize images in different weathers while preserving the geometry and structure of the reference images. They then trained adapters to transform images from specific domains so using the new images can achieve better performance on related vision tasks.

Object detection, despite its importance, has received less attention in recent works on driving in adverse weathers. A related work in this direction is from He et al. [8], where the authors developed a multi-adversarial Faster R-CNN framework for domain-adaptive object detection in driving scenario. Their source and target domain pairs involve regular and foggy Cityscapes, synthetic and real data from two different driving datasets with similar weather conditions. AugGAN [11] aims to combine an image parsing network to enhance object detection performance in nighttime images through day-to-night translation on synthetic datasets. However, it requires paired auxiliary annotations (*e.g.* semantic segmentation maps), which are sometimes expensive or hard to acquire, to regularize

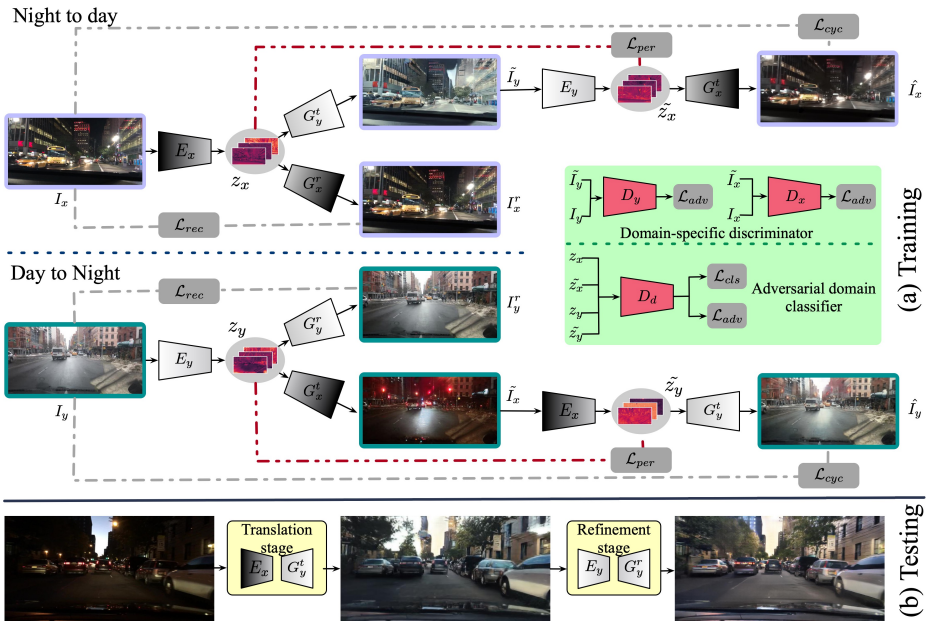

Fig. 2. The framework of ForkGAN. (a) shows the training stage while (b) represents the testing inference. I_x and I_y denote random image from night domain $\mathbb{X}$ and day domain $\mathbb{Y}$, separately. E_x and E_y are the encoders to encode the night and day images, separately. G_x^t and G_y^t are responsible to achieve domain translation based on the domain-invariant representations z and $\tilde{z}$. G_x^r and G_y^r aim to reconstruct the input images based on the representations. D_x and D_y are the discriminators of $\mathbb{X}$ and $\mathbb{Y}$, while D_d is the domain adversarial classifier.

the image parsing network. Our ForkGAN addresses object detection under more challenging weather conditions - driving scenes at nighttime with reflections and noise from rain and even storms, without any auxiliary annotations.

3 Proposed Method

3.1 ForkGAN Overall Framework

Our ForkGAN performs image translation with unpaired data using a novel fork-shape architecture. The fork-shape module contains one encoder and two decoders. Take night-to-day translation in Fig. 2 as an example, first we feed a nighttime image I_x to the encoder E_x and obtain the domain-invariant representation z_x. Then the two decoders G_x^r (reconstruction decoder) and G_x^t (translation decoder) have the same input z_x. G_x^r aims to synthesize the original nighttime image I_x^r from the invariant representation and we perform a pixel-level l_1-norm based reconstruction loss $\mathcal{L}_{rec}$ between I_x^r and I_x. G_x^t is responsible

