Image and Vision Computing 151 (2024) 105257

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Nighttime scene understanding with label transfer scene parser

Thanh-Danh Nguyen [a,d], Nguyen Phan [a,d], Tam V. Nguyen [e,*], Vinh-Tiep Nguyen [a,d], Minh-Triet Tran [b,c,d]

[a] University of Information Technology, Ho Chi Minh City, Viet Nam
[b] University of Science, Ho Chi Minh City, Viet Nam
[c] John von Neumann Institute, VNU-HCM, Viet Nam
[d] Vietnam National University, Ho Chi Minh City, Viet Nam
[e] University of Dayton, Dayton, OH 45469, United States

ARTICLE INFO

Keywords:
Semantic segmentation
Nighttime scene parser
Domain adaptation
Generative adversarial network

ABSTRACT

Semantic segmentation plays a crucial role in traffic scene understanding, especially in nighttime conditions. This paper tackles the task of semantic segmentation in nighttime scenes. The largest challenge of this task is the lack of annotated nighttime images to train a deep learning-based scene parser. The existing annotated datasets are abundant in daytime conditions but scarce in nighttime due to the high cost. Thus, we propose a novel Label Transfer Scene Parser (LTSP) framework for nighttime scene semantic segmentation by leveraging daytime annotation transfer. Our framework performs segmentation in the dark without training on real nighttime annotated data. In particular, we propose translating daytime images to nighttime conditions to obtain more data with annotation in an efficient way. In addition, we utilize the pseudo-labels inferred from unlabeled nighttime scenes to further train the scene parser. The novelty of our work is the ability to perform nighttime segmentation via daytime annotated labels and nighttime synthetic versions of the same set of images. The extensive experiments demonstrate the improvement and efficiency of our scene parser over the state-of-the-art methods with a similar semi-supervised approach on the benchmark of Nighttime Driving Test dataset. Notably, our proposed method utilizes only one-tenth of the amount of labeled and unlabeled data in comparison with the previous methods. Code is available at https://github.com/danhntd/Label_Transfer_Scene_Parser.git.

1. Introduction

In recent years, the technological development of autonomous vehicles has manifested its conveniences towards human life. Instead of manually driving for miles among destinations, autonomous cars help drivers save their energy for other purposes such as working or relaxing. Moreover, with the assistance of intelligent machines, automated vehicles can themselves get over traffic accidents or find the way to a specific destination. Therefore, these vehicles are being investigated to be more accurate and efficient. In fact, autonomous vehicles receive information as images and signals from cameras, sensors, or radars around the vehicle to compute the next appropriate movement. In this task, computer vision is leveraged to handle related image processing issues in order to assist the computer in image understanding Janai, Güney, Behl and Geiger [1]; Chen, Seff, Kornhauser and Xiao [2]; Haque, Islam, Alam, Iqbal and Shaik [3]; Nassi, Ben-Netanel, Elovici and Nassi [4].

Apart from image classification or object detection, semantic image segmentation outputs the semantic labels of pixels in images. In other words, segmentation is a task of pixel classification, which provides pixel-level accuracy. Applying computer vision to automobiles, we take street views from a front-camera of the vehicles as inputs of the semantic image segmentation model. Such tasks have been done under the good weather conditions of daytime Tao, Sapra and Catanzaro [5]; Mohan and Valada [6]; Liu, Chen, Schroff, Adam, Hua, Yuille and Fei-Fei [7]; Chen, Papandreou, Kokkinos, Murphy and Yuille [8,9]; Chen, Papandreou, Schroff and Adam [10]; Chen, Zhu, Papandreou, Schroff and Adam [11]; Zhao, Shi, Qi, Wang and Jia [12]; Zhang, David, Foroosh and Gong [13]; Wang, Gao and Li [14]; Shi, Li, Meng, Wu, Xu and Ngan [15]. However, driving outside in low light conditions of nighttime takes a significant proportion of operating time. Under the context of autonomous driving, segmentation in nighttime conditions is even more necessary. Meanwhile, there is a variety of challenges in nighttime conditions. This work focuses on *semantic image segmentation in nighttime conditions for urban driving scenes*. In fact, the differences between

* Corresponding author.
E-mail address: tamnguyen@udayton.edu (T.V. Nguyen).

https://doi.org/10.1016/j.imavis.2024.105257
Received 14 May 2022; Received in revised form 16 August 2023; Accepted 29 August 2024
Available online 8 September 2024

nighttime and daytime semantic segmentation on urban images lie in the conditions of input images. Instead of a clear and detailed urban image, the nighttime semantic segmentation model takes an intense input to understand, which is whether low contrast, low brightness, or sometimes suffers from light flare or noise artifact (qualitative evidence can be observed in Fig. 3). The task of the semantic segmentation model retains pixel-wise semantic labeling but becomes more difficult as it needs to be functional in adverse conditions.

The task of semantic segmentation in urban nighttime scenes addresses *two challenges*. Firstly, there hardly exists nighttime urban scenes datasets which are specific to the segmentation issue. Secondly, our concern is how to effectively leverage such data in the given task if we are supplied with suitable targeted datasets. Details are discussed below:

The most considerable challenge of our work is related to the dataset. We are facing the problem of *lacking training annotated nighttime city images* for the task of semantic image segmentation. Actually, despite the fact that there already exists many datasets of driving urban scenes, their specifications are mostly in daytime Cordts, Omran, Ramos, Rehfeld, Enzweiler, Benenson, Franke, Roth and Schiele [16]; Neuhold, Ollmann, Rota Bulo and Kontschieder [17]; Yu, Chen, Wang, Xian, Chen, Liu, Madhavan and Darrell [18] and rare in nighttime Tan, Xu, Cao, Zhang, Ma and Lau [19]. To the best of our knowledge, it is such an expensive process to produce a nighttime dataset with annotation for semantic segmentation. Honestly, the poor illumination condition of nighttime makes it hard to annotate a segmentation dataset. To remedy this problem, there are many approaches such as: manually creating a dataset facing the mentioned difficulty or applying the few-shot learning method Liu, Zhang, Lin and Liu [20]; Rakelly, Shelhamer, Darrell, Efros and Levine [21,22]; Abdel-Basset, Chang, Hawash, Chakrabortty and Ryan [23] with a small amount of annotated images. Therefore, we come up with the solution of automatically creating nighttime datasets based on existing daytime ones, which is known as a domain adaptation method. Among methods, GAN recently yields exceptional performance in the field of image generation. Thoroughly, we decide to leverage GAN to transfer daytime images to the nighttime domain while maintaining the contextual structure of the images. By the way, we can also make use of their segmentation annotation as their semantic contents are maintained regardless of daytime or nighttime, which is named after label transfer. To this end, our proposed framework has the ability to segment in the dark without training on any annotated nighttime images.

Though GAN helps translate images among domains to form a suitable dataset for our task, another arising challenging question is *how to utilize this amount of data efficiently*. Normally, we leverage prior knowledge by using a pre-trained model as a base model and perform refinement with new data. However, pre-training in that way does not always improve the performance Zoph, Ghiasi, Lin, Cui, Liu, Cubuk and Le [24]. Consider the idea of the teacher and student model, we apply a self-training approach in our work. Particularly, we use the created nighttime images combined with existing daytime data to train our segmentation model and make use of another amount of unlabeled nighttime scenes to boost the performance. In this way, we take advantage of the trained model to learn advanced knowledge itself.

To address the aforementioned problems, in this paper, we propose a novel framework - Label Transfer Scene Parser to tackle the problem of semantic image segmentation in the dark with the assistance of the domain adaptation method along with making use of existing daytime annotated images. In particular, we target applying GAN-based methods to do the task of domain adaptation. This idea allows utilizing daytime segmentation datasets to address nighttime scene segmentation. Besides, in order to improve the segmentation performance, we employ the self-training technique. Altogether, we aim to optimize the performance of our proposed framework.

