GPS-GLASS: Learning Nighttime Semantic Segmentation Using Daytime Video and GPS data

Hongjae Lee Changwoo Han Jun-Sang Yoo Seung-Won Jung *
Korea University
{jimmy9704, hcwoo329, junsang7777, swjung83}@korea.ac.kr

Abstract

Semantic segmentation for autonomous driving should be robust against various in-the-wild environments. Nighttime semantic segmentation is especially challenging due to a lack of annotated nighttime images and a large domain gap from daytime images with sufficient annotation. In this paper, we propose a novel GPS-based training framework for nighttime semantic segmentation. Given GPS-aligned pairs of daytime and nighttime images, we perform cross-domain correspondence matching to obtain pixel-level pseudo supervision. Moreover, we conduct flow estimation between daytime video frames and apply GPS-based scaling to acquire another pixel-level pseudo supervision. Using these pseudo supervisions with a confidence map, we train a nighttime semantic segmentation network without any annotation from nighttime images. Experimental results demonstrate the effectiveness of the proposed method on several nighttime semantic segmentation datasets.

1. Introduction

Semantic segmentation, which classifies each pixel of an image into a semantic class, is a fundamental problem in computer vision and has been widely used in various applications, including autonomous driving, robotic navigation, and medical imaging. In particular, for autonomous driving applications, it is necessary to design a segmentation method that is robust against domain changes such as illumination and weather changes. In order to design such a method, especially with convolutional neural networks (CNNs), a large amount of pixel-level annotated data is required for supervised learning. However, acquiring pixel-level annotation in poor illumination environments such as nighttime is very challenging beyond the cost of annotations. Therefore, most semantic segmentation datasets focus primarily on daytime environments [2,7], but a semantic segmentation model trained on these datasets fails in nighttime semantic segmentation, as shown in Fig. 1. Although some datasets [3, 45] provide nighttime image annotations, their quantity and quality are insufficient to be used for semantic segmentation network training. In this paper, we propose a training methodology for nighttime image semantic segmentation networks without requiring pixel-level annotation of nighttime scenes.

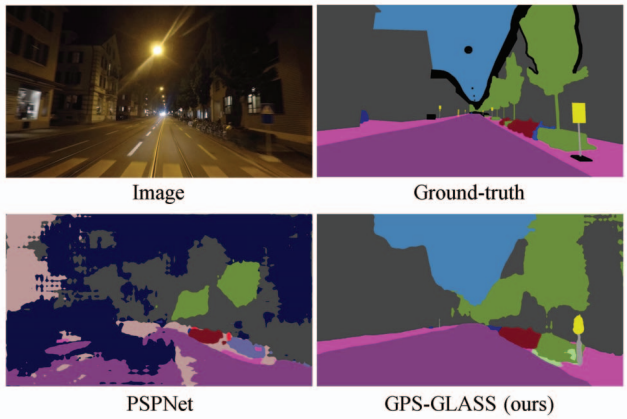

Figure 1: Visual comparison of the nighttime semantic segmentation results between the PSPNet [47] without domain adaptation and our proposed GPS-GLASS.

Several methods have been developed to adapt daytime segmentation networks to nighttime scenes without using annotated nighttime images. For example, the twilight domain between daytime and nighttime has been introduced for gradual domain adaptation [3, 28, 29]. Image translation has also been attempted to obtain synthetic annotations of nighttime images that can help train semantic segmentation networks [26, 32]. However, these methods require additional training data in the twilight domain or several preprocessing stages. Several recent methods [39,43] have presented pseudo-supervised loss terms using coarsely aligned daytime and nighttime image pairs. These recent methods require neither additional domain data nor pre-processing stages, but they have not attempted to align daytime and

*corresponding author

nighttime image pairs precisely.

