COMP4471 Project Final Report
Day-night Image Transformation Using CycleGAN and D2-net

CHENG, Ting-kai
HKUST
tchengad@connect.ust.hk

HSU, Chia-hong
HKUST
chsuae@connect.ust.hk

WU, Chi-hsuan
HKUST
cwuau@connect.ust.hk

Abstract

Illumination variability is arguably one of the most crucial factors in image matching, a task in which aligning structure, pattern, and content between photos remain a hurdle to overcome. A sharp difference in illumination between images can significantly compromise the matching performance, in particular the matching of the images taken at day and those at night. In this study, we propose a GAN-based, image-to-image translation network to tackle the day-night feature matching challenge. Our modified CycleGAN model transforms day images to night ones and vice versa, followed by analysis using D2-net descriptors. By comparing our modified CycleGAN model to the vanilla CycleGAN and the OpenCV models, the proposed model produces greater image quality and delivers better performance on feature matching.

1. Introduction

Matching corresponding structure, pattern, and content between two or more images has been a fundamental hurdle to overcome in the realm of computer vision. Image matching is typically applied to align scenes from different perspectives, locations, and even time points. Its application ranges from image stitching in panoramic pictures to tracking and modeling 3D objects, pose estimation, robot localization, and mapping, etc. In addition to the traditional approach, improvements in key feature repeatability and matching quality are attributable to deep-learning-based approaches. However, it has indicated that the matching performance can be severely tampered given the illumination difference between images taken at day and those at night [1]. In our study, we propose a GAN-based, image-to-image translation network, which serves as an image preprocessing tool to tackle the day-night feature matching problem.

1.1. Day-night Image Matching

Illumination variability remains one of the most crucial issues influencing the matching performances. Zhou et al. have experimented with several traditional detectors and descriptors methods on day-night image matching [2]. Their work suggests a potential improvement in both feature detectors and descriptors under drastic illumination changes. With the current progress in deep-learning networks, convolutional neural networks (CNN) are extensively implemented to extract visual features and background information of an image. So far, D2-Net, proposed by Dusmanu et al. [3], serves as one of the most advanced networks in day-night visualization. Such CNN architecture merges the detection and description process for image matching, which enhances the detection of feature points. The algorithm allows more accurate examination among images, particularly those with varying illumination conditions.

1.2. GAN-based Image-to-image Transformation

With attribution to CNN, various generative adversarial networks (GAN) have been proposed to perform image-to-image translation by encoding and decoding the feature information from different domains. In particular to day-night image matching, it has shown that feature descriptors are subject to non-uniform light sources in night images and influenced by over-exposure in day images. In this regard, we aim to specifically tackle the problem using CycleGAN [4], a renowned GAN-based image-to-image translation network, as a preprocessing tool prior to the image matching task. Such architecture will be utilized to transform day images to night ones and vice versa, and then we will compare them using D2-net descriptors. As a result, we anticipate that the feature matching quality of day-night images could be further enhanced by the proposed method.

2. Related Work

The task of image matching has engaged substantial interest over the past years. In 2017, "Image-to-Image Translation with Conditional Adversarial Networks" [5] ad-

dressed this problem by implementing the conditional GAN alongside the typical generator and discriminator models, named after Pix2Pix. As we condition on an input image and subsequently generate an output image, it allows remarkable specification of the generator, the discriminator, and the model optimization. On that account, this approach renders Pix2Pix one of the most renowned methods to refine the image-to-image translation tasks. Nevertheless, the network requires paired datasets, which may not be available across different realms of research.

Aside from the Pix2Pix network, other researchers have also discussed the cross-domain relationships in their literature. In 2017, T. Kim et al. introduced the DiscoGAN network [6], which succeeded in style transfer across different domains and ensured key attribute preservation. Without the necessity of dataset pairing, this method greatly reduces the costly computational power, a problem present in other networks requiring images with extra labels. Both CycleGAN and DiscoGAN networks adopt two cross-domain transformation functions and are built on the idea of reconstruction loss. The main difference between the two lies in how such losses are utilized within the architectures. In short, DiscoGAN uses two L2 losses for each domain; CycleGAN, in contrast, applies single L1 cycle consistency loss, between the input images and the reconstructed ones.

