

South Dakota State University
**Open PRAIRIE: Open Public Research Access Institutional
Repository and Information Exchange**


Electronic Theses and Dissertations

2019

Development of Semantic Scene Conversion Model for Image-based Localization at Night

Dongyoun Kim
South Dakota State University

Follow this and additional works at: <https://openprairie.sdstate.edu/etd>

 Part of the [Computer Engineering Commons](#), and the [Electrical and Computer Engineering Commons](#)

Recommended Citation

Kim, Dongyoun, "Development of Semantic Scene Conversion Model for Image-based Localization at Night" (2019). *Electronic Theses and Dissertations*. 3382.
<https://openprairie.sdstate.edu/etd/3382>

This Thesis - Open Access is brought to you for free and open access by Open PRAIRIE: Open Public Research Access Institutional Repository and Information Exchange. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of Open PRAIRIE: Open Public Research Access Institutional Repository and Information Exchange. For more information, please contact michael.biondo@sdstate.edu.

DEVELOPMENT OF SEMANTIC SCENE CONVERSION METHOD FOR
IMAGE-BASED LOCALIZATION AT NIGHT

BY

DONGYOUN KIM

A Thesis submitted in partial fulfilment of the requirements for the

Master of Science

Major in Computer Science

South Dakota State University

2019

DEVELOPMENT OF SEMANTIC SCENE CONVERSION METHOD FOR
IMAGE-BASED LOCALIZATION AT NIGHT

This Thesis is approved as a creditable and independent investigation by a candidate for the Master of Science in Computer Science degree and is acceptable for meeting the Thesis requirements for this degree. Acceptance of this does not imply that the conclusions reached by the candidates are necessarily the conclusions of the major department.

Sung Y. Shin, Ph.D.
Thesis Advisor

Date

George Hamer, Ph.D.
Acting Head, Department of Electrical
Engineering and Computer Science

Date

Dean, Graduate School

Date

This Thesis is dedicated to my friends who go with me.

ACKNOWLEDGEMENTS

I would like to appreciate my advisor, Dr. Sung Shin for his considerable and continuous advice for me during my degree program. Dr. Shin gave me valuable feedback not just for my research, but also on my way to become a researcher.

I also cannot appreciate enough of the committee members of my thesis, Dr. Kwanghee Won, Dr. Yi Liu, and my graduate faculty representative, Dr. Lan Xu for their help on my thesis in the final defense. Due to their contributions, I was able to improve the quality of my thesis deeply and rationally. I would also like to thank the senior CCT Lab members, and I will never forget the days we studied together in the lab.

Finally, I cannot express enough appreciation towards my family members for their invaluable support for me. Without their help, I was not able to continue my academic career.

CONTENTS

ABBREVIATIONS	vi
LIST OF FIGURES	vii
LIST OF TABLES	viii
ABSTRACT	ix
1 INTRODUCTION	1
2 BACKGROUND	4
2.1 Generative Adversarial Network	4
2.2 Pix2Pix	5
2.3 Mask R-CNN	6
3 RELATED WORK	8
4 MATERIAL and METHOD	11
4.1 Instance Segmentation	12
4.2 Image Conversion	14
5 RESULT AND ANALYSIS	16
5.1 Experimental Setup	16
5.2 Evaluation	21
6 CONCLUSION	24
REFERENCES	25

ABBREVIATIONS

BMP Bitmap

CNN Convolutional Neural Network

GAN Generative Adversarial Network

HOG Histogram of Oriented Gradients

LBP Local binary patterns

ML Machine Learning

ORB Oriented Fast and Rotated BRIEF

ROI Machine

RGB Red, Green and Blue.

RPN Region Proposal Network

SIFT Scale-Invariant Feature Transform

SLAM Simultaneous Localization and Mapping

SURF Speeded-Up Robust Features

SVM Support Vector Machine

VLAD Vector of Locally Aggregated Descriptors

LIST OF FIGURES

1	The example of the appearance difference depending on the light condition - (a) day scene (b) night scene	1
2	Diagram of standard Generative Adversarial Network	4
3	The overview procedure of Mask R-CNN	7
4	Overview of the proposed model	11
5	The part of result of the day image on Mask RCNN	12
6	The result of the night image on Mask RCNN	13
7	The example of cropped traffic sign in the night images	14
8	The route of collecting dataset	16
9	The example of dataset and labeled dataset	17
10	The example results of ToDayGAN model with global images.	21
11	The example results of Pix2Pix model with global images.	21
12	The example results on traffic signs translation with 128 by 128 resolution images.	22

LIST OF TABLES

1	HYPER-PARAMETER SETTINGS FOR INSTANCE LEVEL SEG- MENTATION	18
2	HYPER-PARAMETER SETTINGS FOR IMAGE CONVERSION .	19
3	THE EXAMPLE RESULTS ON TRAFFIC SIGNS TRANSLATION WITH 128 BY 128 RESOLUTION IMAGES.	23

ABSTRACT

DEVELOPMENT OF SEMANTIC SCENE CONVERSION METHOD FOR
IMAGE-BASED LOCALIZATION AT NIGHT

DONGYOUN KIM

2019

Developing an autonomous vehicle navigation system invariant to illumination change is one of the biggest challenges in vision-based localization field due to the fact that the appearance of an image becomes inconsistent under different light conditions even with the same location. In particular, the night scene images have greatest change in appearance compared to the according day scenes. Moreover, the night images do not have enough information in Image-based localization. To deal with illumination change, image conversion methods have been researched. However, these methods could lose the detail of objects and add fake objects into the output images.