to generate plausible image $\tilde{I}_y$ that looks like night images but under daytime illumination. We leverage adversarial training through one domain-specific discriminator D_y and compute the adversarial loss $\mathcal{L}_{adv}$ (same as the one in CycleGAN [28]), which aims to distinguish the random real night image I_y and the synthesized night image $\tilde{I}_y$. Then E_y extracts the domain-invariant feature $\tilde{z_x}$ from $\tilde{I}_y$. Here, we perform a perceptual loss $\mathcal{L}_{per}$ (to be detailed in 3.2) between $\tilde{z_x}$ and z_x to force I_x and $\tilde{I}_y$ to have similar content representation. Finally we obtain the reconstructed night image $\hat{I}_x$ using the translation decoder G_x^t. The cycle-consistency loss $\mathcal{L}_{cyc}$ is computed between $\hat{I}_x$ and I_x. Note, here we omit the reconstruction decoder G_y^r, which is used to reconstruct the day image based on the domain-invariant feature z_y. Moreover, we adopt one additional adversarial domain classifier D_d, which has two branch outputs: one for adversarial training and another for domain classification to obtain the cross-entropy classification loss $\mathcal{L}_{cls}$ based on the content representations. The total loss of ForkGAN is a weighted sum of all the losses mentioned above:

$$\mathcal{L}(E, G^r, G^t) = \mathcal{L}_{adv} + \mathcal{L}_{cls} + \mathcal{L}_{per} + \gamma\mathcal{L}_{cyc} + \epsilon\mathcal{L}_{rec}, \tag{1}$$

and we set $\gamma = \epsilon = 10$ in our experiments. With the total loss, the three components E, G^r, and G^t are optimized together so that the learned model is unbiased and can disentangle the domain-invariant content and domain-specific style. During inference time, our ForkGAN provides a two-stage translation procedure as shown in Fig. 2. Take night-to-day translation as an example, the input night image is translated to a daytime image using E_x and G_y^t, and the output is regarded as input of the refinement stage. E_y and G_y^r synthesize more precise translation output, which gives the final output of our ForkGAN.

3.2 ForkGAN - Disentanglement Stage

Previous Cycle-GAN based methods target to preserve the appearance of input images through an indirect pixel-level cycle-consistency loss and generate plausible translated image by leveraging adversarial loss to the translated images. However, some weak but informative domain-invariant characteristics are usually ignored during translation stage. Sometimes, the generator $\mathcal{F} : \mathbb{X} \longrightarrow \mathbb{Y}$ can fool the discriminator and minimize the adversarial loss by changing the global conditions that dominate more pixels, while ignoring local features such as cars and pedestrians. This leads to a trivial translation solution that throws away some informative signals. In the opposite direction, $\mathcal{G} : \mathbb{Y} \longrightarrow \mathbb{X}$ has a strong ability to remap the translated image to the original domain under a strong pixel-level cycle-consistency loss. In the two stages, there is no guarantee that the domain-invariant and domain-specific feature can be disentangled. Our fork-shape generator can achieve better disentanglement because:

1) E aims to extract the domain-invariant content and discard the domain-specific style. G^r and G^t target to reload the style representation of the source and target domain separately. The three components have same parameter quantity, which ensures comparable network capacity so that there is no explicit bias

or dependency among the three components. If E is too weak and fails to extract the informative content, the reconstruction loss $\mathcal{L}_{rec}$ is large. If the domain-invariant representation z_x is still mixed with the domain-specific information, the translation decoder will fail to generate reasonable translated output.

2) We impose a perceptual loss as

$$\mathcal{L}_{per} = \tau(\sum_{n=1}^{N} \lambda_n ||\Phi_n(\tilde{z_x})) - \Phi_n(z_x)||_1), \tag{2}$$

which makes $\tilde{z_x}$ perceptually similar to z_x (designed according to the perceptual loss in [4]). Here, Φ_n denotes the feature extractor at the n_{th} level of the pre-trained VGG-19 network on ImageNet. The hyper-parameter λ_n controls the influence of perceptual loss at different levels and here we set λ_n all 1. Different from the way perceptual loss is typically used (feeding image data to the VGG network), we rearrange the feature maps of z_x and $\tilde{z_x}$ through bilinear interpolation to fit into only the last three layers of VGG. Such a modification enables an effective perceptual consistency check between z_x and $\tilde{z_x}$ at the feature level. If E_x and E_y fail to eliminate the domain-specific information completely, the perceptual loss between z_x and $\tilde{z_x}$ will be large. The perceptual loss can also help preserve the content information during translation stage.

