Daydriex: Translating Nighttime Scenes towards Daytime Driving Experience at Night

Euihyeok Lee 1 and Seungwoo Kang 2, *

1 Department of Computer Science and Engineering, Graduate School, KOREATECH, Cheonan-si, Chungcheongnam-do 31253, Korea; euihyeok.lee@misl.koreatech.ac.kr
2 School of Computer Science and Engineering, KOREATECH, Cheonan-si, Chungcheongnam-do 31253, Korea
* Correspondence: swkang@koreatech.ac.kr

Abstract: What if the window of our cars is a magic window, which transforms dark views outside of the window at night into bright ones as we can see in the daytime? To realize such a window, one of important requirements is that the stream of transformed images displayed on the window should be of high quality so that users perceive it as real scenes in the day. Although image-to-image translation techniques based on Generative Adversarial Networks (GANs) have been widely studied, night-to-day image translation is still a challenging task. In this paper, we propose Daydriex, a processing pipeline to generate enhanced daytime translation focusing on road views. Our key idea is to supplement the missing information in dark areas of input image frames by using existing daytime images corresponding to the input images from street view services. We present a detailed processing flow and address several issues to realize our idea. Our evaluation shows that the results by Daydriex achieves lower Fréchet Inception Distance (FID) scores and higher user perception scores compared to those by CycleGAN only.

Keywords: daytime driving experience; image translation; daytime image; road view; GAN

1. Introduction

Imagine that you are riding in an autonomous car and passing through the Amalfi Coast (Figure 1), one of the world’s most scenic roads in Italy listed in the UNESCO World Heritage. You do not have to worry about driving your car, but could enjoy the breathtaking view through the window of your car. Unfortunately, however, it is night. You can see only the tiny part of such a beautiful scene. What if your car window is a magic window, which transforms dark views outside of the window into bright ones as you can see in the daytime?

To realize such a window, one of the important requirements is that the stream of transformed images displayed on the window should be of high quality so that users perceive it as real scenes in the day. This problem can be seen as a problem of image-to-image translation focusing on nighttime to daytime image translation. Image-to-image translation based on Generative Adversarial Networks (GANs) has been widely studied in the deep learning field. Among others, there have been research works for unsupervised image-to-image translation including CycleGAN [1], UNIT [2], MUNIT [3], EGSC-IT [4]. They are advantageous since they do not require paired training images to learn the mapping between two domains of images. Existing image translation models have been applied to diverse applications such as style transfer, season transfer, and photo enhancement, and they have shown impressive translation results. However, night-to-day image translation is still a challenging task. In many cases, too dark areas of the night images inevitably exist, which often causes poor quality of translated images. From our preliminary study with CycleGAN, we find that the CycleGAN model hardly provides quality results for dark areas in nighttime images. The areas rarely have information on geographical features,
objects, or structures that actually exist in those areas. It is quite difficult, if not impossible, to translate such dark areas by using input images only.

To address the problem, we propose a new approach, i.e., supplementing the missing information in dark areas of input image frames by using existing daytime images corresponding to the input images, which has not been studied yet. This is possible for our target scenario since there are readily available daytime street view images in several cities and rural areas worldwide from street view services. We briefly introduce the problem and our initial idea in [5]. However, it is not trivial to realize the approach.

![Figure 1. An example picture of the Amalfi Coast (Photo by Mike Morgan on Unsplash).](image)

In this paper, we present Daydriex, a processing pipeline to generate enhanced daytime translation focusing on road views to study the feasibility of the proposed approach. Its unique feature is to employ a data supplement processing step that supplements hint information for the dark areas of nighttime images. This processing step retrieves a daytime image from a street view service for an input nighttime image and augments the input image with hint information from the street view image. A resulting image from the data supplement processing step is used as an input of the daytime image generation. Such supplemented input images improve the quality of translation results. For daytime image generation, we adopt an existing image-to-image translation model. To realize the idea, we address several issues which we discuss in detail in Section 4.

To evaluate Daydriex compared to an existing translation model, we conduct an experiment that assesses the reality of Daydriex outputs in two aspects. First, we use the Fréchet Inception Distance (FID) [6], which is one of key metrics widely used to measure GAN performance. Second, we conduct a user perceptual study with 30 people. We show that the results by Daydriex achieve lower FID scores and higher user perception scores compared to those by CycleGAN only. We also present average computation time taken to run main operations of Daydriex.