In this thesis, we proposed the semantic objects conversion model using the change of local semantic objects by categories at night. This enables the proposed model to obtain the detail of local semantic objects in image conversion. As a result, it is expected that the proposed model has a better result in image-based localization. Our model uses local semantic objects (i.e., traffic signs and street lamps) as categories. The model is composed of two phases as (1) instance segmentation and (2) semantic objects conversion. Instance segmentation is utilized as a detector for local semantic objects. In translation phase, the detected local semantic objects are translated from the appearance of the night image to day image.

In evaluation, we prove that models using a set of paired images show higher accuracy compared to the models using a set of unpaired images. Our proposed method will be compared with pix2pix and ToDayGAN. Moreover, the result quantitatively evaluates the best matching score with a query image and the converted images using ORB matching descriptor.

1 INTRODUCTION

Image-based localization is used to figure the relative position of cameras [18], [21], [23], [26], [28]. Image-based Localization is one of the most important fields in autonomous vehicle and robotics systems since precise localization can assist vehicles to make appropriate decisions. For image-based localization, the system requires for information retrieval and image matching [11]. Information matching problem requires correspondences between the query image and the reference images from the maps [39]. However, one of the challenges is the inconsistent visual appearance of objects under different lighting conditions [3], [22], [25], [27], [29]. The reason is that the same location shows different appearance in the images between the query image and reference images from the maps. In particular, there is severe appearance change between images taken at night and day. Figure 1 is an example of the appearance difference depending on the light condition [28].



Figure 1: The example of the appearance difference depending on the light condition - (a) day scene (b) night scene

The typical methods such as SIFT [20], SURF [4], and LBP [36] with image descriptors have been proposed for different viewpoints and intensity changes [11] on

image correspondence. However, the typical methods could not deal with the large illumination change between the day and night images because the typical methods rely on the gradient value [29]. Many approaches have been studied for the image matching with illumination change problem. For example, Milford and Wyeth proposed SeqSLAM [24] that utilizes multiple consecutive images to obtain enough visual features for localization at night. Nelson et al. used artificial light sources to identify the location [27], but the approach requires additional maps of the night image sequences and light sources.

Machine learning, especially deep learning, techniques can be used to develop a model that studies how visual appearances change depending on lighting conditions. Recently, Anoosheh et al. proposed a deep generative model that converts a night image to a corresponding day image [3]. The localization is done through the utilization of image-level descriptors since the conversion has been done globally rather than recovering local details. The trained translation (image conversion) model also generates fake information such as cars and trees for the dark area of the input night images. These converted images look real, but it will make the localization task hard to accomplish. Stenborg et al. have used deep neural network-based semantic segmentation for robust image-based cross-seasonal localization [34]. Semantic point feature map has been used for localization. However, this approach requires consistent semantic segmentation results across-season which can also be a challenging task.

In this thesis, a semantic local image conversion method is proposed. To the best of our knowledge, our proposed model is the first attempt to convert particular semantic objects during image translation from a night to a day image. As image-based localization becomes more important in autonomous vehicle and robotics system, the restoration of specific objects is necessary to conduct a precise localization which can assist vehicles to make appropriate decisions. Instead of applying a global image-level translation, a category-wise image conversion model is developed by utilizing both a semantic instance segmentation method and a deep generative model. Semantic labels are useful since it

excludes dynamic objects such as cars and pedestrian from localization tasks. The category-wise image conversion model can recover the details. Thus, this enables the feature-level correspondence search using conventional image features. In the proposed model, the semantic local objects are traffic signs and street lamps, because the signs and lamps are visible even at night scene. The proposed approach utilizes two off-the-shelf deep learning architectures. First, Mask R-CNN has been trained and used for instance segmentation of night and day images. The classes in instance segmentation are traffic signs and street lamps as semantic local information. Then, Pix2Pix [16] model, a Generative Adversarial Network has been trained for category-wise night-to-day image conversion. Before evaluating the proposed model, the result is demonstrated that one of the GAN methods, Pix2Pix, using a set of paired datasets can keep the more global structure in output image compared to ToDayGAN using a set of unpaired datasets with our dataset. To evaluate the proposed model, we show the evaluation that compares the global converted image with the semantic object converted image through visual inspection. Also, we quantitatively compared the number of ORB matching points with existing models [17], [32] to show that our model has more ability to recover the detail of semantic objects from night images. Therefore, it is expected that our model.

The rest of this paper is organized as follows: BACKGROUND section briefly introduces fundamental of related algorithms; RELATED WORK section briefly reviews related methodologies and applications; MATERIAL AND METHODS section describes the proposed algorithm; RESULT AND ANALYSIS presents the evaluated results; CONCLUSION propose overall of the paper at the end.

2 BACKGROUND

2.1 GENERATIVE ADVERSARIAL NETWORK

Generative Adversarial Network was proposed in 2014 by the authors of [9]. The main idea of GAN is to analyze the distribution of real data and to generate synthetic distribution like the real data [9]. This enables us to translate images from the source domain to the target domain as realistic images.