3) The adversarial domain classifier targets to distinguish the real/fake distribution and classify the content representation. We aim to match the distribution of z and $\tilde{z}$ through adversarial training. Specifically, we assign an opposite label to z and $\tilde{z}$ to implement an classification training to obtain $\mathcal{L}_{cls}$. We perform the classification loss using both z and $\tilde{z}$. If the classifier could not distinguish which domain the content representation is from, it indicates that the extracted representation does not carry any domain-specific style information.

Based on above reasons, the design of our model and training objectives can provide strong constraints to achieve disentanglement.

3.3 ForkGAN - Refinement Stage

In the fork-shape module, the generator has two branches: the translation branch and the reconstruction branch. We apply an additional refinement stage to the translated output using autoencoder E_y and reconstructions decoder G_y^r. This pair is trained during disentanglement stage and therefore the refinement does not introduce new parameters. During this stage, the reconstruction branch G_y^r can refine the fake outputs ($\tilde{I_x}$ and $\tilde{I_y}$) by knowledge learned from reconstructing the real images, thus generating more realistic images and strengthening weak signals. We adopt additional pixel-wise Gaussian noise disturbance to the domain-invariant content representation z to improve the robustness of the reconstruction branch and make it less input-sensitive. We also hope that the reconstructed decoder can generate complementary information from additional noise even if some domain-invariant content feature is missed. In this way, the two-stage translation shown in Fig. 2 can obtain better translation performance even in adverse environment.

3.4 Dilated Convolution & Multi-scale Discriminator Architecture

Considering the occlusion and reflection of images captured in adverse weathers, it is difficult to recognize the objects essential to the task of navigation (*e.g.* traffic signs, lanes and other vehicles). A possible solution is to adopt a large receptive field to alleviate the occlusion issue. To do that, we use dilated residual networks [26] for the generator with fewer parameters. The dilated convolution can help the three components of our generator understand the relationship of different parts. To achieve high-resolution image-to-image translation, we adopt the multi-scale discriminator architecture [25, 12, 16] to improve the ability to distinguish the fake images and real images. The proposed architecture could fuse the information from multiple scales and generate more realistic outputs.

4 Experiments

4.1 Datasets and evaluation metrics

Datasets: **Alderley** is originally proposed for the SeqSLAM algorithm [19], which collected the images for the same route twice: once on a sunny day and another time on a stormy rainy night. Every frame in the dataset is GPS-tagged, and thus each nighttime frame has a corresponding daytime frame. The images collected at nighttime are blurry with a lot of reflections, which render the front vehicles, lanes and traffic signs difficult to be recognized. For this dataset, we use the first consecutive four fifths for training and others for evaluation. Since this dataset has day-night correspondences, we use it for quantitative evaluation on image localization task. Unfortunately, it doesn't provide ground-truth annotations for semantic segmentation and object detection, so we use another dataset instead for those two tasks. **BDD100K** [27] is a large scale high-resolution autonomous driving dataset, which collected 100,000 video clips in multiple cities and under various conditions. For each video, it selects a key frame to provide detailed annotations (such as the bounding box of various objects, the dense pixel annotation, the daytime annotation and so on). We reorganized this dataset according to the annotation, and obtained 27,971 night images for training and 3,929 night images for evaluation. We obtained 36728/5258 train/val split for day images. We inherit the data split from the BDD100K dataset. We perform semantic segmentation and object detection on this dataset.

Image Quality Metric: **FID** [9] evaluates the distance between the real sample distribution and the generated sample distribution. Lower FID score indicates higher image generation quality.