The rest of this paper is organized as follows. Section 2 discusses related work. Section 3 presents a motivational study. Section 4 presents the proposed processing pipeline. Section 5 shows evaluation results and Section 6 concludes the paper.

2. Related Work

Thanks to the advance of deep learning, there have been research works on low-light image enhancement based on deep neural networks [7–12]. Some works build paired training data, i.e., low/normal-light image pairs, and use them to generate light-enhanced images. Chen et al. introduced a See-in-the-Dark (SID) dataset of short-exposure low-light raw images with corresponding long-exposure reference raw images [9]. Their proposed
model achieves high quality visual results. However, they require raw image data as well as paired low/normal-light training images. Cai et al. built a multi-exposure image dataset including under/over-contrast and normal-contrast encoded image pairs, and proposed multi-exposure image fusion or high dynamic range algorithms to generate the contrast enhanced images [12]. Jiang et al. proposed EnlightenGAN, which is an unsupervised GAN trained without paired data [10]. It shows the feasibility of learning unpaired data for low-light enhancement. Although these techniques can be used to make low-light images look bright, it is difficult to use them for our purpose of transforming nighttime images to daytime images. It is because light-enhanced nighttime images generated by these techniques would be still nighttime images of which brightness is improved.

There have been several studies on image-to-image translation techniques based on GANs [1–4,13,14]. These techniques are largely divided into supervised and unsupervised learning depending on how the training data set is constructed. Image-to-image translation based on unsupervised learning that trains a model through unpaired data sets can be used more flexibly because of the advantage of being able to collect data sets easily. Widely known unsupervised image-to-image translation models include CycleGAN [1] and DiscoGAN [14]. These models share the same concept, using a cycle-consistent loss with unpaired data to translate between two different domains. The cycle-consistency loss is a constraint for maintaining consistency with the original domain when generating a fake image of another domain. Although they present a similar concept, they have a different focus. The main goal of CycleGAN is to translate the style of high-resolution images, so the network size is large. Although it shows a good performance of translating appearance such as color and texture, it is limited to translate the shape of an object involving geometric changes. On the other hand, DiscoGAN is designed with a relatively simple network structure for finding cross-domain relations. It shows good performance in translating the style of different shapes of objects (e.g., shoes to handbags), but its resolution is not large. Another well-known model is a UNIT employing an architecture that allows information to be shared between domains by applying a shared latent space [2]. In the shared latent space, corresponding images of different domains can be mapped to the same latent representation. UNIT shows a good performance in translation of human faces and animal breeds. In addition, there are many works of using unsupervised image translation models for cartoons, characters, and art works [15–17].

Our aim in this paper can be seen as a specific problem of image-to-image translation focusing on nighttime to daytime image translation. Recently, there have been a couple of works targeting this scope. Anokhin et al. presented a high-resolution daytime translation model to generate daytime timelapse videos [18] exhibiting different times of the day and different lighting for a given input image. It presents a new architecture that combines an image-to-image translation model and a new upsampling scheme for high-resolution image generation. It minimizes the training time and generates output daytime images of high quality. Anoosheh et al. proposed a ToDayGAN to translate nighttime images to daytime images for a visual localization purpose [19]. In this work, the model for night-to-day translation is designed based on ComboGAN [20], which is equivalent to CycleGAN for the case of two domains. Although the generator networks are identical to the ones in CycleGAN, the discriminators are modified to contain three networks specialized for different aspects of input image, blurred RGB, luminance, and gradients. They show that the proposed model improves the performance of localizing nighttime query images given a set of daytime reference images. However, night-to-day image translation is still a challenging task. When changing nighttime images to daytime ones, too dark parts of the night images often exist, resulting in poor translation quality with existing translation models. To address the problem, we propose a new approach in this paper, using already existing daytime images (e.g., street view images) as hint information for translation. Although we briefly introduce the problem and our initial idea in [5], it is not trivial to realize the approach. In this paper, we present a detailed processing pipeline
and solutions to realize the idea which were not addressed previously. Also, we present in-depth evaluation results. Table 1 compares our work with the related works.