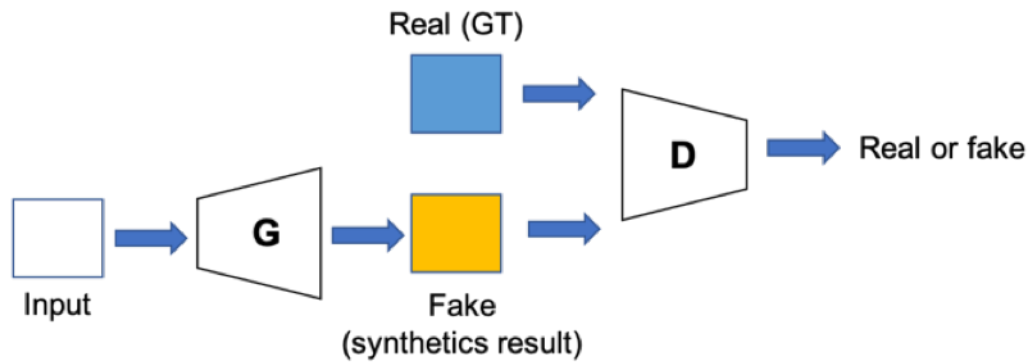


Figure 2: Diagram of standard Generative Adversarial Network

GAN is composed of two phases as discriminator (D) and generator (G). The G generates the fake image, the synthetic image, from the latent vector or source image. Then G makes the distribution of source images to be more similar to the real image. D determines whether the image is fake or not. The key point of GAN is the idea of an adversarial loss that generates a synthetic image to be real. In the training phase, the discriminator will be trained first with real and fake images and then the generator will be trained after discriminator is trained. Figure 2 shows the standard structure of generative adversarial networks. The general GAN uses adversarial losses for training as follows:

$$Loss_{GAN(G,D)} = \min_G \max_G \log(D(y)) + \log(1 - D(G(x)))$$

Where D indicates discriminator, G indicates generator, and x and y indicate input image

and output image, respectively.

2.2 PIX2PIX

Image-to-Image Translation with Conditional Adversarial Networks (Pix2Pix) proposed by Phillip Isola et al. [16] provides general software for effective image-to-image translation such as converting semantic label map to a photo-realistic image, reconstructing objects from edge maps, and colorizing images, among other tasks. The objectives of Pix2Pix is to convert an input image to a corresponding out image. Pix2Pix mitigate the issues from existing framework's loss of function which generates the image to be realistic. CNN-based framework has been used a wide variety of image prediction problems. The machines of CNN-based framework are trained for minimizing the loss function which means estimate a simple Euclidean distance between predicted and ground truth pixels. It will tend to produce blurry results [38]. To address the blurry results, GAN-based models [5], [9], [30], [33] are proposed. They are applied to an adversarial loss that trains the generator and discriminator. However, many of such concentrated on specific applications. Therefore, Pix2Pix combines the above two loss functions as follows:

$$Loss_{P2P(G,D)} = \min_G \max_G \log(D(y)) + \log(1 - D(G(x))) + Loss[||y - G(x)||]$$

Where D indicates discriminator, G indicates generator, and x and y indicate input image and ground truth image, respectively. The loss function of Pix2Pix is composed of two parts. The left part is the objectives of the existing GAN model for fooling the discriminator that we mentioned before. The rightmost part is the crucial idea of Pix2Pix, which is minimizing the difference between the ground truth image and output image, and it leads to the output image like the ground truth image. As a result, Pix2Pix is effective for a wider range of problems with considerably simple setup, which can be used for

general purposes that do not require domain-specific setup.

2.3 MASK R-CNN

Mask Region-based Convolutional Neural Network (Mask R-CNN) proposed by Facebook AI team [1]. Mask RCNN provides highly efficient and accurate pixel-wise mask information of objects for instance segmentation. The framework of Mask R-CNN is based on previous Region-based Convolutional Neural Network (R-CNN) [8], Fast-R-CNN [7], and Faster R-CNN [31]. First, R-CNN is proposed by Girshick et al. [8], which provide the object detection finding specific objects with bounding box. However, they require extra time for training three different models such as CNN for region proposal model, SVM for a classifier, and linear regression for tightening bounding box. Therefore, the author of Fast R-CNN proposed the RoIPooling that take the maximum value of filter in image features when the extracted feature by region proposal convert the fixed size. This enables to reduce the computational time. Moreover, they used Fully connected layer [19] architecture in classifier and bounding box instead of using SVM, which also reduce the computational time. Faster R-CNN is proposed by Ren, S. et al. for a more efficient framework [31]. This modified region proposal with the selective search algorithm to region proposal network with CNN, because the selective search algorithm is slow.

Unlike R-CNN based methods, Mask R- CNN proposed the mask prediction of objects which provides a spatial layout simultaneously with label classification and bounding box regression branches. Mask R-CNN provides rich feature extraction techniques without involving feature engineering by domain experts with the deep learning backbone architecture such as ResNet [13]. Figure 3 shows the overview of Mask R-CNN segmentation process from [15].