Our main contributions in this work are three-fold:

- We propose a novel Label Transfer Scene Parser (LTSP) framework for nighttime semantic segmentation by label transfer. Our framework can perform segmentation in the dark without training on real nighttime annotated data.
- We obtain the free annotation for nighttime scenes for the task of semantic segmentation. This process saves the annotation cost of conducting such real nighttime annotated segmentation data.
- We propose a new fusion loss function that improves the semantic image segmentation model performance. As cross-entropy loss targets at measuring the difference between each couple of pixels, and focal loss aims to balance the dominance of major objects, fusing two loss functions can better address the segmentation problem.

The rest of this paper is organized as follows. Sec. 2 reviews related work of semantic image segmentation problem along with the method of GAN and domain adaptation method. Sec. 3 presents our proposed framework that leverages domain adaptation method and self-training technique to solve nighttime scene parsing. In Sec. 4, we provide experimental evidence to verify our hypothesis and ablation study to demonstrate the effectiveness of each additional component. Finally, we conclude our paper in Sec. 5.

2. Related works

2.1. Image domain adaptation

Domain adaptation is a field associated with machine learning and transfer learning. Normally, the training and testing processes are performed on the same data distribution. However, there exist many cases that the domains between the training set and the test set are different. Domain adaptation method is utilized to solve such problem via minimizing the differences between training and target domains. In our case, we have annotated daytime images for semantic image segmentation purpose, meanwhile, our target is to segment nighttime images. Thus, the domain adaptation method adapts available daytime images to the required domain. Previous work have been done with methods that acquire achievements Sun, Wang, Yang and Xiang [25]; Romera, Bergasa, Yang, Alvarez and Barea [26]; Dai and Van Gool [27]; Cho, Baek, Koo, Arsalan and Park [28]; Nag, Adak and Das [29]. These work focus on combining models to address model adaptation. Some data augmentation techniques such as random cropping, random rotation, and flipping are leveraged to stably adapt in unrelated domains Yang, Bergasa, Romera and Wang [30]. There are researches studying the effective use of synthetic data Saleh, Aliakbarian, Salzmann, Petersson and Alvarez [31]; Sadat Saleh, Sadegh Aliakbarian, Salzmann, Petersson and Alvarez [32]; Xu, Wang, Yang, Sun and Fu [33]. Pre-processing input images is also used to prevent performance degradation Porav, Bruls and Newman [34].

2.2. Generative adversarial network

Generative Adversarial Network (GAN) is a class of machine learning frameworks proposed by Goodfellow et al. Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville and Bengio [35] in 2014. Since then, GAN opens a new innovative horizon of deep learning, particularly computer vision. Specifically, GAN is an approach of a generative model using adversarial methods. Adversarial learning is a technique of machine learning that tries to fool models by providing deceptive data. In this work, we take GAN as the method to perform image domain translation. In the problem of Image-to-Image translation, Pix2Pix Isola, Zhu, Zhou and Efros [36] addresses this task with a paired image dataset. CycleGAN Zhu, Park, Isola and Efros [37] solves the same problem, yet performs without the requirement of paired data thanks to cycle consistency. UNIT Liu, Breuel and Kautz [38] works with the assumption of shared latent space. Shared latent space forms a relationship between two different domains and allows them to transform from side to side. However, one-to-one mapping models like UNIT fail to generate diverse outputs with a single input image. To overcome this

problem, MUNIT Huang, Liu, Belongie and Kautz [39] is proposed by assuming that the image representation can be decomposed into content code and style code. BicycleGAN Zhu, Zhang, Pathak, Darrell, Efros, Wang and Shechtman [40] aims to generate multiple possible outputs with a single input image by encouraging the relationship between output and the latent code. StarGAN Choi, Choi, Kim, Ha, Kim and Choo [41] addresses the limitation of handling the GAN model for more than two domains. This approach uses only a single model to train on multiple datasets with different domains.

2.3. Semantic scene parsing

Under the context of autonomous driving, semantic image segmentation supports understanding the surroundings to avoid obstacles. The early works made use of superpixel over-segmentation and extracted feature classification Tighe, Niethammer and Lazebnik [42]; Nguyen, Lu, Sepulveda and Yan [43]; [44]. Meanwhile, most of the recent segmentation networks rely on end-to-end fully convolutional networks Long, Shelhamer and Darrell [45]. A semantic segmentation model is usually a combination of encoder and decoder, the encoder is in charge of a feature extractor, and the decoder plays the role of reconstructing the output of the encoder to form a label map of the input image.

Typical networks like DeconvNet Noh, Hong and Han [46] follow the mentioned architecture of Long et al. [45], while UNet Ronneberger, Fischer and Brox [47] has an improvement by adding skip connections to preserve the global information. Other work of PSPNet Zhao et al. [12] and RefineNet Lin, Milan, Shen and Reid [48] are proposed with an aim at focusing on the spatial information of the image. Panoptic FPN Kirillov, Girshick, He and Dollár [49] itself focuses on learning spatial information by pyramid features which are essential to the problem of scene understanding. DeepLab Family Chen et al. [8–11] yield significant performance on semantic image segmentation mainly thanks to atrous convolution, atrous spatial pyramid pooling (ASPP) module.

In terms of nighttime semantic segmentation, Christos et al. Sakaridis, Dai and Gool [50]; Sakaridis, Dai and Van Gool [51] also leverage the concept of gradually transferring the domain of daytime to nighttime with a buffer stage of twilight images at the same GPS location. Wu et al. Wu, Wu, Guo, Ju and Wang [52] propose DANNet with an image relighting network to force the distributions of source domains (source daytime, a pair of target daytime and target nighttime domain) closer to each other. The weight-sharing segmentation module then performs its task with the grouth-truth of the source daytime domain only. Dai et al. Dai and Van Gool [27] propose a method training on daytime and twilight urban scenes. The key idea of this approach is to narrow down the distance between daytime and nighttime domains thanks to the bridge of twilight images. Thus, they take several steps to transfer the knowledge from nighttime annotated images to the nighttime segmentation model. Moreover, this work utilizes a large amount of data for segmentation: 8 K daytime labeled images and more than 26 K unlabeled twilight images. To this end, our proposed LTSP framework sets new state-of-the-art results with a smaller amount of data: 2.9 K daytime images and 1.6 K nighttime unlabeled images. Recently in 2021, Lengyel et al. Lengyel, Garg, Milford and van Gemert [53] proposed the approach of zero-shot learning to the task of nighttime scene segmentation. Vobecky et al. Vobecky, Hurych, Siméoni, Gidaris, Bursuc, Pérez and Sivic [54] conducted a work on cross-modal distillation for this task. In particular, we propose directly translating daytime images to nighttime conditions by using GAN in order to obtain the free annotation data. We utilize the pseudo-labels inferred from unlabeled nighttime scenes to further train the scene parser. We also introduce a new comprehensive loss function to improve the performance of the nighttime scene semantic segmentation.

2.4. Datasets

As to the need for urban semantic scene understanding, the research community introduced several datasets to serve this task. Popular benchmarks on urban images are in the daytime, compared to a limited number of datasets in nighttime conditions. Cityscapes Cordts et al. [16] contains 5000 images of urban scenes with high-resolution pixel-level annotations of $2,048 \times 1,024$. Its training, validation, and testing set includes 2975, 500, and 1525 images, respectively, divided into 19 semantic categories. CamVid Brostow, Shotton, Fauqueur and Cipolla [55]; Brostow, Fauqueur and Cipolla [56] is a small urban dataset with around 700 images and 32 classes. The number of images and annotations for each set is 367, 101, and 233 samples. Apart from the aforementioned datasets, Vistas Mapillary Neuhold et al. [17] is a huge dataset with up to 25,000 high-resolution images annotated into 66 categories, different from the definitions of Cityscapes. Thus, it is not suitable for merging this dataset for training. The train-val-test numbers of images are 18,000, 2000, and 5000 images, respectively. The collected images are from various daytime conditions and from different capturing devices. With regard to video, BDD100K Yu et al. [18] on the other concept provides 100,000 videos on urban scenes to serve around 10 tasks on autonomous driving. Under nighttime conditions, the datasets of Tan et al. [19]; Sakaridis et al. [51]; Dai and Van Gool [27] are the three common datasets. Dark Zurich Sakaridis et al. [51] consists of 2416 nighttime, 2920 twilight, and 3041 daytime images, which are all unlabeled with a resolution of $1,920 \times 1,080$. This dataset provides segmentation annotations for only 201 nighttime images, including 50 for validation and 151 for testing. NightCity Tan et al. [19] provides 4297 nighttime images of diverse complexity, with pixel-wise annotations. NightCity is separated into 2998, and 1299 images for training and testing purposes. Nighttime Driving Test Dai and Van Gool [27] is a testing dataset containing 50 fine-grained annotated images at $1,920 \times 1,080$ resolution. The labels of this test set are compatible with the Cityscapes and Dark Zurich testing dataset.