Vision Task Metrics: For **Localization**: SIFT [18] is good measure to find the feature matching points between two images. We measure the localization performance by the SIFT interesting points matching. **Semantic Segmentation**: Intersection-over-Union(IoU) is a commonly used metric for semantic segmentation. For each object class, the IoU is the overlap between predicted segmentation map and the ground truth, divided by their union. In the case of multiple classes, we take the average IoU of all classes (*i.e.*, mIoU) to indicate

the overall performance of the model. **Object Detection**: We use mean average precision (mAP) to evaluate the performance and also report the average precision scores for individual classes to have a more thorough evaluation.

4.2 Experiment settings and implementation details

We compare our proposed method with other state-of-the-art image translation methods such as UNIT [17], CycleGAN [28], MUNIT [12], DRIT [16], UGATIT [15] and StarGAN [1]. Additionally, we also compare with low-light enhancement methods such as EnlightenGAN [14] and ToDayGAN [1]. We follow the instructions of those methods and make a fair setting for comparison.

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 discriminator architecture and we set the number of scales as 2. For the adversarial domain classifier, the backbone has 4 Conv-Ins-ReLU blocks, the adversarial branch has one additional convolution layer to get adversarial output, while the classification branch has one more fully-connected layer to obtain a domain classification output. We adopt Adam optimizer and set learning rate to 0.0002.

4.3 Localization by SIFT point matching

We aim to perform translation at an extremely difficult setting on Alderley dataset. Fig. 3 shows the qualitative translation result comparisons. UNIT and MUNIT fail to perform reasonable translation and generate plausible objects. DRIT has lost the detailed information and missed some objects after domain translation. The result of EnlightenGAN fails to provide meaningful visual information and it only changed the illumination slightly. ToDayGAN and UGATIT obtain better translation results and have captured the visual objects in the darkness. But they cannot preserve the visual objects (*e.g.*, traffic signs and cars) well. In contrast, our method has stronger ability to capture these weak signals and preserve them better. For this dataset, we perform experiments using 512*256 resolution. We compute the number of SIFT matching points between the translated daytime images and the corresponding natural daytime images. Table 1 reports the quantitative comparison. Our ForkGAN obtains the best SIFT result through precise night-to-day image translation. It has also achieved the best image generation quality with lowest FID score. By improving the ability to maintain and enhance the SIFT matching, it can benefit place recognition and visual localization.

Ablation studies Several experiments are designed for ablation studies. Firstly, we remove the Fork-shape architecture (denoted as w/o Fork-shape) of the generator, and follow the setting of Cycle-GAN methods to optimize the model. The

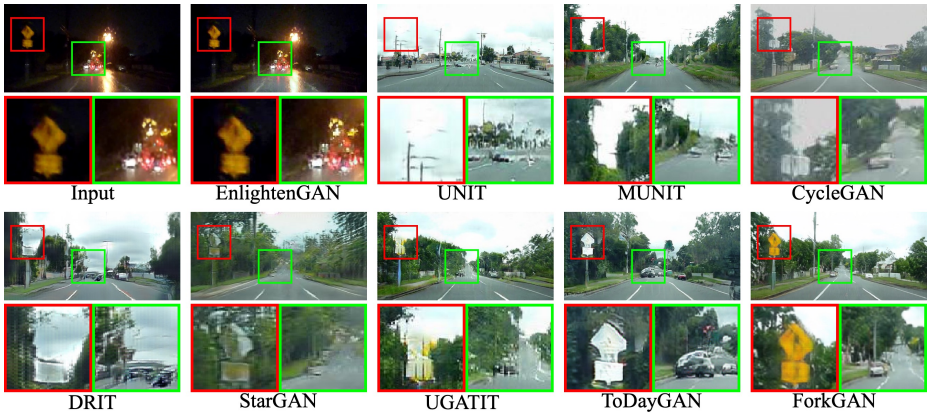

Fig. 3. The visual/qualitative translation result comparison of different methods. Please zoom in to check more details on the content and quality. The parts covered by red and green boxes show the enlarged cropped region in the corresponding image.

Table 1. Evaluation metric results of different methods for night → day translation task on Alderley dataset [19]. The original denotes the scores of the original real night images. FID reports the visual image quality (lower is better) while SIFT reports the localization performance (higher is better).