Similar with Faster R-CNN, Mask R-CNN adopted convolutional backbone for feature extraction, which is shared with Region Proposal Network (RPN) and RoIAlign

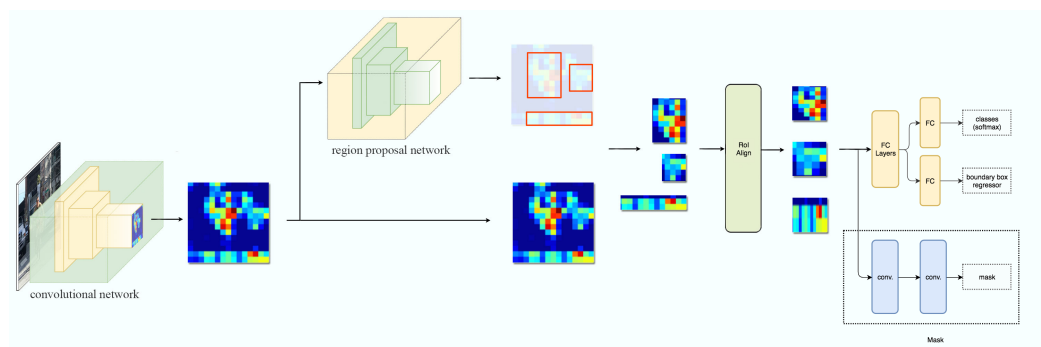


Figure 3: The overview procedure of Mask R-CNN

layer. RPN proposes a set of box-shaped ROIs which will be converted into a fixed-size feature map. Unlike RoIPool [31], RoIAlign converts each ROI into a fixed-size feature map using bi-linear interpolation to eliminate quantization problem. Additionally, RPN in Mask R-CNN proposes an expected region of the object as bounding box, which is kept trained through the extracted features and three branches and it is called head of the network: (1) mask branch, (2) box regression branch, and (3) label classification branch.

3 RELATED WORK

Over the past few decades, visual localization has been extensively researched. For tackling the illumination change problem, using feature descriptors [11], [17] have been developed using global and local features such as gradient, histogram, and etc. Typical methods using image feature descriptors are SIFT [20] and SUFT [4]. The scale-invariant feature transform (SIFT) was introduced by [20]. SIFT detects a number of interest points using Difference of Gaussian operator by building a histogram of gradient orientations. Although it was a more accurate result than other methods, it takes much computational time than others. The speeded-up robust features (SURF) was introduced by [4]. SURF detects a number of interest points using 2-Dimensional box which is using second-order Gaussian derivatives for efficient performance. However, the typical ways of using the descriptors could not deal with a large illumination change [11].

For dealing with illumination change problem, SeqSLAM [24] was proposed for addressing the problem of navigating a vehicle at night. SeqSLAM found the best-matched location candidate based on the information of image sequences in an illumination changing environment. However, the SeqSLAM [24] relied on the viewpoint direction and sequence information [14], [22]. Maddern et al. [22] proposed a mapping and scene classification with exploiting full knowledge of the spectral properties for 24 hours. The method improved performance in typical outdoor environments. However, it required adding mapping and observation during daylight hours. Nelson et al. [27] used artificial light sources for the localization at night. The model [27] tracked the changed appearance of the light sources depending on distances between the camera and the light sources. However, it required additional maps of artificial lights.

One of the ways to tackle the illumination problem is image conversion methods. Image to image translation was oriented from [37] that proposed an end-to-end algorithm and improved in [6] by using nonparametric texture. Generative Adversarial Network

(GAN) is one of the translation image methods from night to day with machine learning technique. Based on the idea of GAN [9], Pix2Pix [16] used convolutional neural network architecture for training the images conversion with paired datasets. The mechanism [16] is used as our base architecture. However, it had a restriction to collect paired dataset for machine training. The authors [40] designed CycleGAN that extends the model [16] to unpaired images translation. The key concept of CycleGAN was style transfer that makes the structure to keep up. This not only used adversarial loss but also used cycle loss for unpaired training. Even though CycleGAN [40] maintain the structures of the images, Pix2Pix [16] is more accurate in image conversion [40]. Based on these methods, GAN-based methods have been proposed for tackling the illumination problem. The authors of [29] presented a method of visual place recognition under our problem. The method [29] compared images under different illumination by visually translating with an unpaired dataset. They added an additional loss for descriptor on top of the conventional GAN based on CycleGAN [40]. The characteristics in [29] are that translated images should have identical SURF features. It can preserve the relevant features from the original image (input) by using cyclic reconstruction. ToDayGAN [3] also used the image to image translation method. The authors [3] used the generative adversarial network models based on ComboGAN [2] for image translation from night image(source) to the day images (target) with unpaired images. It was the first research of applying image translation to the problem of retrieval-based localization. The characteristics in ToDayGAN is using the three discriminators such as RGB-channel, gray-channel, and gradient information for keeping the structure of images and then to use DenseVLAD [35] descriptor for image retrieval method. Even though these GAN-based models [3], [29], [40] could generate the feasible fake images (output), they would lose the detail of some objects and add the fake objects because night is not enough to recover the information of some objects due to the dark side. This might result in failure of images matching on image-based localization.

Compared with previous works, the proposed model converts the local semantic objects of the night image by categories. It can alleviate loss of the detail of local objects and can ignore adding the fake information of invisible objects in images, which can lead to improvement of the image matching at night.