| Target Input          | Raw Format Input | Target Output          | Hint Supplemented Input Preprocessing |
|-----------------------|------------------|------------------------|--------------------------------------|
| Chen et al. (SID) [9] | Low-light image  | Yes                    | Light-enhanced image                 | No                                   |
| Cai et al. [12]       | Low-light image  | No                     | Light-enhanced image                 | No                                   |
| Jiang et al. (EnlightenGAN) [10] | Low-light image  | No                     | Light-enhanced image                 | No                                   |
| Anokhin et al. (HiDT) [18] | Different times of day image | No                   | Different times of day image         | No                                   |
| Anoosheh et al. (ToDayGAN) [19] | Nighttime image  | No                     | Daytime image                        | No                                   |
| Zhu et al. (CycleGAN) [1]  | Not specific     | No                     | Not specific                         | No                                   |
| Ours (Daydriex)       | Nighttime image  | No                     | Daytime image                        | Yes                                  |

To implement our idea, it is necessary to address a correspondence problem between input nighttime images acquired from a vehicle’s camera and hint images acquired through the street view services. The correspondence problem is a very common problem in the field of computer vision. This is mainly caused by spatial and temporal differences between images or distortion that occurs in the process of expressing three-dimensional reality in two dimensions. In our target scenario, there are two issues, position matching and perspective matching. First, the locations where nighttime images are acquired are highly likely to be different from the locations where the corresponding street view images are acquired. To provide appropriate hint information, we need to figure out a certain area of the street view image to match the input nighttime image. Although it could be possible to apply image registration techniques, we adopt a simple method of comparing the coordinates from which the two images are acquired and calculating the distance between them. Image registration is generally used in the correspondence problem in diverse domains such as medical images [21,22]. We expect that we can elaborate a solution to the position matching problem through image registration techniques, which we leave as future work. Second, the perspective of those two sets of images would be different since the position of the camera attached to the vehicle for acquiring street view images is different from the position of the camera to be used for our application scenario. To address this problem, perspective transformation can be used, which is a linear projection where three-dimensional objects are projected on a two-dimensional picture plane. It is used to solve the perspective problem in various domains such as road driving, industrial sites, and traffic [23–26]. In our proposed solution, we apply a warping technique based on perspective transformation.

3. Motivation

We investigate the limitation of an existing image-to-image translation model in our target scenario in terms of image quality. We use CycleGAN, a representative model for unsupervised image-to-image translation. We trained a model using 2467 daytime and 2376 nighttime images that we collected. We used the PyTorch code downloaded from the authors’ website with the same parameters in CycleGAN [1]. Given nighttime images, the trained model outputs translated images at $256 \times 256$ resolution. The experiment was
conducted using a computer with Intel i7-7700 CPU, NVIDIA GTX 1080Ti, and 32GB RAM, running Windows 10 (see Table 2).

Table 2. Experimental H/W setup.

|              |                  |
|--------------|------------------|
| CPU          | Intel i7-7700    |
| GPU          | NVIDIA GeForce GTX 1080Ti |
| RAM          | 32GB             |
| Compiler     | pyTorch (with CUDA(9.0) + cuDNN) |

Figure 2 shows three example nighttime images used as inputs and Figure 3 shows the corresponding outputs of translated images. As shown in the figure, the dark part of the input images mainly results in low quality translation. Such dark areas hardly have any information on geographical features, structures, and/or objects that actually exist in those areas. Thus, the translation model fails to provide high quality of translated images. As seen in the output images (Figure 3), the boundary between the sky and building or trees is not clear. A part of the building and trees are incorrectly translated to the blue sky in Figure 3c.

Figure 2. Examples of nighttime images. Images (a–c) were taken at night and used as input for translation by the CycleGAN model.

Figure 3. Examples of output images translated by the CycleGAN model. Images (a–c) correspond to (a–c) of Figure 2, respectively.

4. Proposed Solution

4.1. Overview

To increase the image quality, it is essential to address the problem of poor translation in the dark part of images. The part of the nighttime image that is dark and cannot be discerned has insufficient information that can be converted into a reasonable quality of
daytime image, so that the conversion is not properly performed. It is quite difficult, if not impossible, to transform such dark areas by only using the input image data.

A proposed idea in this work is to augment the obscured information in dark areas of input image frames using the past daytime images corresponding to the input. Such augmentation is possible since there are readily available daytime street view images in several cities and rural areas worldwide from street view services and we can use them.