While the above-mentioned networks only tackle two domain image translation, multimodal image transformation requires independent models for each pair of image domains. In 2018, StarGAN was presented by Y. Choi [7], entailing only one single model for multiple domains. With inputs and outputs residing in variational domains, StarGAN allows concurrent training of several datasets within the same network, which has been proven efficacious in dealing with such proposition [8, 9].

In addition to a wide variety of GAN networks, OpenCV, on the other hand, plays a crucial role in real-time image processing. Despite a simpler and less disparate transformation, the open-source library for computer vision enables faster image transformation than most machine learning and deep learning models. In this project, we will implement it as a comparison with the generated images to cross-check our model's performance.

3. Dataset

The datasets utilized in this project are the 24/7 Tokyo dataset [10], the Aachen Day-Night dataset [11], as well as the street view image taken in Hong Kong. The 24/7 Tokyo dataset is composed of paired day-time and night-time street-view photos [10], while the Aachen Day-Night dataset depicts the city view of Aachen, Germany, with 20 relevant day-time pictures for each of the 98 night-time query images [11].

Prior to the training in CycleGAN, we first resize photos from the 24/7 Tokyo dataset and those taken in Hong Kong (800 pixels as the narrow side). For the purpose of data augmentation, these pictures are later randomly cropped into patches of 256 pixels times 256 pixels, with a subtotal of 1626x2 training images and 160x2 validation images for day-to-night and night-to-day transformations. They are subsequently applied to train the CycleGAN network, which then learns the mapping in both directions. Specifically, the input of the network would be pairs of day-night images, whereas the output is the feature point correspondences of the given two pictures.

Once the network has learned how to conduct the transformation, the Aachen dataset of day-night paired images would be fed into the D2-net [3], where a further evaluation of the proposed architecture's feature matching performance is carried out.

4. Method

In this section, an overview of our algorithm's pipeline is provided. In addition, the implementation details of each stage will be discussed individually.

4.1. Overview

Improvements in image matching quality between day-night photos are the primary focus of our study. Compared to matching images in the same domain, it has been shown in previous works [1–3] that the quality of direct matching day-night images is reduced. Therefore, an intuitive algorithm that performs day-to-night (or night-to-day) transformation is proposed in this study, in which the transformed images and the other night/day photos will serve as the input for the downstream matching task.

For the image transformation stage, a CycleGAN model is selected for training (with some modifications, please refer to section 4.2) since it is known for learning style transferring between images. Its strength is exemplified by mapping horses to zebras, Van Gogh's painting to Monet's, and, in our case, day images to night images. Our CycleGAN model will be trained on the day images and night images from the 24/7 Tokyo dataset [10]. As a result, it will adopt day-to-night and night-to-day style translators. For the feature matching algorithm, the pre-trained D2-net model is chosen, given its excellent performance in day-night localization tasks. The Aachen Day-Night dataset [11] is used for our final evaluation, where a night query image is paired with a transformed-day image and a transformed-night query is paired with a day image. In the end, the results from the two pipelines will be aligned, and those with a matching score above a given threshold will be retained.

4.2. CycleGAN with identity loss

Image-to-image translation is the task of transforming images from one domain to another. In doing so, the im-