4 MATERIAL AND METHOD

In this section, our proposed algorithm and its preliminary are introduced. Unlike other conversion models that convert global image [16] and [3], the contribution of the proposed method is the categorical conversion that detects and convert specific objects in night scene. This is expected to improve image-based localization. The specific objects are defined as semantic objects, which are visible even at night scene and not changed frequently such as trees, vehicles, and people in the scene. The proposed model employed the Mask R-CNN [12] and Pix2Pix [16] as instance segmentation and image conversion method, respectively. Figure 4 shows an overview of the proposed model.

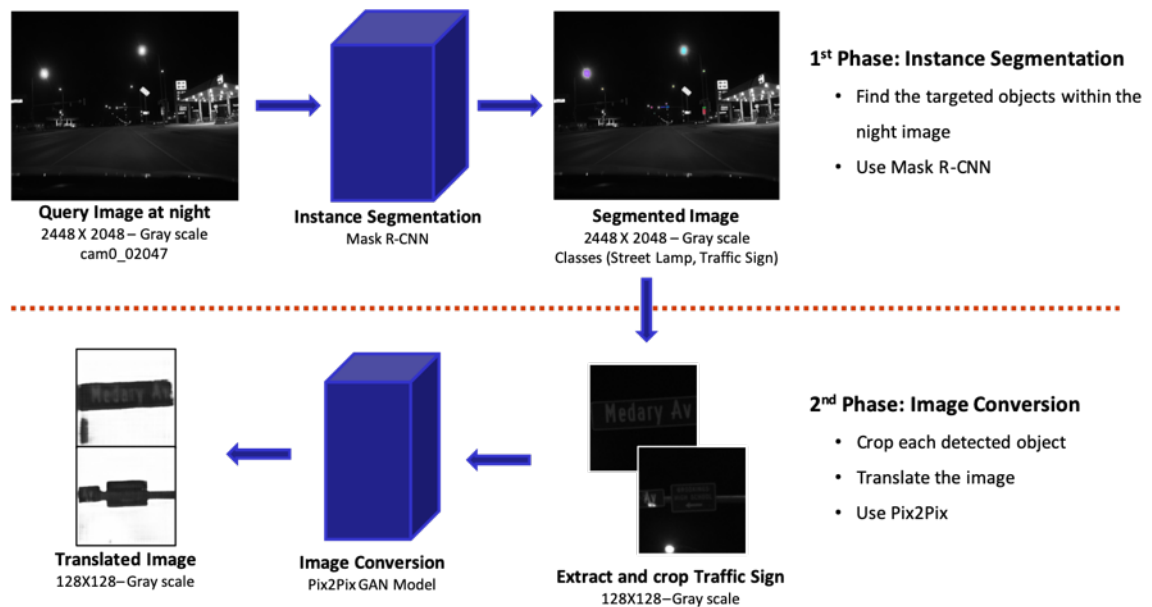


Figure 4: Overview of the proposed model

4.1 INSTANCE SEGMENTATION

Instance level segmentation task aims at identifying meaningful objects presented in a given image. In our model, this phase detects the semantic objects, traffic signs and street lamps, because they are visible even at night and static objects. we used state-of-the-art instance segmentation method, Mask RCNN [12], because it effectively detects a set of objects in a reasonable time. The according bounding boxes surrounding the detected local objects are suited as an input for the following objects conversion function. More precisely, Mask R-CNN extends Faster R-CNN [31] by adding mask prediction branches which provides a spatial layout of the object performed simultaneously with label classification and bounding box regression. Another notable modification in Mask R-CNN is that it mitigates the misalignments between the Region of Interest (RoI) and the extracted features using RoIAlign layer. Unlike other standard feature extracting techniques such as RoIPool and RoIWarp, RoIAlign is free from quantization that brings positive effect on finding more accurate RoIs. Based on this framework, the outputs of instance segmentation are class label, bounding box, and predicted mask.



Figure 5: The part of result of the day image on Mask RCNN



Figure 6: The result of the night image on Mask RCNN

With the aforementioned improvements, Mask R-CNN outperforms the variants of all previous state-of-the-art models, including the winner of the COCO 2016 Detection Challenge. It is worth noting that Mask R-CNN achieves good results even under challenging conditions, for instance, images containing overlapping instances like our dataset. Figure 5 and 6 show the result of Mask R-CNN with given day and night image, respectively.

4.2 IMAGE CONVERSION

Our main purpose is to translate the detected semantic objects from the appearance of night scene(query image) today (target image). After using instance segmentation, we crop the detected semantic objects image by 128 x 128 with empirical implementation from the centroid point of object. Figure 7 is shown as the output of cropped segmented image. The patch images from cropped results are to be the input images of pix2pix that is trained from traffic signs.



Figure 7: The example of cropped traffic sign in the night images

In image conversion phase, we employ Pix2Pix [16] architecture to train a generator in order to translate a set of semantic objects as a base. Note that even though Pix2Pix method requires a set of pair images (i.e., day and night at the same location), making data acquisition more cumbersome. Zhu et al. [3] showed that it gives better image translation results compared to other methods using unpaired image training

methods, such as CycleGAN or ToDayGAN [2], [3], [40]. Also, in our visual inspection, we confirmed that Pix2Pix generates more intuitively reasonable results (i.e., detailed edge) compared to ToDayGAN.