3. Proposed framework

3.1. Framework overview

In this paper, our overarching goal is to address the problem of lacking annotated nighttime urban scenes dataset for semantic segmentation. Fig. 1 shows the overview of the proposed framework, named after Label Transfer Scene Parser - LTSP framework. Our LTSP framework consists of two main components. The first component is responsible for GAN-based Day-Night image domain translation. Meanwhile, the second component handles the nighttime scene semantic segmentation.

Just a quick glimpse, the first component takes the responsibility of converting daytime images to nighttime domain with an aim at preparing to train our segmentation model. In the next semantic segmentation component, we train our model to make predictions on nighttime images. Furthermore, we make use of the self-training technique in this module to refine our segmentation model. In particular, our LTSP framework consists of five main steps. *Firstly*, two sets of daytime and nighttime images are taken to train our GAN-based image translation model. *Secondly*, the trained model translates daytime images into the nighttime domain. Both daytime and nighttime images share the same ground-truth and become our segmentation training dataset. *Thirdly*, our semantic segmentation model is trained on the newborn dataset. *Fourthly*, a set of extra unlabeled nighttime images is used to infer pseudo-labels. *Finally*, the combination of the inferred images with pseudo-labels and the previous training data are utilized to perform the self-training stage, which we call *re-train* our model.

Let us denote an image by $\mathbf{x}$, and $\mathbf{x}^{ds}, \mathbf{x}^{ns}$ indicate daytime, and nighttime scenes for training an image domain translator, respectively. To train an image segmentor, we indicate images taken at daytime and nighttime by $\mathbf{x}^0$, $\mathbf{x}^1$, respectively. The corresponding annotation for $\mathbf{x}^0$ is available and denoted by $\mathbf{y}^0$, where $\mathbf{y}^0(a,b) \in \{1,\ldots,C\}$ is the label of

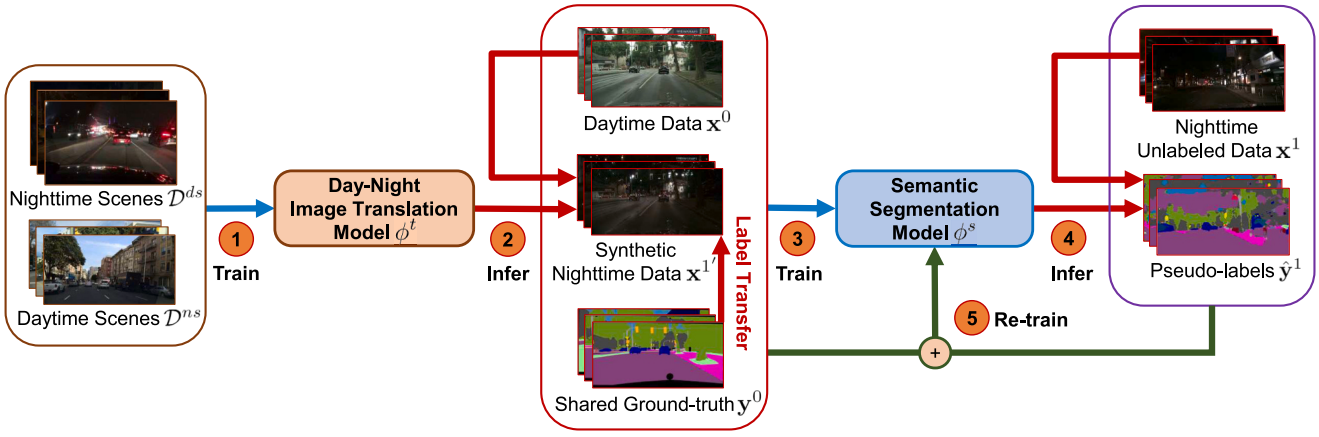

Fig. 1. Our proposed LTSP framework for scene parsing via domain adaptation. The framework is divided into two components: image domain translation and semantic image segmentation. The first component synthesizes nighttime images from given daytime samples. The second component performs nighttime scene segmentation with the self-training method. In other words, the proposed LTSP can complete the task of segmentation in the dark without training on any real nighttime annotated images. The pipeline is carefully described in Sec. 3.1.

pixel (a, b) and C is the number of semantic classes. Besides, the synthetic nighttime images are denoted as $\mathrm{x}^{1'}$. Afterwards, the training data of the image domain translator include two sets of images $\mathscr{D}^{ds} = \{\mathrm{x}_i^{ds}\}_{i=1}^{l^{ds}}$ and $\mathscr{D}^{ns} = \{\mathrm{x}_j^{ns}\}_{j=1}^{l^{ns}}$. The training data of the image segmentor includes a set of annotated daytime images $\mathscr{D}^0 = \{(\mathrm{x}_k^0, \mathrm{y}_k^0)\}_{k=1}^{l^0}$, nighttime synthetic images $\mathscr{D}^{1'} = \{(\mathrm{x}_m^{1'}, \mathrm{y}_m^{1'})\}_{m=1}^{l^{1'}}$ and real unlabeled nighttime images $\mathscr{D}^1 = \{\mathrm{x}_n^1\}_{n=1}^{l^1}$. In our case, we share the segmentation annotations from daytime images, so $\mathrm{y}^{1'} = \mathrm{y}^0$ and $l^{1'} = l^0$. Note that l^{ds}, l^{ns}, l^0, l^1 and $l^{1'}$ are the number of images in correspondence with each set of data.

The whole framework, as mentioned, consists of five steps as below:

1. *Day-Night Image Translator Training:* Two sets of data $\mathscr{D}^{ds}$ and $\mathscr{D}^{ns}$ are used to train the GAN-based image domain translator ϕ^t:

$$\min_{\phi^t}\Big(\mathscr{L}_{GAN}\big(\phi^t\big(\mathscr{D}^{ds}\big), \mathscr{D}^{ns}\big) + \mathscr{L}_{GAN}\big(\phi^t\big(\mathscr{D}^{ns}\big), \mathscr{D}^{ds}\big) + \mathscr{L}_p\big(\phi^t\big(\mathscr{D}^{ds}\big), \mathscr{D}^{ns}\big) + \mathscr{L}_p\big(\phi^t\big(\mathscr{D}^{ns}\big), \mathscr{D}^{ds}\big)\Big), \tag{1}$$

where $\mathscr{L}_{GAN}$ is the loss of the applied GAN-based model and $\mathscr{L}_p$ is the perceptual loss presented in Sec. 3.3.

2. *Day-Night Image Translation:* The set of daytime images passes through the domain translator ϕ^t to generate their nighttime versions: $\mathrm{x}_k^{1'} = \phi^t(\mathrm{x}_k^0)$.
3. *Nighttime Image Segmentation Training:* The daytime images $\mathscr{D}^0$ and synthesis nighttime images $\mathscr{D}^{1'}$ together train the segmentation model ϕ^s:

$$\min_{\phi^s}\Big(\frac{1}{l^0}\sum_{k=1}^{l^0}\mathscr{L}_{CL}\big(\phi^s\big(\mathrm{x}_k^0\big), \mathrm{y}_k^0\big) + \frac{1}{l^{1'}}\sum_{m=1}^{l^{1'}}\mathscr{L}_{CL}\big(\phi^s\big(\mathrm{x}_m^{1'}\big), \mathrm{y}_m^{1'}\big)\Big), \tag{2}$$

where $\mathscr{L}_{CL}$ is our proposed comprehensive loss function defined in Sec. 3.2.