Method	EnlightenGAN [14]	UNIT [17]	MUNIT [12]	CycleGAN [28]	DRIT [16]
FID / SIFT	249 / 2.00	155 / 2.68	138 / 2.75	167 / 3.36	145 / 3.71
Method	StarGAN [5]	UGATIT [15]	ToDayGAN [1]	ForkGAN	Original
FID / SIFT	117 / 3.28	170 / 2.51	104 / 4.14	**61.2** / **12.1**	210 / 3.12

result generated by the vanilla generator has artifacts on cars as there is no guarantee on the disentanglement between the domain-invariant and domain-specific information. Then we investigate the effectiveness of the Fork-shape generator itself only (with a name of "Fork-shape"). Note, we do not compute $\mathcal{L}_{per}$ and the adversarial domain classification loss $\mathcal{L}_{cls}$ in this setting. Due to the reconstruction loss, the synthesized images have fewer artifacts having all but the Fork-shape. Based on this, we aim to explore the improvement from the dilated residual blocks (DRB for abbreviation) by evaluating "Fork-shape+DRB". A larger receptive field can help the generator to better capture the objects in the dark. Then we adopt a multi-scale discriminator architecture (MSD for abbreviation), here we set $n = 2$. As reported in Table 2, the MSD architecture can also lead to the improvement on FID and SIFT matching (as shown by the results of "Fork-shape+DRB+MSD"). In the next experiment, we use everything we have covered for training ForkGAN, and just exclude the refinement stage when

Table 2. Quantitative comparisons for ablation studies on Alderley dataset. [19]

Method	w/o Fork-shape	Fork-shape	Fork-shape+DRB
FID / SIFT	146 / 4.26	131 / 7.12	113 / 8.12
Fork-shape+DRB+MSD	w/o refinement	w/ shared decoders	ForkGAN
95.3 / 9.14	70.7 / 11.5	73.8/ 9.29	**61.2** / **12.1**

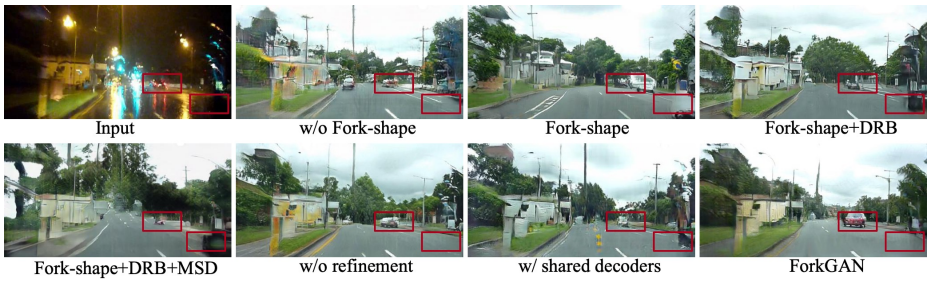

Fig. 4. Visual results for ablation studies on Alderley. Red boxes highlight some details.

we use it for testing, which is denoted by "w/o refinement". As shown in Fig. 4 and Table 2, adding $\mathcal{L}_{per}$ and $\mathcal{L}_{cls}$ to "Fork-shape+DRB+MSD" can achieve better disentanglement, which leads to significantly better translation outputs. Finally, we apply the "ForkGAN" with the refinement stage at the testing stage and observe that the refinement can greatly improve the detailed part generation. Last but not the least, we also evaluate a twisted version of ForkGAN by letting the translation decoder and reconstruction decoder of the same domain (*e.g.*, G_y^t and G_y^r) share the same parameters (basically using the same decoder instead of two different ones), and have it denoted by "w/ shared decoders". As showed, it results in a significant performance drop when compared with "ForkGAN" which doesn't have shared decoders. The two decoders for the same domain look similar, but they are constrained by different losses and thus have different duties, which complement each other. Putting the loads on one single decoder makes it much harder to achieve the goal and leads to inferior model. All the quantitative comparisons of above different settings are listed in Table 2 and qualitative ones are given in Fig. 4.