5 RESULT AND ANALYSIS

5.1 EXPERIMENTAL SETUP

Our image datasets for the experiment were collected in Brookings, SD, United States on March 25th, 2019. The traversals contain the sequence images of the same 4.9-mile route in Brookings as shown in Figure 8 from the Google map [10]. The stereo camera was mounted inside the vehicle's front window on the left and right side. Each side of the camera in the collected data set contains approximately 14,000 day and night images, respectively. The image is in grayscale BMP format with dimension of 2,448 by 2,048 pixels. 9 demonstrates the example of collecting dataset and labeled dataset for training instance segmentation.



Figure 8: The route of collecting dataset



Figure 9: The example of dataset and labeled dataset

For the instance level segmentation, the subset of 500 images from each day and night datasets were used as a training set. The selected datasets are labeled with two target classes - street lamps and traffic signs - which are visible local static objects at night. Figure 8 shows the example matched images from day and night dataset with ground truths.

For the semantic object conversion, we cropped 1,000 local semantic object that contains traffic signs only from the dataset used in the instance level segmentation. Because street lamps vary visually in the image depends on the conditions, we extracted only the location information of the street lamps and excluded them from the training set

of the translation model. The size of cropped image is set as 128 by 128 pixels, which is empirically found to be the best size after segmentation.

To optimize the model, we fine-tuned our instance-level segmentation and image translation model as indicated in Table 1 and Table 2. Both models offered options for selecting hyper-parameters, but we chose some parameters which fit for our dataset. The omitted parameters for the table are set as indicated in the original implementation [1], [16], [40].

Table 1: HYPER-PARAMETER SETTINGS FOR INSTANCE LEVEL SEGMENTATION

Instance Level Segmentation	
Model	Mask R-CNN
Size of training set	500
Target classes	[Street Lamp, Traffic Sign]
Pre-trained weights	COCO dataset
Backbone architecture	ResNet 101
# of epochs	90
Learning rate	0.01
Minimum confidence	0.6

The hyper-parameter setting is epoch, which is how many iterations completed to pass through the whole training set or batch size which is also hyper-parameter for training. The pre-trained weight is using MS-COCO dataset to initialize the weights of the network with weights from the pre-trained model using public natural scene image database. Backbone architecture is architecture of convolution neural network used for feature extraction. Minimum confidence represents a confidence threshold which indicates how confident the model is to predict the object for corresponding classes. The number of iterations with learning rate decay which dynamically reduce the learning rate to zero

Table 2: HYPER-PARAMETER SETTINGS FOR IMAGE CONVERSION

Image Conversion	
Model	Pix2Pix
Crop size	128 by 128
Size of training set	1,000
Direction	Night to day
# of channels for input	1
# of channels for output	1
Learning rate	0.0002
# of iteration with original learning rate	200
# of iteration with learning rate decay	200
# of epochs	400

during the iteration. For example, 200 iterations with original learning rate (0.0002) and 200 iterations with learning rate decay in our implementation.

The models were implemented with the image processing library in Tensorflow and PyTorch, tested on the environment of 2.8GHz Intel Core i7 with 16GB 2133 MHz LPDDR3 memory.

5.2 EVALUATION

We demonstrated the experimental result of the proposed model with image translation for the semantic objects. Before the evaluation of the proposed method, the result is shown as Figure 10 and Figure 11 with global images - input images are the entire scene per frame.



Figure 10: The example results of ToDayGAN model with global images.



Figure 11: The example results of Pix2Pix model with global images.

The visual inspection shows that Pix2Pix using a set of paired datasets keep more

their structure than ToDayGAN using a set of unpaired datasets [3]. In training, the result of Pix2Pix was trained with 1, 200 paired images and the result of ToDayGAN was trained with 10, 000 unpaired images because the paired images require human-resource for matching each image at the same location. On the other hands, Pix2Pix and ToDayGAN with global images lost the detail and added fake objects.