4. *Pseudo-Labels Inference:* A set of unlabeled nighttime images $\mathscr{D}^1$ goes through the segmentor to infer pseudo-labels: $\hat{\mathrm{y}}_n^1 = \phi^s(\mathrm{x}_n^1)$
5. *Nighttime Image Segmentation Re-Training:* Together, the three sets of $\mathscr{D}^0$, $\mathscr{D}^1$ and $\mathscr{D}^{1'}$ are used to re-train the segmentor $\phi^{s'} \leftarrow \phi^s$:

$$\min_{\phi^{s'}}\Big(\frac{1}{l^0}\sum_{k=1}^{l^0}\mathscr{L}_{CL}\big(\phi^{s'}\big(\mathrm{x}_k^0\big), \mathrm{y}_k^0\big) + \frac{1}{l^{1'}}\sum_{m=1}^{l^{1'}}\mathscr{L}_{CL}\big(\phi^{s'}\big(\mathrm{x}_m^{1'}\big), \mathrm{y}_m^{1'}\big) + \frac{1}{l^1}\sum_{n=1}^{l^1}\mathscr{L}_{CL}\big(\phi^{s'}\big(\mathrm{x}_n^1\big), \hat{\mathrm{y}}_n^1\big)\Big). \tag{3}$$

Each module in our framework performs a specific task that satisfies the common target of semantic image segmentation in the dark, which is clearly described in Eq. (1), (2) and (3). Specifically, without the domain translator, the framework cannot capture nighttime features as the inputs are all in daytime condition. In the next sections, we present the work of the two components in our proposed framework and their specifications.

3.2. Nighttime scene parser

In this module of semantic image segmentation, we leverage the Panoptic Feature Pyramid Networks (Panoptic FPN) Kirillov et al. [49] as our backbone model to perform the semantic segmentation task. To the aspect of training strategy, we apply the self-training technique Zoph et al. [24].

Panoptic feature pyramid networks. Panoptic FPN is similar to a so-called Mask R-CNN He, Gkioxari, Dollar and Girshick [57] approach with a shared Feature Pyramid Network Lin, Dollár, Girshick, He, Hariharan and Belongie [58] (FPN) backbone. As mentioned by the authors, this Panoptic FPN yields a lightweight model and a top-performance method for both semantic and instance segmentation. To generate the semantic segmentation output from FPN features, there is a simple design in this network (as shown in Fig. 2) whose proposal is to merge the information from all levels of the pyramid network into a single output (segmentation map). Panoptic FPN is able to capture fine structures of images thanks to the multiple resolutions encoder stage. Moreover, the encoder can extract sufficiently rich semantic information at each resolution level to predict class labels. Based on FPN, Panoptic FPN with some refinement of the network can adapt well to the segmentation tasks.

Comprehensive loss function. In this subsection, we propose a comprehensive loss function that compromises the specifications of both cross-entropy loss and focal loss which have own impacts to the task of

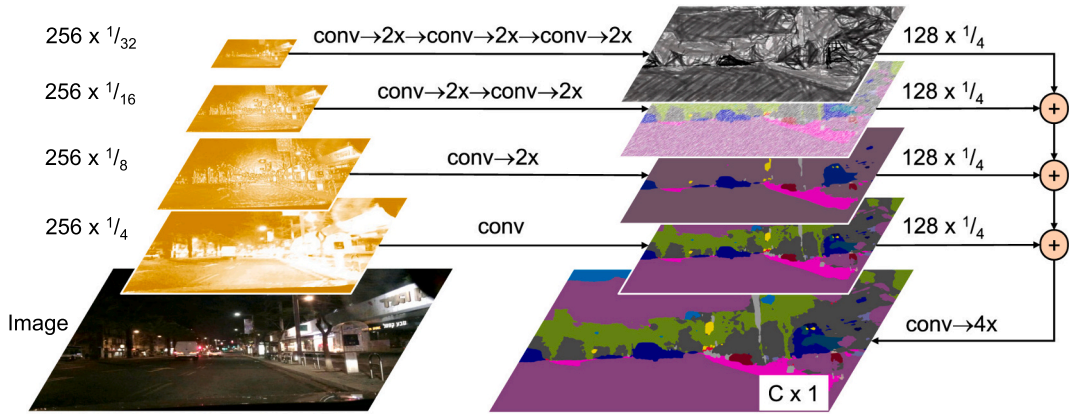

Fig. 2. The architecture of Panoptic FPN model for semantic image segmentation in our LTSP framework (Our re-implementation).

semantic image segmentation.

Either cross-entropy loss or focal loss maintains good specifications that benefit our segmentation model. From that point, we propose a comprehensive loss function that is a combination of the two mentioned functions. Cross-entropy loss measures the differences between each couple of pixels in the ground-truth and the predicted mask regardless of whether they are the major or the minor objects in the image. Here, major objects are those accounting for a large area of the image and minor objects are vice versa. Meanwhile, focal loss focuses on balancing the importance of small-area objects to be equally treated as large-area ones. Altogether, we form our novel loss function as below:

$$\mathscr{L}_{CL}(p(\mathrm{x}_i), \mathrm{y}_i) = \lambda \mathscr{L}_{ce}(p(\mathrm{x}_i), \mathrm{y}_i) + (1-\lambda)\mathscr{L}_f(p(\mathrm{x}_i), \mathrm{y}_i), \tag{4}$$

where $p(\mathrm{x}_i)$ is the predicted probability and y_i is the label of the input x_i.

Our proposed Comprehensive Loss function is presented in Eq. (4). $\mathscr{L}_{ce}$ denotes the cross-entropy loss function since this loss measures the differences among a couple of pixels. And, $\mathscr{L}_f$ denotes focal loss since this loss is able to balance the importance of minor and major pixels in the image. The effect of each component is decided by the weight of λ. From experiments, we figure out $\lambda = 0.6$ yields the finest performance of our framework with a UNIT-based image translator (see the ablation experiments in Sec. 4.3.2 for more information).

3.3. Annotation transfer

Due to the lack of annotated nighttime scenes, we adopt the annotation transfer to train the semantic scene parser mentioned above. The details are presented in the following paragraphs.

Daytime annotation transfer. In this work, we use the previously mentioned UNIT Liu et al. [38] framework as the translation converter to do the task of image domain translation. Unsupervised image-to-image translation (UNIT for short) is based on Generative Adversarial Networks (GAN) and Variational Autoencoders (VAEs). Each image domain is modeled by the VAEs. For people who are interested, please refer Liu et al. [38] to read for more information. In addition, we also leverage perceptual loss to synthesize more semantic features in the translation module. The perceptual loss formula is given as follows:

$$\mathscr{L}_p(\mathrm{x}, \hat{\mathrm{x}}) = \left\| G_j(\mathrm{x}) - G_j(\hat{\mathrm{x}}) \right\|^2 \tag{5}$$

In Eq.(5), x denotes the real image and $\hat{\mathrm{x}}$ represents the generated image. $G_j(\mathrm{x})$ are the extracted features of image x after feeding through j-th layer of the pre-trained network.

In particular, the perceptual loss takes three components as input including real image, generated image, and pre-trained network (feature extractor). The real and generated images are preprocessed before being fed through the network. We did follow the work of Johnson, Alahi and Fei-Fei [59], which returns features after all activation layers. However, to optimize the training time, we only target the output at the final activation layer ($Relu5\_3$) as output to calculate the perceptual loss. It is also worth mentioning that we used VGG-16 pre-trained as our feature extractor.

Perceptual loss helps the model to generate nighttime images which look more plausible. Qualitatively, Fig. 4 shows the results without perceptual loss (in the second column) compared to the applied ones (in the third column). The mismatched vehicles and traffic lights are appropriately diminished. For the implementation, we use the standard daytime dataset of Cityscapes Cordts et al. [16] as a resource for the image translation module to generate nighttime images. Together they become the Day-Night dataset of our work.