4.4 Semantic segmentation

Moreover, we perform high-resolution (1024×512) night-to-day image translation to boost the semantic segmentation performance. Fig. 5 presents the translated results and corresponding segmentation outputs of various methods. For

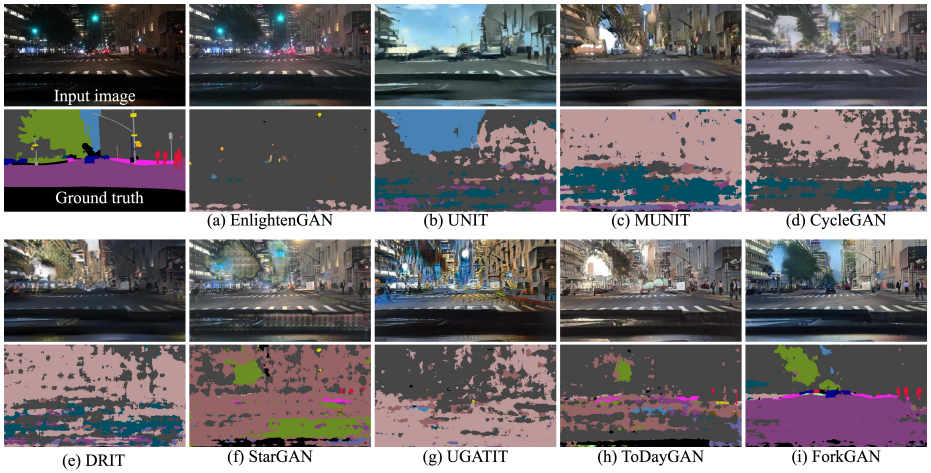

Fig. 5. The visual translation (the first row) and segmentation performance (the second row) comparison of different methods, with models pre-trained on Cityscapes [6].

semantic segmentation, we use a pre-trained Deeplab-v3 model [4] on Cityscapes dataset [6]. The BDD100K dataset provides segmentation ground truth of 137 night images. So we compute the IOU metric between the segmentation outputs of the 137 translated daytime images and corresponding segmentation ground truth. The quantitative comparison is listed in Table 3. Since there is no night image on Cityscapes dataset, the segmentation performance of night image has a drastic performance drop shown in Table 3. The mIoU is only 7.03 percent if we directly perform the semantic segmentation on the real night images. Night-to-day translation model provides a powerful tool to improve segmentation performance, where stronger translation model should lead to larger performance boost. As shown, our ForkGAN achieves the highest mIoU among all the methods, almost doubling the original night image segmentation result. We also observe that the synthesized daytime images produced by some comparative translation methods obtain worse segmentation performance than the original night images. MUNIT and DRIT methods both fail to synthesize plausible outputs from challenging night images and thus obtain poor mIoU scores. ToDayGAN, while achieving reasonable night-to-day translation, obtains higher mIoU score than original night images. Our ForkGAN preserves detailed information during night-to-day image translation, especially the small traffic signs and pedestrians. So our method can boost the segmentation performance by preserving and enhancing the crucial detailed information. To quantitatively compare the translation quality, we also compute the FID score to measure the distance between the generated sample distribution and the real image distribution in Table 3.

[4] https://github.com/srihari-humbarwadi/DeepLabV3_Plus-Tensorflow2.0

Table 3. Quantitative comparison of different methods for night → day translation task on BDD100K dataset [27]. The original denotes the outputs of the original real night images. FID reports the visual image quality (lower is better) while mIoU (percentage) reports the segmentation performance (higher is better).

Method	EnlightenGAN [14]	UNIT [17]	MUNIT [12]	CycleGAN [28]	DRIT [16]
FID / mIoU	90.3 / 6.03	62.1 / 2.47	61.1 / 2.44	51.7 / 1.88	53.1 / 2.45
Method	StarGAN [5]	UGATIT [15]	ToDayGAN [1]	ForkGAN	Original
FID / mIoU	68.3 / 6.63	72.2 / 3.83	43.8 / 8.19	**37.6** / **14.4**	101 / 7.03