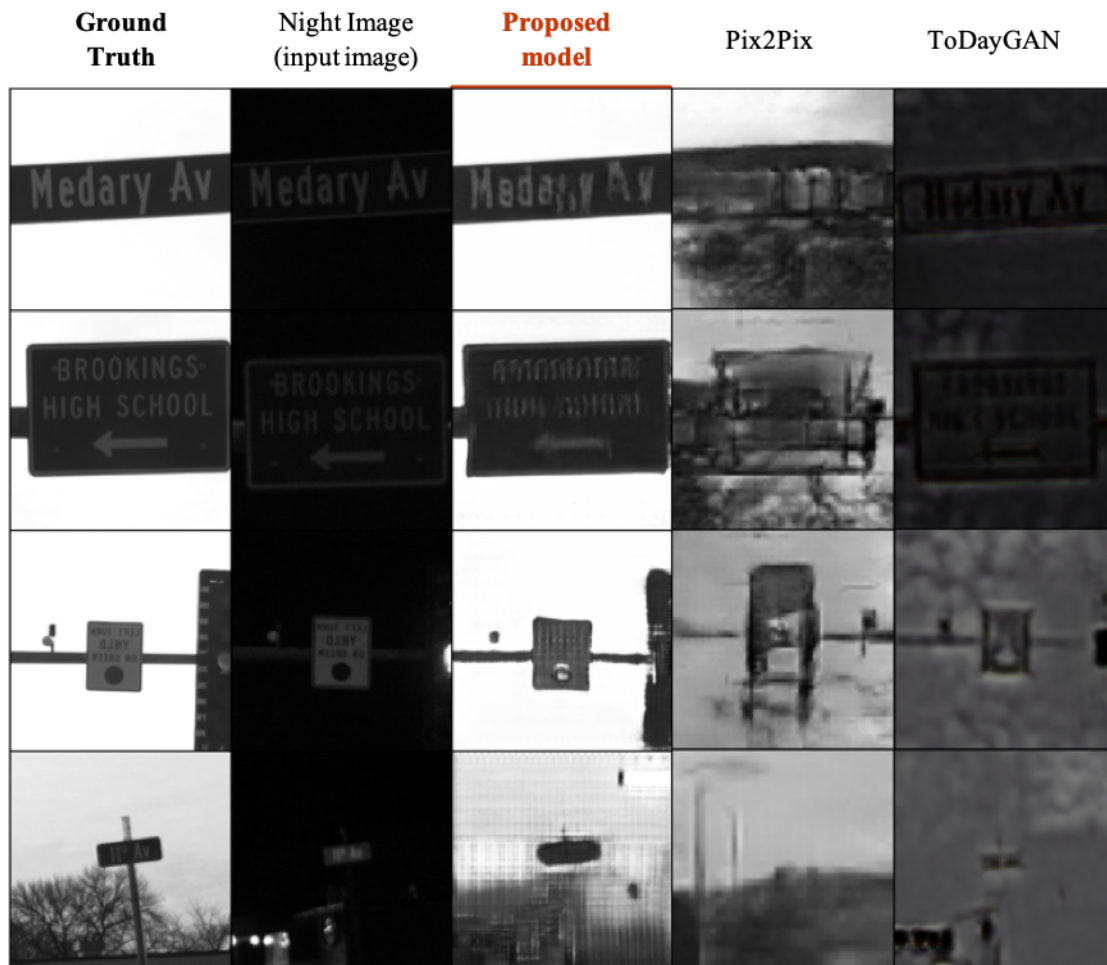


Figure 12: The example results on traffic signs translation with 128 by 128 resolution images.

Figure 12 shows the example results of our proposed model and existing models with query images and its expected outputs. We observed that even though the existing method could restore the scratch shape of the query images, they lose many of the details such as edges or color. This result makes each unique traffic sign indistinguishable.

We quantitatively measure the performance of the models by counting the number of matching points using ORB [32] feature descriptor. The ORB feature descriptor is one of the widely used descriptor in image matching method [17], [32]. It computes the intensity weighted centroid of the patch with located corner at center [17]. Table 3 demonstrates that average of matching results on ORB feature descriptor with 50 images from test sets. The results show that the validity of the proposed model to reconstruct details by showing that the proposed model has more matching points compared to the existing methods.

Table 3: THE EXAMPLE RESULTS ON TRAFFIC SIGNS TRANSLATION WITH 128 BY 128 RESOLUTION IMAGES.

Models	Average of matching
Night (query image)	20.8
Pix2Pix	100.2
TodayGAN	40.4
Proposed method	110.8

6 CONCLUSION

In this paper, the proposed model which is called the semantic object conversion is designed to deal with the appearance change problem in images matching with night image [18], [26], [28]. The proposed model includes instance level segmentation method and generative adversarial network, which work as local objects detector and conversion by categories, respectively.

By converting the local semantic object image from night to day, the model could recover the detail information as parts of the image. In first evaluation, the result demonstrates that models using a set of paired images show higher accuracy compared to models using a set of unpaired images. The visual inspection results of detail restoration in image show that the semantic object converted image maintains better detail compared to the global converted image. Moreover, the proposed model quantitatively achieves the best matching score with a query image and the converted images using ORB matching descriptor.

Based on the experiment, we will expand the model in image-based localization field by using the street lamp information. The extracted location information of the street lamp will be used to evaluate the 3-dimensional coordinate pose estimation with stereo cameras. We will also use open data sets to evaluate the model to show the robustness of the model for real-world applications.

REFERENCES

- [1] W. Abdulla, *Mask R-CNN for object detection and instance segmentation on Keras and Tensorflow*, Mar. 2017.
- [2] A. Anoosheh, E. Agustsson, R. Timofte, and L. Van Gool, “Combogan: Unrestrained scalability for image domain translation,” Dec. 2017.
- [3] A. Anoosheh, T. Sattler, R. Timofte, M. Pollefeys, and L. Van Gool, *Night-to-day image translation for retrieval-based localization*, Sep. 2018.
- [4] H. Bay, T. Tuytelaars, and L. Van Gool, “Surf: Speeded up robust features,” in *European conference on computer vision*, Springer, 2006, pp. 404–417.
- [5] E. L. Denton, S. Chintala, R. Fergus, *et al.*, “Deep generative image models using a $\frac{1}{4}$ laplacian pyramid of adversarial networks,” in *Advances in neural information processing systems*, 2015, pp. 1486–1494.
- [6] A. Dosovitskiy and T. Brox, “Generating images with perceptual similarity metrics based on deep networks,” Feb. 2016.
- [7] R. Girshick, “Fast r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.
- [8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.
- [9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in neural information processing systems*, 2014, pp. 2672–2680.
- [10] Google, “Google maps directions for driving from vancouver, bc to nelson, bc,” 2018. [Online]. Available: <https://www.google.ca/maps/dir/vancouver,+bc/nelson,+bc/@48.90494,-121.3409231,8z/data=!3m1!4b1!4m13!4m12!1m5!1m1!1s0x548673f143a94fb3:0xbb9196ea9b81f38b!2m2!1d-123.1207375!2d49.2827291!1m5!1m1!1s0x537cb41f1c6bb871:0x6d0054861620bcc!2m2!1d-117.2948343!2d49.4928119>.
- [11] M Hassaballah, A. Ali, and H. Alshazly, “Image features detection, description and matching,” in. Feb. 2016, vol. 630, pp. 11–45, ISBN: ISBN 978-3-319-28852-9. DOI: 10.1007/978-3-319-28854-3_2.
- [12] K. He, G. Gkioxari, P. Dollar, and R. Girshick, “Mask r-cnn,” in *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