Pseudo-labeled nighttime scenes. In the previous section, our semantic segmentation model is trained on the combination of labeled daytime Cityscapes images and labeled synthetic nighttime images. Note that the nighttime images of Cityscapes were generated by our GAN-based image translation module. After a number of iterations training the model, we perform the label inference on the unannotated nighttime images. This step takes a set of unlabeled data as its input and uses the assumed-fine-trained model to generate pseudo-labels. These generated labels are assumed as the ground-truth of the extra data. Then, we combine the true-labeled data with the pseudo-labeled data together to re-train our model. We carefully attach the process of training a nighttime scenes parser in Fig. 1.

4. Experiments

4.1. Experimental settings

4.1.1. Dataset for daytime annotation transfer

After carefully reviewing the available datasets, we choose two datasets, namely, Cityscapes and NEXET.[1] Both are captured by the front-camera of the transportation which almost moves around the urban. Thus, the captured scenes are under consideration as they have the same distribution. However, NEXET is collected in daytime and nighttime conditions compared to the only daytime of Cityscapes.

We first use NEXET to train the UNIT model to translate daytime images to the nighttime domain in order to leverage the similarity between the two mentioned datasets. These translated images are prepared for training the segmentation model. NEXET dataset simultaneously contains 50 K images including daytime and nighttime instances. Before training, we pre-process data by filtering images with histogram equalization via a threshold β to split images into day and night categories. We divide the NEXET dataset into two sets: daytime and

[1] The NEXET dataset can be downloaded at https://www.kaggle.com/solesensei/nexet-original

nighttime instead of three sets as default (daytime, twilight, and nighttime). If the histogram of an image is lower than β_{night}, it seems to be a nighttime image. If the histogram of an image is greater than β_{day}, it seems to be a daytime image. In this case, we experimentally set $\beta_{night} = 65$, $\beta_{day} = 80$. As a result, we finally obtain 19.8 K and 19.5 K images for the daytime domain and nighttime domain to train UNIT, respectively. Visually, we provide exemplary images for daytime and nighttime conditions in Fig. 3. For the training process, we split this customed NEXET into a training set and a validation set with a ratio of 3: 1. So, we pick out 4.9 K and 4.6 K images for the daytime and nighttime validation set, respectively. And, both 14.9 K and 14.8 K images are considered as two sets of daytime and nighttime to train the domain translator.

Details of training the domain translator. We trained UNIT on the pre-processed NEXET dataset using a GPU GeForce GTX Titan X. The pre-processing step resizes images to 512×512 and randomly crops with a patch 256×256 to train the translator. The training process performs with learning rate $lr = 10^{-4}$ and learning rate scheduler step after each 100 K iterations with learning rate decay 0.5. In our experiments, the inferred results used the model at 330 K-th iteration which took about three days of training on a single mentioned GPU.

4.1.2. Dataset for nighttime scene parser

Our segmentation task uses the Cityscapes dataset Cordts et al. [16] as the main training data of our scene parser. For evaluation purpose, we choose Nighttime Driving Test Dataset Dai and Van Gool [27]. This benchmark contains 50 fine annotated nighttime urban images. As Cityscapes is a segmentation dataset containing urban daytime scenes, while our need is a set of those in nighttime conditions, we take the domain translator to create their nighttime versions. Our result is a training set of nighttime scenes which can be used for the segmentation training process. In particular, we use 2, 975 urban daytime scenes in the training set of Cityscapes and translate them into nighttime conditions to obtain 2, 975 nighttime synthetic images. Similarly, our validation set consists of 500 daytime images and 500 synthetic nighttime images. As to the fact that translating image domain only changes their day or night condition, images can share the same annotations. Altogether, original daytime and generated nighttime images are our annotated training data. In the self-training phase, we leverage nighttime scenes in NEXET to perform pseudo-labels inference.

Details of training the image segmentor. Our experiments on semantic segmentation with the self-training method were established with the Panoptic FPN model and the ResNet-101 is used as its backbone architecture (ResNet-101-FPN). Our experiments use a GPU GeForce GTX Titan X and a GPU Nvidia RTX 2080 Ti. The process of training uses a Stochastic Gradient Descent optimizer with learning rate $lr = 0.01$, $momentum = 0.9$ and $weight\_decay = 10^{-5}$ and cosine learning rate scheduler. Beside comparing our approach with other methods, we present experiments in ablation study section in order to demonstrate the effectiveness of each individual configuration.

4.2. Nighttime scene translation

To do the task of segmentation in the dark, the LTSP framework first translates daytime images to the nighttime domain. This stage prepares data for training the segmentor later. Fig. 4 visualizes the results of image-to-image translation task in the whole framework. Almost features from daytime images are relatively converted to the nighttime domain. Especially, the light of vehicles or traffic lights looks more realistic, which is solved by the additional perceptual loss. The initial results without perceptual loss (the second column) show the mismatches of various sparkling vehicle lights or traffic lights. After applying the perceptual loss, the results look more plausible and the wrongly placed vehicle lights are significantly diminished. The following segmentation model leverages the nighttime synthetic data to address nighttime scene segmentation.

4.3. Semantic segmentation

This section presents a series of experiments related to our segmentation module as well as the overall performance of our system. The segmentation module is established and enhanced by a self-training technique. In this work, we evaluate our system with the two main common segmentation evaluation metrics, which are *pixel accuracy* and *mean intersection-over-union (mIoU)*. We mainly focus on evaluating the performance of our framework by mIoU metric to compare to other methods.

4.3.1. State-of-the-art comparison

Our segmentation experiment proves the effectiveness of our model adaptation approach, using GAN to translate images between domains and leveraging the generated images to teach the model to segment in the dark. In Table 2, the first two rows are fully supervised models trained on daytime and Day-Night dataset, respectively. The rest configurations are those with self-training refinement on unlabeled images

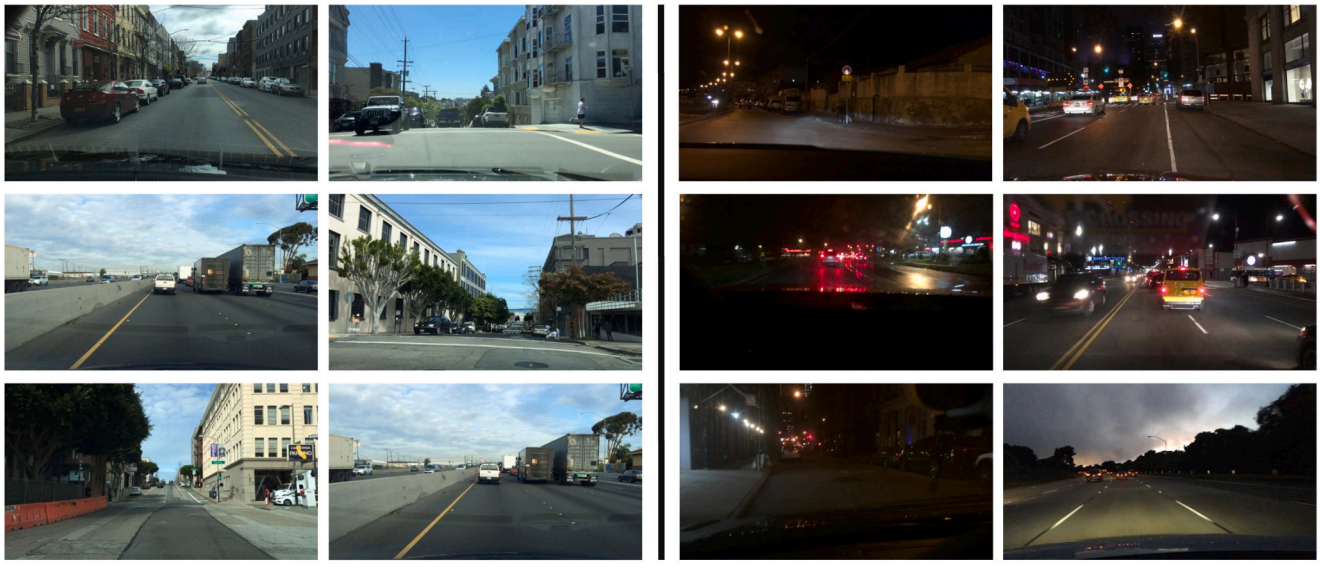

Fig. 3. Exemplary images of the dataset in daytime and nighttime conditions, split from the original NEXET dataset.