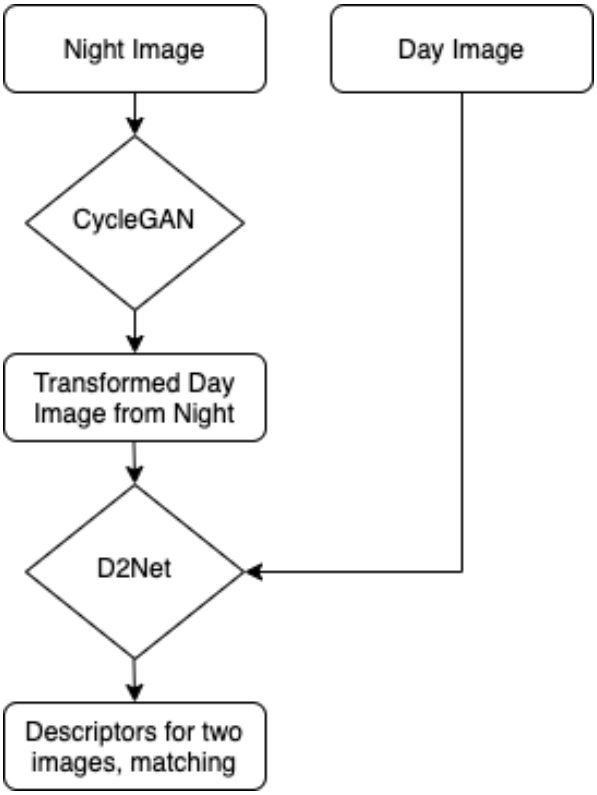


Figure 1. Pipeline of the algorithm

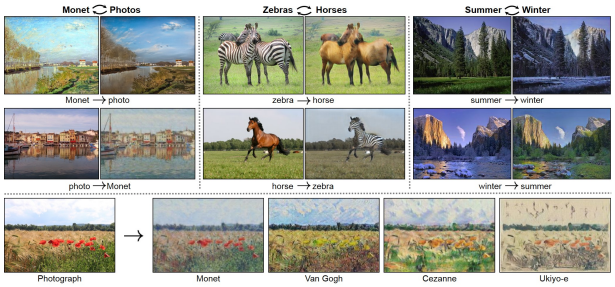


Figure 2. Examples of CycleGAN transformation

ages adopt styles or characteristics of images from another domain. Our architecture is shown in Figure 4. In the generator part (Fig. 4 (top two rows)), the input image first passes through an encoder stage which consists of three convolutional layers. Subsequently, it undergoes several residual blocks to refine the feature learning. Finally, the feature passes through a series of upsampling layers and generates the transformed image. The discriminator (Fig. 4 (bottom row)) analyzes the input image and classifies it into day or night.

As shown in Figure 5, G_{day} and G_{night} represent the mapping from night to day and from day to night, respectively. In the CycleGAN architecture, we also have D_{day}

Figure 3. D2-net feature matching

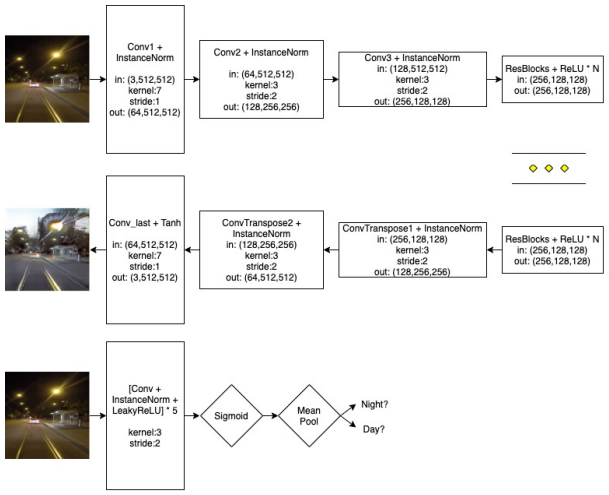


Figure 4. Proposed modified CycleGAN architecture

and D_{night}, the two discriminators that discriminate between the styles of images, i.e., classify the image into "day style" and "not day style", and vice versa for night style.

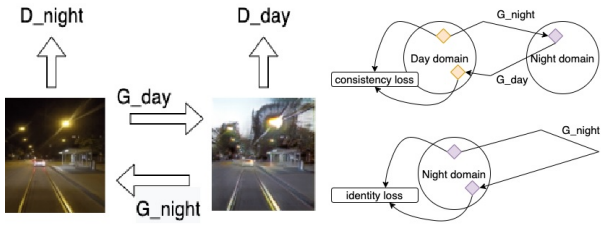


Figure 5. CycleGAN generators, discriminators, and losses

The loss consists of three parts: the adversarial loss, the consistency loss, and the identity loss. Similar to other GAN networks, the adversarial loss is a two-man game scheme where the generator aims to deceive the discriminator. The cycle consistency loss is an L1 distance function that computes the difference between the original input image x and the transformed then inverted image $F(G(x))$.