- [13] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [14] K. L. Ho and P. Newman, “Detecting loop closure with scene sequences,” *International Journal of Computer Vision*, vol. 74, pp. 261–286, 2006.
- [15] J. Hui. (2018). Image segmentation with mask r-cnn, [Online]. Available: https://medium.com/@jonathan_hui/image-segmentation-with-mask-r-cnn-eb6d793272 (visited on 06/30/2019).
- [16] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976, 2017.
- [17] E. Karami, S. V. H. Prasad, and M. S. Shehata, “Image matching using sift, surf, brief and orb: Performance comparison for distorted images,” *ArXiv*, vol. abs/1710.02726, 2017.
- [18] P. Kim, B. Coltin, O. Alexandrov, and H. Jin Kim, “Robust visual localization in changing lighting conditions,” May 2017, pp. 5447–5452. DOI: 10.1109/ICRA.2017.7989640.
- [19] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
- [20] D. G. Lowe *et al.*, “Object recognition from local scale-invariant features.,” in *iccv*, vol. 99, 1999, pp. 1150–1157.
- [21] S. M. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. D. Cox, P. I. Corke, and M. Milford, “Visual place recognition: A survey,” *IEEE Transactions on Robotics*, vol. 32, pp. 1–19, 2016.
- [22] W. P. Maddern, A. D. Stewart, C. McManus, B. Upcroft, W. Churchill, and P. Newman, “Illumination invariant imaging : Applications in robust vision-based localisation , mapping and classification for autonomous vehicles,” 2014.
- [23] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, “A comparison of affine region detectors,” *International Journal of Computer Vision*, vol. 65, pp. 43–72, Nov. 2005. DOI: 10.1007/s11263-005-3848-x.
- [24] M. Milford and G. Wyeth, “Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights,” May 2012, pp. 1643–1649, ISBN: 978-1-4673-1403-9. DOI: 10.1109/ICRA.2012.6224623.

- [25] T. Naseer, M. Ruhnke, C. Stachniss, L. Spinello, and W. Burgard, “Robust visual slam across seasons,” *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2529–2535, 2015.
- [26] T. Naseer, B. Suger, M. Ruhnke, and W. Burgard, “Vision-based markov localization across large perceptual changes,” Sep. 2015, pp. 1–6. DOI: 10.1109/ECMR.2015.7324181.
- [27] P. Nelson, W. Churchill, I. Posner, and P. Newman, “From dusk till dawn: Localisation at night using artificial light sources,” *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2015, pp. 5245–5252, Jun. 2015. DOI: 10.1109/ICRA.2015.7139930.
- [28] N. Piasco, D. SidibÃ©, C. Demonceaux, and V. Gouet-Brunet, “A survey on visual-based localization: On the benefit of heterogeneous data,” *Pattern Recognition*, vol. 74, Sep. 2017. DOI: 10.1016/j.patcog.2017.09.013.
- [29] H. Poray, W. Maddern, and P. Newman, “Adversarial training for adverse conditions: Robust metric localisation using appearance transfer,” May 2018, pp. 1011–1018. DOI: 10.1109/ICRA.2018.8462894.
- [30] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” *arXiv preprint arXiv:1511.06434*, 2015.
- [31] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, Jun. 2015. DOI: 10.1109/TPAMI.2016.2577031.
- [32] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “Orb: An efficient alternative to sift or surf,” Nov. 2011, pp. 2564–2571. DOI: 10.1109/ICCV.2011.6126544.
- [33] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training gans,” in *Advances in neural information processing systems*, 2016, pp. 2234–2242.
- [34] E. Stenborg, C. Toft, and L. Hammarstrand, “Long-term visual localization using semantically segmented images,” Jan. 2018.
- [35] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, “24/7 place recognition by view synthesis,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1808–1817.
- [36] X. Wang, T. X. Han, and S. Yan, “An hog-lbp human detector with partial occlusion handling,” in *2009 IEEE 12th international conference on computer vision*, IEEE, 2009, pp. 32–39.

- [37] Y. Xia, D. He, T. Qin, L. Wang, N. li Yu, T. Liu, and W.-Y. Ma, “Dual learning for machine translation,” in *NIPS*, 2016.
- [38] R. Zhang, P. Isola, and A. A. Efros, “Colorful image colorization,” in *European conference on computer vision*, Springer, 2016, pp. 649–666.
- [39] W. Zhang and J. Kosecka, “Image based localization in urban environments.,” in *3DPVT*, Citeseer, vol. 6, 2006, pp. 33–40.
- [40] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251, 2017.