Fig. 4. Image-to-Image translation results with/without additional perceptual loss (VGG16 backbone). The results with perceptual loss better imitate nighttime conditions and get over the problem of messy light dots.

using the histogram-based method (HIS-based) and FID-based method. The results show that the combination of all refinements of our framework (the last row) achieves the finest result of 42.3% mIoU. This result sets the new state-of-the-art performance on the Nighttime Driving Test Dataset Dai and Van Gool [27] in the approach of self-training. In comparison with the initial experiment of the baseline, our improvement in mIoU increases the number of *14.8%* and in accuracy increases the number of *9.1%*. This experiment proves that our proposed loss function actually helps train our segmentation rather than using only each of its components. Besides, applying the self-training method on extra unlabeled data also contributes to the improvement of our model. The details are mentioned in the ablation study section below. To end up, we significantly improve the baseline of Panoptic FPN from 27.5% mIoU to **42.3%** and from 73.6% accuracy to **85.4%**.

As mentioned in Sec. 1, we aim to address the problem of lacking real nighttime segmentation datasets. Therefore, we do not follow the fully supervised approach, the semi-supervised approach is utilized instead. In Table 1, we report the results of our proposed LTSP framework and compare our results with other methods. The results of the compared methods are referenced from their original publications. Dai et al. Dai and Van Gool [27] leveraged self-training on twilight images to minimize the gap between the daytime and nighttime domains. Sun et al. Sun et al. [25], in contrast, performed their work with a fully supervised method which means they utilize totally annotated images and thus is not comparable to our work. Vobecky et al. Vobecky et al. [54] applied a cross-model distillation approach and Lengyel et al. Lengyel et al. [53] went with a zero-shot learning approach. Note that our Label Transfer Scene Parser uses a smaller around one-tenth of the total number of annotated images and unlabeled images compared to others. Qualitatively, Fig. 5 presents the visualization results. As can be seen, the output semantic masks do not really satisfy the visual purpose. To the best of our knowledge, there exists a gap between the synthesis images and the real samples that leads to the limitation in performance of the model.

4.3.2. Ablation study

In this section, we present a series of experiments to explain our improvement process that leads to our final results. Besides, we provide experimental evidence to customize our proposed comprehensive loss function. Table 4 represents a checklist that expresses the importance of each change to our model. Below are the explanations for each configuration we performed.

Self-training performance verification. In the first experiment, we verify the efficiency of the self-training technique in the semantic segmentation problem. Our Panoptic FPN model is trained on daytime images to result in 73.6% accuracy and 27.5% mIoU which are our initial baseline results. Then, we do two more configurations of self-training, one performs from scratch and another is from the trained checkpoint. The result of self-training from checkpoint increases an amount of 1.5% to reach 29.0% mIoU meanwhile model with self-training from scratch suffers from a slight decrease of 0.4%. It can be explained that the pseudo-labeled images have negative effects on the initial model, meanwhile, the checkpoint one has already got prior knowledge. This is equivalent to the conclusion of Zoph et al. Zoph et al. [24] about pre-training and self-training. Here, we conclude that self-training from a checkpoint gives better result compared with doing from scratch. Thus, our remaining experiments of self-training are set up from checkpoints to fine-tune our segmentation model.

Appropriate number of unlabeled data for self-training. In our LTSP framework, the self-training method is applied to improve the performance of the segmentation model. To perform self-training refinement, we use unlabeled nighttime images in NEXET as our extra data. We first use the total amount of nighttime images in NEXET as our unlabeled data. However, they show decreases in self-training results. Thus, we investigate the effect of the domination of unlabeled images on

Table 1
State-of-the-art Comparison on Nighttime Driving Test Dataset Dai and Van Gool [27].

Method	# Real Daytime	# Synthetic Nighttime	# Unlabeled Data	Training Method	mIoU (↑)
Segmenter, ViT-S/16 Vobecky et al. [54]	24,500	–	–	Unsupervised	18.9
RefineNet Lin et al. [48]	2975	–	–	Fully supervised	35.2
Dark Model Adaptation (1-step) Dai and Van Gool [27]	8000	–	26,250 twilight	Self-training 1-step twilight	39.1
W-RefineNet Lengyel et al. [53]	2975	–	–	Zero-shot learning	41.6
Dark Model Adaptation (3-step) Dai and Van Gool [27]	8000	–	8750 civil twilight + 8750 nautical twilight + 8750 astronomical twilight	Self-training 3-step twilight	41.6
LTSP (Ours)	**2975**	**2975**	**1653 nighttime**	**Self-training 1-step nighttime**	**42.3**

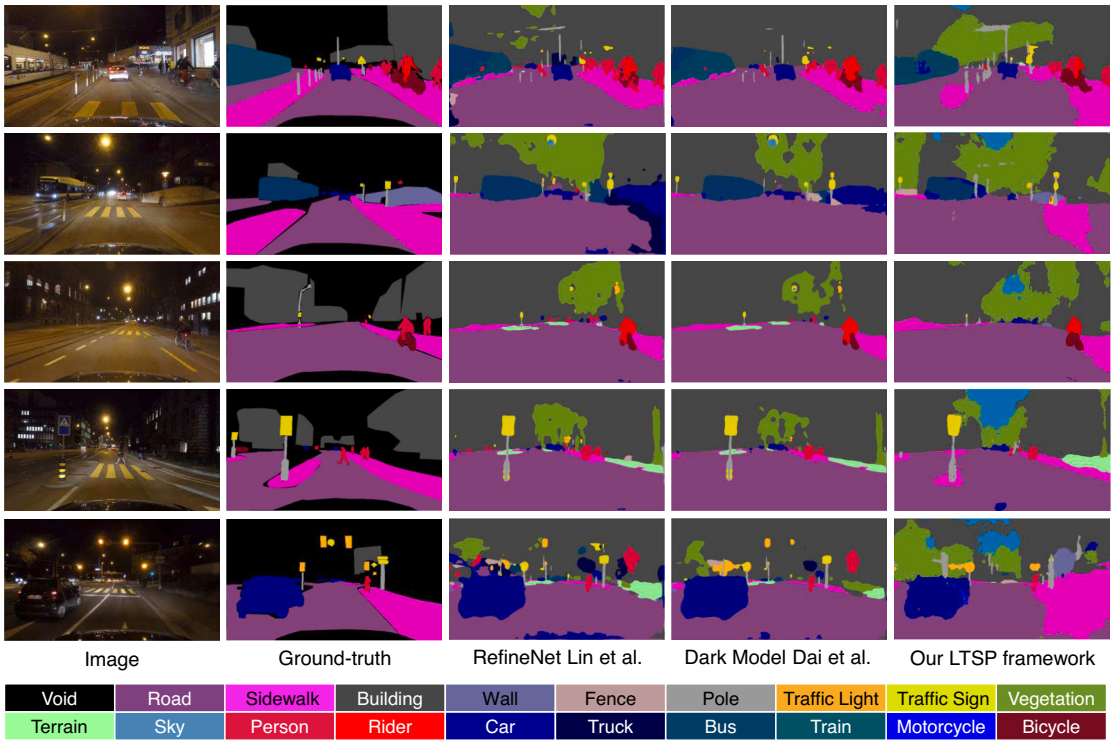

Fig. 5. Visual comparison between RefineNet Lin et al. [48], Dark Model Adaptation Method Dai and Van Gool [27] and our LTSP framework. The results of Lin et al. [48]; Dai and Van Gool [27] are referenced from its publication. Our LTSP framework yields the finest results among comparative methods despite maintaining mislabeled and over-segmented areas.

Table 2
Our variant configurations of Panoptic FPN with our proposed framework.

Configuration	# Real Daytime	# Synthetic Nighttime	# Unlabeled Nighttime	Unlabeled Data Selection	λ in Loss Function	Acc.	mIoU
w/o self-training	2975	–	–	–	1.0	73.6	27.5
		2975	–	–	1.0	81.2	34.7
	2975	2975	1600	HIS-based	1.0	83.3	37.8
					1.0	83.4	39.5
w/ self-training			1653	FID-based	0.5	83.7	40.7
					0.6	**85.4**	**42.3**

Note: All the configurations perform with fixed learning rate $lr = 0.01$, the last is with cosine learning rate scheduler.