Figure 4 shows the overall processing pipeline of Daydriex. It consists of two main steps: the data supplement processing step that supplements hint information for the dark areas of nighttime images and the daytime image generation step that translates the processed nighttime images to daytime images. The data supplement processing step retrieves daytime images from a street view service for input nighttime images and augments the input images with hint information from the street view images. Resulting images from the data supplement processing step are used as an input of the daytime image generation step, which generates daytime images corresponding to the input nighttime images through an image translation model.

Figure 4. Overview of the Daydriex processing pipeline including two main steps, data supplement processing and daytime image generation.

However, it is not straightforward to perform the aforementioned processing. We address several issues to realize our idea. Figure 5 shows the flowchart of detailed processing steps to realize the idea and implement the proposed processing pipeline. The hint image retrieval stage acquires a 360-degree street view image and its location information using the vehicle location coordinates. From the image, a panoramic image is constructed to provide hint images. The position matching stage finds a partial area in the panoramic image matching the input nighttime image with the vehicle location. The perspective matching stages crops the matched image area and performs perspective warping to derive a final hint image. The image merging stage generates a blended image with the hint image and input nighttime image, which is used as an input of the image translation stage. Finally, an output daytime image is obtained through image translation. In the following subsections, we present problems to tackle and detailed descriptions of our solutions as shown in Figure 5. Sections 4.2–4.5 present position matching, perspective matching, panoramic image construction for hint image retrieval, and image merging, respectively.
4.2. Obtaining Hint Images Matching Real-Time Locations

To provide hint information using the existing street view services (e.g., Google Street View (https://www.google.com/streetview/, accessed on 24 February 2021), Naver Road View (https://map.naver.com/, accessed on 24 February 2021), a street view image matching the current location and orientation is required. We can compare GPS coordinates to find a matching image for the current location. However, the street view image that exactly matches the current location may not exist since street view images are captured at regular intervals. Thus, it is necessary to obtain hint information to be supplemented to a nighttime image using the street view image closest to the current location.

Figure 6 shows an example of image taken by a camera on a car driving on a road at night (Figure 6a) and the corresponding street view image for the matching location (Figure 6b). As we can see in Figure 6b, the street view image is a 360-degree image and it has relatively wide field of view. Thus, only a part of the image that matches the current nighttime image should be used to provide hint information. Accordingly, it is necessary to find and crop such apart from an entire street view image.
We devise a method to pinpoint a part of street view image that matches an image captured at the vehicle location. The matching procedure is as follows: (1) Using the vehicle’s GPS coordinates, the street view image closest to the vehicle location is obtained. An example is specified in Figure 7. (2) The distance between the GPS coordinates of the vehicle (1) and the coordinates where the street view image was taken (2) is calculated (see Figure 7). The physical distance between the two coordinates is transformed to the pixel distance (3). (3) A window to crop the hint image (4) is moved from the middle of the street view image by the distance. (4) the part of image inside the window (4) is cropped.

**Figure 7.** Procedure of obtaining an image part to be used to provide hint information from a street view image.

4.3. Matching Perspective of Images

Obtaining a part of street view image that matches the current location is not sufficient for providing appropriate hint information. It is because a street view image has different perspective from that taken in a normal car. Such perspective difference mainly results from the fact that street view images are captured using a specialized equipment and vehicle as shown in Figure 8a. The camera that takes street view images is installed at a high position as shown in the figure, which causes a big difference in the vertical position where a camera can be mounted in a normal car that we consider for our target scenario.

To make the perspective of street view image match that of nighttime image input, it is necessary to perform perspective transformation. For the purpose, we need to use a trapezium-shaped window to crop the part to be used as hint, not a rectangle-shaped window shown in Figure 7. Figure 8b shows an example of trapezium-shaped window overlaid on a street view image. We empirically determine the size and initial position of the window. We apply perspective transformation to the cropped image using the window. We use OpenCV library to perform perspective transformation.
Figure 8. A vehicle to capture street view images (Photo from Pixabay) (a) and an example of trapezium-shaped window overlaid on an example street view image (b).