The identity loss is the difference between the original input image x and the transformed image $G(x)$ that maps to its own domain.

For consistency loss, it is to prevent the transformation function G_{day} and G_{night} from overfitting the data. Without the consistency loss, a generator G could potentially seize an image from the database of night images and transform all of the day images to a single night image in the database. In other words, consistency loss helps preserve the content of the image during style transformation. This is completed by computing the L1 loss of the inverse-transformed image and the input image. In addition, we define an identity loss as the L1 difference between the input image and the mapped image to its own domain. The identity loss increases the mapping complexity of the generator by feeding images from different domains. In our own experiments, it is also shown that the identity loss significantly raises our training efficiency. For more details regarding the experiment results, please refer to section 5.

4.3. D2-net

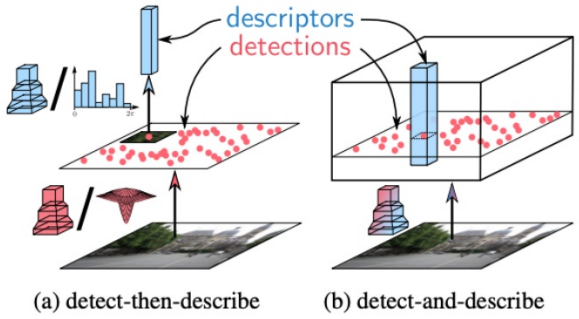


Figure 6. Comparison between traditional (left) training pipeline and D2-Net (right)

Traditional feature matching algorithms consist of two parts: the detecting stage and the describing stage. In detail, we will use SIFT to locate feature points and derive their descriptors by constructing a histogram, which counts the contributions of several gradient bins. This pipeline is called "detect-then-describe". D2-net is the first to propose a "detect-and-describe" pipeline that simultaneously finds the feature points and their descriptors using CNN.

The core idea behind detecting features and descriptors is to iterate each pixel and distinguish whether they are feature points one by one. For each pixel, there exists at least one channel that has the maximum value for all the feature channels encoded. Next, we evaluate if that pixel is the regional maximum in that feature channel map. If the pixel is the local maximum, it is considered a feature point. However, identifying the hard maximum is not practically trainable. Instead, the author utilizes a softmax feature detection

$$(i,j) \text{ is a detection } \iff D_{ij}^k \text{ is a local max. in } D^k \text{ with } k = \arg\max_t D_{ij}^t \ .$$

$$\alpha_{ij}^k = \frac{\exp\left(D_{ij}^k\right)}{\sum_{(i',j')\in\mathcal{N}(i,j)} \exp\left(D_{i'j'}^k\right)} \qquad \beta_{ij}^k = D_{ij}^k \Big/ \max_t D_{ij}^t$$

$$\gamma_{ij} = \max_k \left(\alpha_{ij}^k \beta_{ij}^k\right) \qquad s_{ij} = \gamma_{ij} \Big/ \sum_{(i',j')} \gamma_{i'j'}$$

Figure 7. Softmax score (s_{ij}) for each pixel

by assigning scores for each pixel. The followings are the mathematical formulations regarding how these scores are computed.

5. Experiment

5.1. Data Preprocessing

Before using the training images to update the weight of the model, images are resized to 256 pixels times 256 pixels, with the value of each pixel tuned between -1 and 1. The rationale for a range of -1 to 1 is to match the value range of generated images, given a Tanh activation function being the last layer of the generator. On the other hand, for the image size, the size of 512 x 512 was originally implemented; nonetheless, blurry structures and edges of the resulting images were frequently observed. Through several trial-and-errors, images resized to 256-pixel for both height and width are eventually selected.

5.2. Training

We use Pytorch to build and train our models with the Tesla P100 GPU. In total, two models are trained, i.e., the CycleGAN model trained without identity loss (vanilla CycleGAN) and that with such loss. The Adam optimizer is applied for both models, and details of the hyperparameters along with the loss curves are provided in Table 1.