4.5 Object detection with data augmentation

In autonomous driving, it is laborious and sometimes difficult to collect abundant data with annotations in a wide variety of weather and illumination conditions for object detection. Most of available datasets contain images mostly from daytime driving. Models trained on those datasets are subject to performance degradation once they are tested on a different domain such as nighttime. One possible solution is to augment nighttime data with annotated daytime images through domain translation such that we can make the most use of available annotations. Our ForkGAN can also perform day-to-night translation to aid off-the-shelf detection model to adapt to different domains. We compare our ForkGAN with the most related ToDayGAN on BDD100K dataset in two settings. In both settings, we have unlabelled images from both day and night domains for training image translation network, as well as bounding box annotations for daytime images for training detection network, either with real or translated images:

1) **Day Labels Only** - No nighttime labeled image is available at training time: We use ForkGAN to translate daytime images to night images and preserve the corresponding bounding boxes. Then we train an object detection network on those translated nighttime images. For comparison, we also train two separate detection networks using raw daytime images(*Day Real*) and translated nighttime images by ToDayGAN. The quantitative results are shown in Table 4. We observe that ForkGAN can improve the detection performance on night images. Visualization of detection results are shown in Fig. 6. The ability to detect small traffic signs in dark has been improved through domain adaptation.

2) **Day + Night Labels** - Both nighttime and daytime labeled images are available for training: We again apply ForkGAN to translate the daytime images to night images for data augmentation. The detection network is trained on both real and translated night images. We also report the performance of the detection network trained only on real night images(*Night Real*) and night image augmentation with ToDayGAN(*Night+ToDayGAN*). Fig. 6 and Table 4 show the visual and quantitative comparison. By combining with translated night images, the detection performance has been improved, which indicates the detection task can benefit from domain translation.

Table 4. Comparisons for object detection on 3,929 validation nighttime images. The first three rows show the results from setting 1), while the rest are from setting 2). We apply faster-rcnn-r50-fpn-1x detector based on MMDetection [3] in all the experiments.

Method	mAP	person	rider	car	bus	truck	bike	motor	traffic light	traffic sign
Day(Real)	22.1	26.1	**14.3**	37.5	29.8	30.7	**18.5**	**16.3**	14.6	33.1
+ToDayGAN	19.5	23.5	10.4	35.9	32.5	29.4	16.0	11.0	9.0	26.7
+ForkGAN	**22.9**	**26.3**	13.0	**41.2**	**33.3**	**32.1**	16.4	15.9	**16.2**	**34.5**
Night(Real)	23.9	26.6	13.0	42.0	33.8	35.0	16.7	16.9	18.2	36.0
Night+ToDayGAN	24.2	26.9	14.1	42.3	36.5	36.8	20.2	19.1	17.6	35.7
Night+ForkGAN	**26.2**	**28.1**	**16.1**	**42.5**	**37.8**	**38.7**	**22.1**	**21.9**	**18.3**	**36.2**

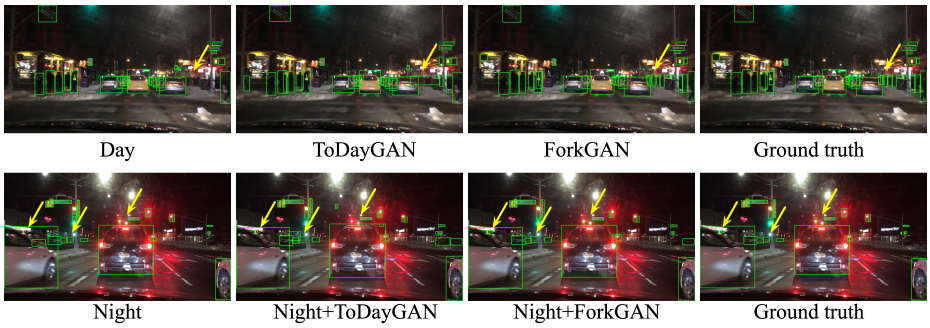

Fig. 6. Visual comparison of detection results on BDD100K, where "Day"/"Night" denotes training with daytime/nighttime images. Areas pointed by yellow arrows are worth attention. ForkGAN can improve the detection of small objects. We show all the results of person, rider, car, bus, truck, bike, motor, traffic light and traffic signs.