Figure 9 shows an example result of perspective transformation. An input nighttime image is shown in Figure 9a and the corresponding street view image is shown in Figure 9b. To compare the difference between the perspective-transformed image and the one without transformation, we use two different cropping windows, the red trapezium-shaped window and the yellow rectangular window as shown in Figure 9b. Two hint images from the windows are also shown, one is cropped using the yellow window without perspective transformation (Figure 9c) and the other is cropped using the red window and perspective-transformed (Figure 9d). We can check the difference between the two hint images based on the area (blue rectangle) that is bright and visible in the input nighttime image. In Figure 9c, the streetlight that exists in the right side of the input image is not shown and the position of the tree in the center of the input image is slightly different. In contrast, Figure 9d shows the streetlight and the tree that are positioned similarly to the input image, which means that perspective transformation results in more appropriate hint images.

4.4. Filling a Missing Area

In the process of cropping a street view image for hint information, a partial area of the cropped image may be missing. Main reasons are two-fold: First, the street view images are not continuous; we find that they are often captured at 10-meter intervals on average or sometimes even longer. Second, the window to obtain a hint image is a trapezium-shaped as shown in Figure 8b.

Figure 10 shows an example case that a missing area can be included in the cropped image. Assume that a car is currently moving past the center coordinates of the i-th street view image. The hint image is cropped as the window for the current location moves in the moving direction of the car. However, if the car moves more than a certain distance from the center coordinates of the street view image, a hint image cropped by the window can include some empty area as shown in Figure 10. In this case, hint information for the area cannot be provided.

To address the problem, we adopt a panoramic image construction method. Since street view services usually provide a 360-degree view of image at a location, we can obtain an image with a view parallel to the road (i.e., front or rear view) as well as an image with a view to the roadside. To fill a missing part from a roadside view image at a certain location, we stitch two images, a side view image and a front (or rear) view image obtained at the same location to construct a panoramic image. Then, we crop a part of image from the constructed panoramic image to get a hint image without any missing area. For panoramic image construction, we use a technique proposed in [27].
Figure 9. An example to show the results of perspective transformation with two different windows. (c,d) show that the red trapezium-shaped window is better than the yellow rectangular window to provide more appropriate hint images.

Figure 10. An example case showing that some part of hint image is missing. This can happen when the distance between the vehicle location and the street view image location is larger than a certain range.
4.5. Supplementing Hint Information

Once a final hint image is obtained for an input nighttime image, it is necessary to supplement hint information to a dark area of the input. In this work, we take a simple approach, merging two images in a pixel-wise manner. A merged image is obtained by calculating weighted averages of corresponding pixels from the two images.

In the merging process, the translated image quality is affected by the weights assigned to the two images, an input nighttime image and a hint image. Since hint images are obtained from the street view images that were captured in the past, they might include some objects such as cars and people that happened to be in the field of view of camera. If the weight of the hint image is too high, the object that does not exist in the input nighttime image but exists in the street view image can appear faintly in translated images, thereby resulting in deterioration in translation quality. On the other hand, if the weight of the hint image is low, information cannot be properly provided for the dark area of the nighttime image. Thus, it is difficult to obtain good quality of translated images.

To determine appropriate weights for image merging, we examine different sets of weights and their effect on translated image quality. Figure 11 shows an example of translated results depending on weights. As the weight of the hint image increases (Figure 11c–h), it can be seen that the car in the red square area is gradually revealed. This means that objects that exist only in the hint image have a significant effect on the translated image. As a result, the quality of the image is deteriorated. In addition, if the weight of the image is too low, hint information for the dark area of input image is hardly supplemented. As a result, it can be seen that the boundary between the tree branch and the sky is not clearly distinguished, resulting in poor image quality. To achieve reasonable quality, we empirically set the weights to 0.8 for an input nighttime image and 0.2 for a corresponding hint image, respectively.

![Figure 11. An example of translated results depending on merging weights (Input: Hint). The weights affect the quality of translated results.](image)

5. Evaluation

We evaluate Daydriex in terms of image quality with two methods. First, we use the Fréchet Inception Distance (FID) \[6\], which is widely used to measure GAN performance. Second, we conduct a user perceptual study with 30 people to assess the reality of Daydriex.
outputs. We compare our results with those of CycleGAN. In addition, we measure computation time taken to run main processing stages of Daydriex.