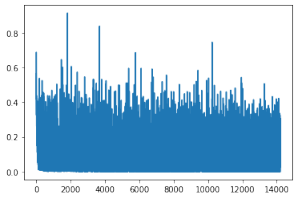
To stabilize the training process, we choose a small learning rate and experiment with different batch sizes. Eventually, the batch size with value 1 yields a more robust color transformation compared to values 4 and 8 with instance normalization. The weight for the consistency loss is designated carefully since it could significantly influence the performance of feature matching. Values of 1, 5, and 10 are experimented with for the consistency loss; ultimately, we decide on a larger value of 10 to emphasize the importance of details in the structure. During the training process, it is also observed that a larger weight for identity loss could more successfully preserve color transformation from dark to bright.

The training processes are halted after 16,000 iterations for vanilla CycleGAN and after 40,000 iterations for the Cy-

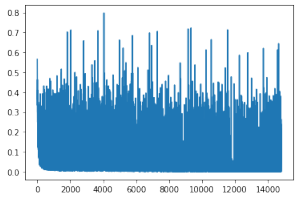
	Vanilla CycleGAN	Our proposed model
Learning rate	5e-4	5e-4
Batch size	1	1
Adam beta1	0.5	0.5
Adam beta2	0.999	0.999
Identity loss weight	0	10
Consistency loss weight	10 with L1 loss	10 with L1 loss
Adversarial loss weight	1	1
Number of iterations	16000	40000

Table 1. Hyperparameter Specification

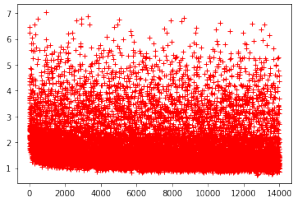
cleGAN model with identity loss. Both models have maintained the performance on the validation set within 5,000 iterations from the ceased iteration. Figure presented below only shows the result up to 16,000 iterations.



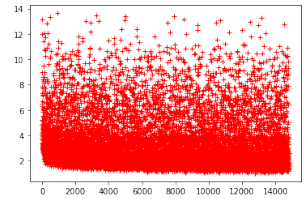
(a) Discriminator loss: real image CycleGAN



(b) Discriminator loss: real image CycleGAN with identity loss



(c) Adversarial loss: day-night-day CycleGAN



(d) Adversarial loss: day-night-day CycleGAN with identity loss

Figure 8. Discriminator and adversarial loss

5.3. Evaluation

To evaluate the performance of our model, night images are arranged in pairs and imported into the night-to-day generator. D2-net is then utilized to perform feature matching of the original day images (from the image pairs) and the generated fake day images. Given an image pair, the pretrained D2-net model could generate (1) a set of matched features, (2) the location of the feature points, and (3) the matching score of each feature match, which rates the confidence level of D2-net. We use both the matching score and the number of matching to evaluate the effectiveness of the model. In terms of the dataset, the Aachen Day-Night dataset is chosen for evaluation, with 98 image pairs in total.

We compare our model performance with a vanilla CycleGAN model, which was trained on the same dataset as our proposed model, as well as OpenCV, by tuning the gamma value of the image. The results of D2-net are evaluated from four different inputs, including (1) the original image, as well as those generated by (2) our proposed model, (3) vanilla CycleGAN, and (4) OpenCV.

It is also perceived that the pretrained D2-net could detect some minor pixel value change even when such values are too insignificant to be detected by human eyes. Therefore, some low-score matchings generated by D2-net are uninterpretable. To avoid such counterintuitive matching, a threshold of scores above 1000 is also experimented with. Given this limitation, a more robust matching result could be acquired.

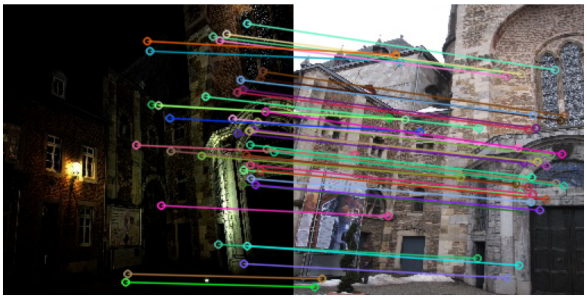


Figure 9. Bad matching demonstration (matching with low scores)

5.4. Result

We match the performance of the night-to-day transformation methods using three sections: image quality, feature matching performance without threshold, and feature matching performance with a score threshold of 1,000. The results suggest that in comparison with the vanilla CycleGAN, our proposed model has a greater image quality and delivers a better performance on feature matching with a score threshold.