Table 3
Ablation study on the proportion of λ to control the impact of each component in our comprehensive loss evaluated on **Nighttime Driving Test Dataset** Dai and Van Gool [27].

Configuration	λ in Loss Function	w/o Self-training		Self-training	
		Acc.	mIoU	Acc.	mIoU
Day-Night UNIT	0.4	81.96	37.97	84.21	38.79
	0.5	82.34	39.68	83.06	40.59
	0.6	82.49	40.98	**85.42**	**42.33**
	0.7	83.46	39.83	83.76	41.10
	0.8	82.82	39.01	83.44	39.47
Day-Night CycleGAN	0.4	82.55	36.69	82.95	35.15
	0.5	83.60	39.62	84.57	37.87
	0.6	83.79	38.61	84.57	36.71
	0.7	85.00	41.44	**85.28**	**41.64**
	0.8	82.44	38.51	84.52	38.14

Note: All the configurations perform with cosine learning rate scheduler with base $lr = 0.01$.

model performance. Actually, we perform self-training on an amount of nearly 15 K unlabeled nighttime images while our annotated data is 5.9 K images. This domination leads to our failure in self-training. To verify this hypothesis, we reduce the amount of unlabeled data to 1.6 K images. We choose this amount as to the ratio of labeled and unlabeled data in our initial daytime experiment, which was a successful case of self-training. Those unlabeled images are chosen based on HIS-based or FID-based methods. With a suitable proportion of unlabeled data, it is observed that the self-training strategy boosts the model performance. Using the FID-based method to choose extra data yields significant improvement since this method leverages semantic concepts of images while the HIS-based method compares hand-crafted color features. The first six rows of Table 4 present the results of this description.

Working with only nighttime domain. After trying on both daytime and nighttime images simultaneously, we experiment on only nighttime images. With the assumption that our model needs to predict nighttime testing images, we just train our model on nighttime images. The results show that training on only generated nighttime images gives a lower performance. Specifically, the mIoU metric observes 29.6% in the first stage and 29.8% with self-training in this configuration. To our explanation, when training on both daytime and nighttime images, these

Table 4
Ablation study results on Unlabeled Data and Loss function.

Configuration	Perceptual Loss	Self-training	Day-Night Dataset	Unlabeled Data Selection		λ in Loss Function	Acc.	mIoU
				HIS-based	FID-based			
Day-Night UNIT			✓			1.0	79.2	31.5
		✓	✓			1.0	78.1	28.8
	✓		✓			1.0	78.0	33.9
	✓	✓	✓			1.0	81.5	32.1
	✓	✓	✓		✓	1.0	84.3	38.8
Day-Night UNIT w/ Perceptual Loss	✓	✓	✓	✓		1.0	80.4	34.2
	✓	✓	✓	✓		0.0	76.0	28.3
	✓	✓	✓	✓		0.5	83.7	40.7
	✓	✓	✓		✓	0.6	**85.4**	**42.3**

Note: All the configurations perform with fixed learning rate $lr = 0.01$, the last is with cosine learning rate scheduler.

two domains contribute to teaching our model the features of objects in various conditions. Then, when training on only nighttime images, our model is lacks of the ability to predict on brighter nighttime images. Moreover, we also try to convert the nighttime domain of the test set to daytime, the results, however, are not higher than the original test set.

Cross-entropy loss vs. focal loss. In the previous experiments, the model is trained with a cross-entropy loss function. Thus, we test the configuration of training with focal loss to see whether focal loss can help better converge our segmentation model by overcoming problems of cross-entropy loss. Specifically, cross-entropy loss does not consider major and minor objects. From the reported results, we observe that focal loss takes more time to converge our segmentation model. However, once again we declare the help of self-training when increasing the mIoU of our system performance as well as slight improvements in accuracy. With this experiment, we conclude that focal loss does not have a great effect on our semantic segmentation model compared to cross-entropy loss. With the assumption that each loss function solves specific cases of the whole problem, we propose a comprehensive loss function (in Sec. 3.2) that combines the strength of both cross-entropy and focal loss to train a segmentation model. The rows sixth to eighth of Table 4 express the effect of the loss functions on our results via the adjustment of λ hyper-parameter.

Comprehensive loss with λ hyper-parameter. In Eq. (4), our proposed loss function includes a hyper-parameter λ that controls the effect of each componential function. To figure out which value of λ provides the finest performance to our LTSP framework, we establish ablation experiments as in Table 3. In detail, we use five different ratios $\lambda = \{0.4, 0.5, 0.6, 0.7, 0.8\}$. Here we apply the same configuration to examine the reaction of both UNIT-based and CycleGAN-based models to different λ ratios. The training and validation data is equal to the previously mentioned Day-Night dataset. The unlabeled data in the self-training stage is 1.6 K nighttime images chosen by the FID-based method. From the reported results, we observe that the ratio $\lambda = 0.6$ gives the finest result among all versions of the Day-Night UNIT model and $\lambda = 0.7$ to the Day-Night CycleGAN model. Besides, our framework with synthetic nighttime images generated by a CycleGAN Zhu et al. [37] model gets lower results than those of UNIT. This can be explained due to the lower visual quality of generated images from CycleGAN compared to UNIT (as illustrated in Fig. 6). UNIT demonstrates its ability to transfer images across the nighttime domain, especially in large areas like *sky* or *road*.

5. Conclusion and future work

In this paper, we propose a novel framework - Label Transfer Scene Parser (LTSP) for semantic image segmentation on nighttime scenes with a domain adaptation method. Our framework solves the problem of lacking annotated nighttime city street datasets for segmentation tasks by using GAN to adapt the existing daytime domain to the nighttime domain. We also leverage the self-training method to improve segmentation performance. Besides, we propose a comprehensive loss function to train our segmentation model. Furthermore, the series of experiments provide readers with lessons through each configuration. Last but not least, relying on the existing daytime dataset, our image translation

Fig. 6. Visual comparison between the results of CycleGAN and UNIT on the task of translating nighttime scenes. The translated samples of UNIT are visually better than those of CycleGAN, specifically in large areas like *sky* or *road*.

model generates a dataset for urban nighttime images which is specific to the segmentation problem.

In the future, we aim to improve this framework so as to deal with real-time video data, which is indeed a prerequisite to support autonomous vehicles. In addition, image enhancement techniques can be applied to our segmentation network to enhance segmentation in low-light conditions.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgement

This research is supported by Vingroup Innovation Foundation (VINIF) in project code VINIF.2019.DA19 and National Science Foundation (NSF) under Grant No. 2025234.