5.1. Setup

To evaluate the technique proposed in this paper, we collected nighttime images as well as corresponding GPS data. For data collection, we implemented a mobile application and deployed on the roof of the car we used for driving. With the setup, we captured images while driving at night in three driving environments: rural (slow), rural (fast), and urban. To obtain hint information for our method, we used Google Street View images corresponding to the captured nighttime images with the GPS data.

For each driving environment in Table 3, images converted using CycleGAN and Daydriex were prepared. As a result, the experiment was conducted with a total of 9 sets of images (3 sets of original images at night and 6 sets of images converted to daytime). Figure 12 is an example of the captured nighttime images (Figure 12a,d,g), the converted images using CycleGAN (Figure 12b,e,h) and the converted images using our method (Figure 12c,f,i). For image translation, we used the same HW setup and CycleGAN model mentioned in Section 3.

To examine FID of translated results, we collected daytime videos recorded while driving along the same roads mentioned above; total length of videos is about 90 min. From these videos, we obtained daytime images which do not overlap each other and grouped them into the three cases corresponding to those in Table 3. We use these groups of images to obtain FID scores of daytime images translated by Daydriex and CycleGAN, respectively.

Also, we conducted the user perception evaluation with 30 participants (ages between 21 and 33, mean: 24.4). We showed them videos made by using translated images and asked them to answer a 7-point Likert scale question as shown in Table 4. Each participant watched total 6 videos, two for each case including one by CycleGAN and the other by our method, and answered the question for each video.

Table 3. Three cases for nighttime image collection.

| Case          | Description                                    |
|---------------|------------------------------------------------|
| Case 1: rural (slow) | Slow driving in a rural area (avg. 20 km/h) |
| Case 2: rural (fast)    | Fast driving in a rural area (avg. 60 km/h)    |
| Case 3: urban         | Driving in an urban area                        |

Table 4. 7-point Likert scale question used for evaluation.

| I Think This Video Shows Natural Daytime Scenes | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|------------------------------------------------|---|---|---|---|---|---|---|
|                                                | Strongly Disagree | Disagree | Slightly Disagree | Neutral | Slightly Agree | Agree | Strongly Agree |

To evaluate the performance of Daydriex, we measure how much time is required for the main operations of the processing pipeline shown in Figure 4. For the purpose, we randomly select 10 nighttime images of case 1 in Table 3 and measure the time taken to translate them into daytime images. We report an average of the measured computation times.
Figure 12. Examples of captured nighttime images (a,d,g), translated images by CycleGAN (b,e,h), and translated images by Daydriex (c,f,i). (b,c), (e,f), and (h,i) are the translations for (a), (d), and (g), respectively.

5.2. Result

Figure 13 shows FID scores of translated images for the three cases by our proposed method, Daydriex, and CycleGAN. Daydriex shows lower FID scores for all the cases, which means translated images by our method are better than those by CycleGAN. For the first case, i.e., Rural (slow), the FIDs of CycleGAN and Daydriex are 232.12 and 229.06, respectively. For the second case, Rural (fast), they are 223.18 and 213.58, respectively. The third case shows the largest difference among the three cases; the FIDs are 248.21 (CycleGAN) and 222.15 (Daydriex), respectively.

Moreover, Figure 14 shows that our proposed method achieves the higher ratio of positive responses (over 4) than CycleGAN for all three cases. Our results get 60%, 80%, and 57% of positive responses for the slow rural, fast rural, and urban case, respectively. In contrast, the CycleGAN results only get 17%, 13%, and 10% of positive responses for those cases. Figure 15 shows the box plots of the Likert scale responses for the three cases. For the slow rural and urban cases, the medians of CycleGAN and ours were 2 and 5, respectively. For the fast rural case, they were 2 and 6, respectively.
Figure 13. Fréchet Inception Distance (FID) for the three cases. Daydriex shows lower FID scores for all the cases.

Figure 14. Diverging stacked bar charts of the Likert scale responses for the three cases. Daydriex gets more positive responses regarding the perceptual quality of translated images.
Figure 15. Box plots of the Likert scale responses for the three cases. Daydriex achieves higher median scores than CycleGAN.