5.4.1 Image Quality

When we analyze the maintained street-view structure and the clearness of the edges in the images from two CycleGAN models and OpenCV, a general pattern shows that OpenCV performs the best while CycleGAN trained with identity loss conserves most details. This is reasonable because image processing using OpenCV only tunes the gamma value without altering any structural information. Regarding the structural quality of our proposed CycleGAN, details such as brick patterns and windows could be clearly identified in the images. Minor problems for this model could be the missing color information in some results. Concerning the performance of vanilla CycleGAN,

blurry structures and edges are displayed in the images, and at times, straight edges are generated to be curved.

In examination of the color transformation of different methods, it has been observed that the CycleGAN model trained with identity loss can properly transform the night image into a day image under most conditions; occasionally, however, images may be transformed into light brown sky color. In regard to the performance of vanilla CycleGAN, it lacks consistency in color transformation, and color themes are often unpredictable. Lastly, for the OpenCV method, the images are simply tuned brighter and no daylight features could be detected.

(a) Original dark image (b) Image generated by OpenCV

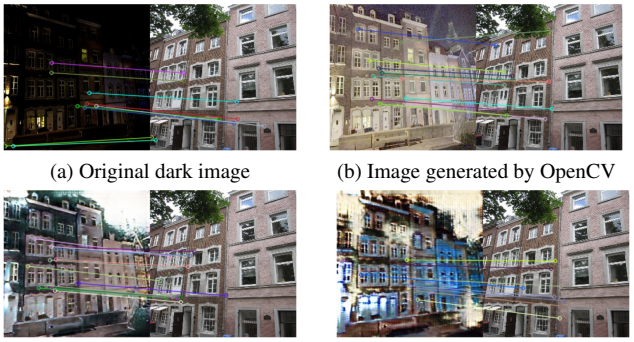
(c) Image generated by the proposed model (with identity loss) (d) Image generated by vanilla CycleGAN

Figure 10. Generated image comparison

5.4.2 Feature Matching without Threshold

We carry out experiments with 3 different generated image pairs and the original pair by D2-net, among the 98 pairs. Resultantly, images generated by CycleGAN model trained with identity loss yields the largest number of feature matching in 48 pairs. On the other hand, the OpenCV method receives 43 pairs in terms of the greatest number of feature matching, whereas the remaining 7 pairs have the most number of feature matching from the vanilla CycleGAN method. The feature matching sample could be found in Figure 11.

The average scores for the feature matching are 1179.9824, 1025.648, and 923.604 for CycleGAN with identity loss, OpenCV method, and vanilla CycleGAN, respectively. The average score for feature matching on the original image pairs is 1036.8812. As for the number of feature matching per image pair, on average 17.54 features could be matched with images generated from CycleGAN with identity loss, while the OpenCV method has an average of 16.72 features and CycleGAN 13.43. The average feature matching found on the original image pairs is 12.2.



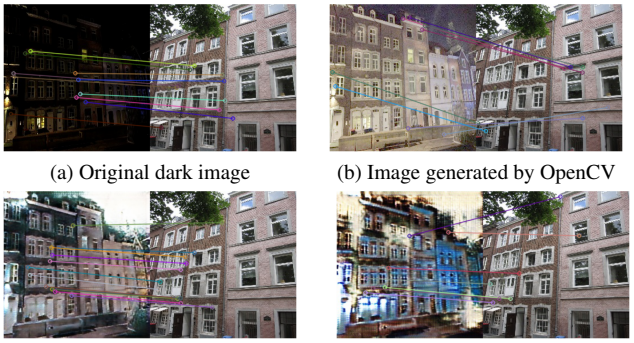
(a) Original dark image (b) Image generated by OpenCV

(c) Image generated by the proposed model (with identity loss) (d) Image generated by vanilla CycleGAN

Figure 11. Feature matching comparison (without threshold)

Despite the fact that images from CycleGAN method with identity loss and OpenCV method are identified with similar numbers of feature matching, average scores of feature matching for CycleGAN-generated images are higher than that from the OpenCV method.