References

[1] J. Janai, F. Güney, A. Behl, A. Geiger, Computer vision for autonomous vehicles: problems, datasets and state of the art, Found. Trends ® Comput. Graph. Vision 12 (2020) 1–308.
[2] C. Chen, A. Seff, A. Kornhauser, J. Xiao, Deepdriving: Learning affordance for direct perception in autonomous driving, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015.
[3] R. Haque, M. Islam, K.S. Alam, H. Iqbal, E. Shaik, A computer vision based lane detection approach, Int. J. Image Graph. Signal Process. 11 (2019).
[4] D. Nassi, R. Ben-Netanel, Y. Elovici, B. Nassi, Mobilbye: Attacking adas with camera spoofing, 2019 arXiv preprint arXiv:1906.09765.
[5] A. Tao, K. Sapra, B. Catanzaro, Hierarchical multi-scale attention for semantic segmentation, 2020 arXiv preprint arXiv:2005.10821.
[6] R. Mohan, A. Valada, Efficientps: efficient panoptic segmentation, Int. J. Comput. Vis. 129 (2021) 1551–1579.
[7] C. Liu, L.C. Chen, F. Schroff, H. Adam, W. Hua, A.L. Yuille, L. Fei-Fei, Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 82–92.
[8] L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected crfs, in: Proceedings of the International Conference on Learning Representations (ICLR), 2015.
[9] L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2017) 834–848.
[10] L.C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, 2017 arXiv preprint arXiv:1706.05587.
[11] L.C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 801–818.
[12] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2881–2890.
[13] Y. Zhang, P. David, H. Foroosh, B. Gong, A curriculum domain adaptation approach to the semantic segmentation of urban scenes, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2020) 1823–1841, https://doi.org/10.1109/TPAMI.2019.2903401.
[14] Q. Wang, J. Gao, X. Li, Weakly supervised adversarial domain adaptation for semantic segmentation in urban scenes, IEEE Trans. Image Process. 28 (2019) 4376–4386, https://doi.org/10.1109/TIP.2019.2910667.
[15] H. Shi, H. Li, F. Meng, Q. Wu, L. Xu, K.N. Ngan, Hierarchical parsing net: semantic scene parsing from global scene to objects, IEEE Trans. Multimed. 20 (2018) 2670–2682, https://doi.org/10.1109/TMM.2018.2812600.
[16] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3213–3223.
[17] G. Neuhold, T. Ollmann, S. Rota Bulo, P. Kontschieder, The mapillary vistas dataset for semantic understanding of street scenes, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4990–4999.
[18] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, T. Darrell, Bdd100k: A diverse driving dataset for heterogeneous multitask learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2636–2645.
[19] X. Tan, K. Xu, Y. Cao, Y. Zhang, L. Ma, R.W. Lau, Night-time scene parsing with a large real dataset, IEEE Trans. Image Process. 30 (2021) 9085–9098.
[20] W. Liu, C. Zhang, G. Lin, F. Liu, Crnet: Cross-reference networks for few-shot segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
[21] K. Rakelly, E. Shelhamer, T. Darrell, A. Efros, S. Levine, Conditional networks for few-shot semantic segmentation, in: Workshop Track Proceedings of the International Conference on Learning Representations (ICLR), 2018.
[22] K. Rakelly, E. Shelhamer, T. Darrell, A.A. Efros, S. Levine, Few-shot segmentation propagation with guided networks, 2018 arXiv preprint arXiv:1806.07373.
[23] M. Abdel-Basset, V. Chang, H. Hawash, R.K. Chakrabortty, M. Ryan, Fss-2019-ncov: a deep learning architecture for semi-supervised few-shot segmentation of covid-19 infection, Knowl.-Based Syst. 212 (2021) 106647.
[24] B. Zoph, G. Ghiasi, T.Y. Lin, Y. Cui, H. Liu, E.D. Cubuk, Q. Le, Rethinking pre-training and self-training, Adv. Neural Inf. Proces. Syst. 33 (2020).
[25] L. Sun, K. Wang, K. Yang, K. Xiang, See clearer at night: towards robust nighttime semantic segmentation through day-night image conversion, in: Artificial Intelligence and Machine Learning in Defense Applications, International Society for Optics and Photonics, 2019, p. 111690A.
[26] E. Romera, L.M. Bergasa, K. Yang, J.M. Alvarez, R. Barea, Bridging the day and night domain gap for semantic segmentation, in: 2019 IEEE Intelligent Vehicles Symposium (IV), 2019, pp. 1312–1318, https://doi.org/10.1109/IVS.2019.8813888.
[27] D. Dai, L. Van Gool, Dark model adaptation: Semantic image segmentation from daytime to nighttime, in: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), 2018.
[28] S.W. Cho, N.R. Baek, J.H. Koo, M. Arsalan, K.R. Park, Semantic segmentation with low light images by modified cyclegan-based image enhancement, IEEE Access 8 (2020) 93561–93585.
[29] S. Nag, S. Adak, S. Das, What's there in the dark, in: 2019 IEEE International Conference on Image Processing (ICIP), 2019, pp. 2996–3000, https://doi.org/10.1109/ICIP.2019.8803299.
[30] K. Yang, L.M. Bergasa, E. Romera, K. Wang, Robustifying semantic cognition of traversability across wearable rgb-depth cameras, Appl. Opt. 58 (2019) 3141–3155.
[31] F.S. Saleh, M.S. Aliakbarian, M. Salzmann, L. Petersson, J.M. Alvarez, Effective use of synthetic data for urban scene semantic segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer, 2018, pp. 86–103.
[32] F. Sadat Saleh, M. Sadegh Aliakbarian, M. Salzmann, L. Petersson, J.M. Alvarez, Effective use of synthetic data for urban scene semantic segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 84–100.
[33] Y. Xu, K. Wang, K. Yang, D. Sun, J. Fu, Semantic segmentation of panoramic images using a synthetic dataset, in: Artificial Intelligence and Machine Learning in Defense Applications, International Society for Optics and Photonics, 2019, p. 111690B.
[34] H. Porav, T. Bruls, P. Newman, Don't worry about the weather: Unsupervised condition-dependent domain adaptation, in: 2019 IEEE Intelligent Transportation Systems Conference (ITSC), 2019, pp. 33–40.
[35] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, in: Advances in Neural Information Processing Systems, 2014.
[36] P. Isola, J.Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1125–1134.
[37] J.Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2223–2232.
[38] M.Y. Liu, T. Breuel, J. Kautz, Unsupervised image-to-image translation networks, Adv. Neural Inf. Proces. Syst. (2017) 700–708.
[39] X. Huang, M.Y. Liu, S. Belongie, J. Kautz, Multimodal unsupervised image-to-image translation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018.
[40] J.Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A.A. Efros, O. Wang, E. Shechtman, Toward multimodal image-to-image translation, Adv. Neural Inf. Proces. Syst. 30 (2017).
[41] Y. Choi, M. Choi, M. Kim, J.W. Ha, S. Kim, J. Choo, Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8789–8797.
[42] J. Tighe, M. Niethammer, S. Lazebnik, Scene parsing with object instance inference using regions and per-exemplar detectors, Int. J. Comput. Vis. 112 (2015) 150–171.
[43] T.V. Nguyen, C. Lu, J. Sepulveda, S. Yan, Adaptive nonparametric image parsing, IEEE Trans. Circuits Syst. Video Technol. 25 (2015) 1565–1575.
[44] T.V. Nguyen, L. Liu, K. Nguyen, Exploiting generic multi-level convolutional neural networks for scene understanding, in: 14th International Conference on Control, Automation, Robotics and Vision, ICARCV 2016, Phuket, Thailand, November 13–15, 2016, pp. 1–6.
[45] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3431–3440.

[46] H. Noh, S. Hong, B. Han, Learning deconvolution network for semantic segmentation, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1520–1528.
[47] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, 2015, pp. 234–241.
[48] G. Lin, A. Milan, C. Shen, I. Reid, Refinenet: Multi-path refinement networks for high-resolution semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1925–1934.
[49] A. Kirillov, R. Girshick, K. He, P. Dollár, Panoptic feature pyramid networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6399–6408.
[50] C. Sakaridis, D. Dai, L.V. Gool, Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7374–7383.
[51] C. Sakaridis, D. Dai, L. Van Gool, Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 44 (2020) 3139–3153.
[52] X. Wu, Z. Wu, H. Guo, L. Ju, S. Wang, Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15769–15778.
[53] A. Lengyel, S. Garg, M. Milford, J.C. van Gemert, Zero-shot day-night domain adaptation with a physics prior, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4399–4409.
[54] A. Vobecky, D. Hurych, O. Siméoni, S. Gidaris, A. Bursuc, P. Pérez, J. Sivic, Drive&segment: Unsupervised semantic segmentation of urban scenes via cross-modal distillation, in: European Conference on Computer Vision, 2022, pp. 478–495.
[55] G.J. Brostow, J. Shotton, J. Fauqueur, R. Cipolla, Segmentation and recognition using structure from motion point clouds, in: Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I 10, Springer, 2008, pp. 44–57.
[56] G.J. Brostow, J. Fauqueur, R. Cipolla, Semantic object classes in video: A high-definition ground truth database, in: Pattern Recognition Letters, Elsevier, 2009, pp. 88–97.
[57] K. He, G. Gkioxari, P. Dollar, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.
[58] T.Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2117–2125.
[59] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14, Springer, 2016, pp. 694–711.