5 Conclusion and Future Work

We propose a novel framework ForkGAN to achieve unbiased image translation, which is beneficial to multiple vision tasks: localization/retrieval, semantic segmentation and object detection in adverse conditions. It disentangles domain-invariant and domain-specific information through a fork-shape module, enhanced by an adversarial domain classifier and an across-translation perceptual loss. Extensive experiments have demonstrated its superiority and effectiveness. Possible future works include designing a multi-task learning network to share the backbone of different vision tasks and performing object detection in the domain-invariant content space, which can be more compact and more efficient.

Acknowledgement

This work was supported by a MSRA Collaborative Research 2019 Grant.

References

1. Anoosheh, A., Sattler, T., Timofte, R., Pollefeys, M., Van Gool, L.: Night-to-day image translation for retrieval-based localization. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 5958–5964. IEEE (2019)
2. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Computer vision and image understanding **110**(3), 346–359 (2008)
3. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., et al.: Mmdetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
4. Chen, Q., Koltun, V.: Photographic image synthesis with cascaded refinement networks. In: IEEE International Conference on Computer Vision. pp. 1511–1520 (2017)
5. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In: CVPR. pp. 8789–8797 (2018)
6. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
7. Halder, S.S., Lalonde, J.F., de Charette, R.: Physics-based rendering for improving robustness to rain. In: IEEE/CVF International Conference on Computer Vision (2019)
8. He, Z., Zhang, L.: Multi-adversarial faster-rcnn for unrestricted object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6668–6677 (2019)
9. 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. In: NIPSs. pp. 6626–6637 (2017)
10. Hu, X., Fu, C.W., Zhu, L., Heng, P.A.: Depth-attentional features for single-image rain removal. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8022–8031 (2019)
11. Huang, S.W., Lin, C.T., Chen, S.P., Wu, Y.Y., Hsu, P.H., Lai, S.H.: Auggan: Cross domain adaptation with gan-based data augmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 718–731 (2018)
12. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: European Conference on Computer Vision. pp. 172–189 (2018)
13. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR. pp. 5967–5976 (2017)
14. 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. arXiv preprint arXiv:1906.06972 (2019)
15. Kim, J., Kim, M., Kang, H., Lee, K.: U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. arXiv preprint arXiv:1907.10830 (2019)
16. Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Diverse image-to-image translation via disentangled representations. In: European Conference on Computer Vision. pp. 35–51 (2018)

17. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: Advances in Neural Information Processing Systems. pp. 700–708 (2017)
18. Lowe, D.G., et al.: Object recognition from local scale-invariant features. In: iccv. vol. 99, pp. 1150–1157 (1999)
19. Milford, M.J., Wyeth, G.F.: Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights. In: 2012 IEEE International Conference on Robotics and Automation. pp. 1643–1649. IEEE (2012)
20. Porav, H., Bruls, T., Newman, P.: Don't worry about the weather: Unsupervised condition-dependent domain adaptation. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC). pp. 33–40. IEEE (2019)
21. Porav, H., Maddern, W., Newman, P.: Adversarial training for adverse conditions: Robust metric localisation using appearance transfer. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 1011–1018. IEEE (2018)
22. Romera, E., Bergasa, L.M., Yang, K., Alvarez, J.M., Barea, R.: Bridging the day and night domain gap for semantic segmentation. In: 2019 IEEE Intelligent Vehicles Symposium (IV). pp. 1312–1318. IEEE (2019)
23. Ros, G., Alvarez, J.M.: Unsupervised image transformation for outdoor semantic labelling. In: 2015 IEEE Intelligent Vehicles Symposium (IV). pp. 537–542. IEEE (2015)
24. Sakaridis, C., Dai, D., Van Gool, L.: Semantic foggy scene understanding with synthetic data. International Journal of Computer Vision **126**(9), 973–992 (2018)
25. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8798–8807 (2018)
26. Yu, F., Koltun, V., Funkhouser, T.: Dilated residual networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 472–480 (2017)
27. Yu, F., Xian, W., Chen, Y., Liu, F., Liao, M., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving video database with scalable annotation tooling. arXiv preprint arXiv:1805.04687 (2018)
28. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: International Conference on Computer Vision. pp. 2223–2232 (2017)