5.4.3 Feature Matching with Threshold



(a) Original dark image (b) Image generated by OpenCV

(c) Image generated by the proposed model (with identity loss) (d) Image generated by vanilla CycleGAN

Figure 12. Feature matching comparison (with threshold)

To make the feature matching on image pairs more robust and intuitive, we apply a threshold of 1,000 to the feature matching scores. This is because the average scores of feature matching for various transformation methods are all approximately 1,000. After the experiment, 77 out of 98 (78.5 percent) image pairs have the largest number of feature matching (with scores above 1,000) using CycleGAN

with identity loss, while only 20 out of 98 (20.4 percent) of image pairs have the largest number of feature matching by the OpenCV method. The result could be seen in Figure 12.

On average, the CycleGAN method with identity loss delivers 12.4 feature matching (with a score above 1,000) and the OpenCV method conveys 9.1 feature matching. Regarding the performance of vanilla CycleGAN, only 6.4 feature matching (with a score above 1,000) is detected per image.

From the 3 sections above, we reach the understanding that the OpenCV method is most effective in maintaining the structural details; nevertheless, the CycleGAN method conserves sufficient details while minimizing the illumination difference. This results in a larger number of feature matching and also a higher score for each matching generated from D2-net.

6. Conclusion

Our study examines the effectiveness of style transformation among images for day-night feature matching. To increase the quality of feature matching for images from distinct domains, we design a unique algorithm pipeline that applies a GAN network to translate images to other domains, prior to matching the corresponding feature descriptors. We have built and trained our model based on the CycleGAN model for image transformation. It is discovered that the addition of the identity-loss function accelerates the training process and improves the translating results.

Our experiment results demonstrate that GAN-based day-night translation enhances the feature matching quality, in contrast to a simple, uniform adjustment of exposure by the OpenCV library. So far, we have implemented two losses, i.e., the consistency loss and the identity loss, to retain the visual content after image translation. However, there still remain limitations and room for improvements. According to the results, some content of the images could still be hidden after translation. Inspired by Hoffman et al. [12], it is recommended that future works focus more on monitoring the feature loss. This is achievable by introducing images and translated images into a pretrained network, like VGG or Resnet, then extracting and comparing different levels of features between them. We are currently experimenting on this aspect, in which we expect to attain a great improvement in translation quality.

7. Acknowledgement

Work for this study is supported by the authors' supervisor, Prof. Qifeng Chen. The authors would also like to express their gratitude to the teaching assistants of COMP4471, Mr. Samuel Cahyawijaya, Mr. Hao Ouyang, Mr. Maosheng Ye, Mr. Chao Zhao, and Mr. Yiyao Zhu, for their assistance throughout the semester. Finally, credit is due to Google Colab for providing us the computational power to conduct experiments on this project.

References

[1] Zhenfeng Shao, Min Chen, and Chong Liu. Feature matching for illumination variation images. *Journal of Electronic Imaging*, 24:033011, 05 2015. 1, 2

[2] Hao Zhou, Torsten Sattler, and David Jacobs. *Evaluating Local Features for Day-Night Matching*, pages 724–736. 11 2016. 1, 2

[3] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint detection and description of local features, 2019. 1, 2

[4] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2020. 1

[5] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2018. 1

[6] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks, 2017. 2

[7] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, 2018. 2

[8] Yangyun Shen, Runnan Huang, and Wenkai Huang. Gd-stargan: Multi-domain image-to-image translation in garment design. *PLOS ONE*, 15(4):1–15, 04 2020. 2

[9] Sefik Eskimez, Dimitrios Dimitriadis, Kenichi Kumatani, and Robert Gmyr. One-shot voice conversion with speaker-agnostic stargan. pages 1334–1338, 08 2021. 2

[10] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, and T. Pajdla. 24/7 place recognition by view synthesis. In *CVPR*, 2015. 2

[11] Aachen day-night dataset, 2021. 2

[12] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation, 2017. 7