To evaluate the difference in the responses, we ran a Wilcoxon Signed-rank test. We found that there is a significant difference for the all three cases: the slow rural case ($W = 4$, $Z = -4.73$, $p < 0.05$, $r = 0.61$), the fast rural case ($W = 0$, $Z = -4.81$, $p < 0.05$, $r = 0.62$), and the urban case ($W = 27$, $Z = -4.13$, $p < 0.05$, $r = 0.53$). From these results, we can see that the results translated by our proposed method were perceived more natural daytime scenes.

Table 5 shows the result of the average computation time required for the main processing stages. The time varies depending on the stages. Please note that the result for the hint image retrieval stage only includes the time of operations needed to generate a panoramic image without the time taken to download an image from the street view service. Although the hint image retrieval takes the longest time, 710 ms, the position mating, perspective matching, and image merging stages take about 9 ms, 2 ms, and 0.02 ms, respectively. The size of front and side images used to construct panoramic images is $1280 \times 720$ pixels. In the case of image translation, the computation time taken is different depending on the image resolution. It takes 6 ms, 14 ms, and 53 ms at the resolutions of $128 \times 128$, $256 \times 256$, and $512 \times 512$, respectively. Please note that we target these resolutions for image translation since those are mainly used in the previous image-to-image translation works including CycleGAN.

### Table 5. Average computation time taken for the main processing stages. Please note that the result for the hint image retrieval stage only includes the time of operations needed to generate a panoramic image without the time taken to download images from the street view service.

| Hint Image Retrieval | Position Matching | Perspective Matching | Image Merging | Image Translation |
|----------------------|-------------------|---------------------|--------------|------------------|
| 710 ms               | 9 ms              | 2 ms                | 0.02 ms      | 6 ms ($128 \times 128$) |
|                      |                   |                     |              | 14 ms ($256 \times 256$) |
|                      |                   |                     |              | 53 ms ($512 \times 512$) |

6. Conclusions

In this paper, we propose Daydriex, a processing pipeline that makes nighttime scenes look like daytime scenes in the context of driving on a road. Dark area in nighttime scenes have insufficient information, thereby resulting in poor translation quality even with widely used image-to-image translation models such as CycleGAN. To address the problem, Daydriex employs a processing stage to use street view images as hint information. We present a technique for constructing hint images corresponding to the location of vehicle.
Our evaluation shows that Daydriex achieves better performance than using CycleGAN only in terms of the FID score and user perception.

We have several topics of future work. First, we will test and compare a range of translation models to generate daytime images in our processing pipeline, which currently uses CycleGAN. Considering the fidelity of translated results and computation time, we can find the best performing model. In addition, it would be possible to design a new translation model specific to the nighttime-daytime image translation task in the proposed pipeline, which we leave as a future work. Second, it is an important future work to develop a prototype system for deployment and conduct field tests. Our study to develop the processing pipeline was based on pre-collected data. However, to deploy a working prototype in actual driving situations, we need to address several other issues such as real-time data acquisition, response time, user interaction, and so on. Third, it is necessary to achieve a certain level of resolution and frame rate for image translation to meet users' quality requirements. This would require high computing power. We will study potential solutions such as cloud/edge-assisted computation and caching to address the problem.

Author Contributions: Conceptualization, E.L. and S.K.; methodology, E.L.; software, E.L.; validation, E.L. and S.K.; investigation, E.L. and S.K.; resources, E.L.; writing—original draft preparation, E.L. and S.K.; writing—review and editing, S.K.; visualization, E.L. and S.K.; supervision, S.K.; project administration, S.K.; funding acquisition, S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was in part supported by the National Research Foundation of Korea (NRF) grant (No. 2020R1I1A3060675). This work was also supported in part by the Education and Research Promotion Program of KOREATECH (2019).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. 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, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
2. Liu, M.Y.; Breuel, T.; Kautz, J. Unsupervised image-to-image translation networks. In Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 700–708.
3. Huang, X.; Liu, M.Y.; Belongie, S.; Kautz, J. Multimodal unsupervised image-to-image translation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 172–189.
4. Ma, L.; Jia, X.; Georgoulis, S.; Tuytelaars, T.; Van Gool, L. Exemplar-guided unsupervised image-to-image translation with semantic consistency. arXiv 2018, arXiv:1805.11145.
5. Lee, E.; Kang, S. Towards the Experience of Daytime Driving at Night (poster). In Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services, Seoul, Korea, 17–21 June 2019; pp. 637–638.
6. 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 Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 6626–6637.
7. Wang, R.; Zhang, Q.; Fu, C.W.; Shen, X.; Zheng, W.S.; Jia, J. Underexposed Photo Enhancement Using Deep Illumination Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
8. Ren, W.; Liu, S.; Ma, L.; Xu, Q.; Xu, X.; Cao, X.; Du, J.; Yang, M.H. Low-light image enhancement via a deep hybrid network. IEEE Trans. Image Process. 2019, 28, 4364–4375. [CrossRef] [PubMed]
9. Chen, C.; Chen, Q.; Xu, J.; Koltun, V. Learning to see in the dark. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3291–3300.
10. 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 2019, arXiv:1906.06972.
11. Yang, W.; Wang, S.; Fang, Y.; Wang, Y.; Liu, J. From Fidelity to Perceptual Quality: A Semi-Supervised Approach for Low-Light Image Enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–18 June 2020.

12. Cai, J.; Gu, S.; Zhang, L. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Trans. Image Process.* 2018, 27, 2049–2062. [CrossRef] [PubMed]

13. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the 28th Annual Conference on Neural Information Processing Systems 2014, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.

14. Kim, T.; Cha, M.; Kim, H.; Lee, J.K.; Kim, J. Learning to discover cross-domain relations with generative adversarial networks. *arXiv* 2017, arXiv:1703.05192.

15. Chen, Y.; Lai, Y.K.; Liu, Y.J. Cartoongan: Generative adversarial networks for photo cartoonization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 9465–9474.

16. Zheng, Z.; Yang, X.; Yu, Z.; Zheng, L.; Yang, Y.; Kautz, J. Joint discriminative and generative learning for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 2138–2147.

17. Tomel, M.; Cornia, M.; Baraldi, L.; Cucchiara, R. Art2real: Unfolding the reality of artworks via semantically-aware image-to-image translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5849–5859.

18. Anokhin, I.; Solovev, P.; Korzhenkov, D.; Kharlamov, A.; Khakhulin, T.; Silvestrov, A.; Nikolenko, S.; Lempitsky, V.; Sterkin, G. High-Resolution Daytime Translation Without Domain Labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 7488–7497.

19. Anoosheh, A.; Sattler, T.; Timofte, R.; Pollefeys, M.; Van Gool, L. Night-to-day image translation for retrieval-based localization. In Proceedings of the IEEE 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 5958–5964.

20. Anoosheh, A.; Agustsson, E.; Timofte, R.; Van Gool, L. Combogan: Unrestrained scalability for image domain translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 783–790.

21. Maintz, J.A.; Viergever, M.A. A survey of medical image registration. *Med. Image Anal.* 1998, 2, 1–36. [CrossRef]

22. Kahaki, S.M.; Wang, S.L.; Stepanyants, A. Accurate registration of in vivo time-lapse images. In Proceedings of the SPIE 10949, Medical Imaging 2019: Image Processing, San Diego, CA, USA, 15 March 2019.

23. Kahaki, S.M.M.; Nordin, M.J.; Ashtari, A.H. Incident detection algorithm based on radon transform using high-resolution remote sensing imagery. In Proceedings of the IEEE 2011 International Conference on Electrical Engineering and Informatics, Bandung, Indonesia, 17–19 July 2011; pp. 1–5.

24. Lin, S.F.; Chen, J.Y.; Chao, H.X. Estimation of number of people in crowded scenes using perspective transformation. *IEEE Trans. Syst. Man Cybern. Part Syst. Hum.* 2001, 31, 645–654.

25. Jiang, Y.; Gao, F.; Xu, G. Computer vision-based multiple-lane detection on straight road and in a curve. In Proceedings of the IEEE 2010 International Conference on Image Analysis and Signal Processing, Zhejiang, China, 9–11 April 2010; pp. 114–117.

26. Wang, C.; Wang, N.; Ho, S.C.; Chen, X.; Song, G. Design of a new vision-based method for the bolts looseness detection in flange connections. *IEEE Trans. Ind. Electron.* 2019, 67, 1366–1375. [CrossRef]

27. Brown, M.; Lowe, D.G. Recognising Panoramas. In Proceedings of the Ninth IEEE International Conference on Computer Vision ICCV ’03, Nice, France, 13–16 October 2003; Volume 2, p. 1218.