In this paper, we present a novel Global Positioning System (GPS)-Guided Learning Approach for nighttime Semantic Segmentation (GPS-GLASS), as illustrated in Fig. 2. Similar to DANNet [39], GPS-GLASS uses image relighting and semantic segmentation modules and two discriminators for the daytime and nighttime domains. Unlike DANNet, GPS-GLASS extracts image features obtained during the segmentation process to estimate the correspondence from the daytime to nighttime and vice versa. Moreover, observing that nighttime images are located between daytime image frames, GPS-GLASS applies intra-domain correspondence matching to daytime image frames and performs GPS-based flow scaling. From these inter-domain and intra-domain correspondences, we construct pseudo-labels for training a nighttime semantic segmentation network. In addition, due to the cross-domain correspondence matching the proposed GPS-GLASS well generalized to both daytime and nighttime.

Our contributions are summarized as follows:

- We introduce a framework called GPS-GLASS that performs inter-domain correspondence matching to construct a pseudo-label for training a nighttime semantic segmentation network.
- We propose to perform intra-domain correspondence matching using daytime video frames and scale the estimated flow field using GPS data, yielding another pseudo-label.
- By combining the two pseudo-labels with a confidence map, GPS-GLASS shows state-of-the-art performance on several nighttime image datasets. Ablation studies also verify the effectiveness of each component of GPS-GLASS.

2. Related Work

2.1. Unsupervised Domain Adaptation for Nighttime Semantic Segmentation

Supervised training of semantic segmentation networks requires pixel-level annotation, which is laborious and time-consuming to obtain. Because ground-truth annotation is publicly available only for some limited domains, e.g., Cityscapes [2] for daytime road scenes and GTA5 [25] for synthetic scenes, unsupervised domain adaptation (UDA) has received significant interest. There have been several approaches [19, 36] to achieve the goal of UDA for semantic segmentation. However, these approaches are focused on reducing the domain gap between synthetic and real.

The existing datasets developed for semantic segmentation of road scenes are biased toward daytime scenes [2, 7], segmentation networks trained without considering UDA tend to fail in handling nighttime scenes. Consequently, efforts have been made to reduce the domain gap between daytime and nighttime scenes.

Motivated by the widely studied image style transfer [12, 13, 21], one can try translating daytime scenes to nighttime scenes. Earlier studies along this direction [26, 32] were overly interested in the auxiliary task of the image style transfer rather than the main semantic segmentation task. Sakaridis *et al.* [28] constructed the *Dark Zurich* dataset, which contains daytime and nighttime images that are coarsely matched with GPS information. These daytime and nighttime image pairs are helpful in guiding the semantic segmentation of nighttime scenes, resulting in many follow-up studies, including their guided curriculum model adaptation (GCMA). Sakaridis *et al.* [29] proposed an improved version of GCMA that uses depth and camera pose information. Wu *et al.* [39] proposed a multi-target domain adaptation network for nighttime semantic segmentation via adversarial learning. Xu *et al.* [43] proposed nighttime domain gradual self-training and patch-level prediction guidance methods. Gao *et al.* [6] proposed a correlation distillation approach for cross-domain between synthetic and real nighttime.

However, the above methods [28, 29, 39, 43] have not attempted to precisely align daytime and nighttime images because such an alignment can even be a more difficult task than semantic segmentation. Xu *et al.* [40] aligned the daytime and nighttime images using an additional optical flow estimation network. However, the above method requires additional datasets and training stages for the optical flow estimation network. We notice that the invaluable information of GPS is obtainable when constructing datasets such as Dark Zurich. Consequently, in this paper, we propose to use the GPS information of daytime and nighttime images to guide the correspondence matching for nighttime semantic segmentation network training.

2.2. Optical Flow and Correspondence Matching

Various studies are being conducted to find a matching point between two images e.g., stereo matching [22, 24, 44], optical flow estimation [5, 10], and semantic correspondence [17, 18]. In the case of learning-based optical flow estimation approaches, Dosovitskiy *et al.* [5] introduced an end-to-end optical flow estimation method with CNNs. Ranjan *et al.* [23] proposed a spatial pyramid network that predicts flow in a coarse-to-fine manner. Sun *et al.* [31] proposed a method of warping the spatial feature pyramid and calculating the cost volume from the warped features. Teed *et al.* [34] proposed a recurrent unit for gradual flow refinement, demonstrating high performance with fewer network parameters. Recent transformer [37]-based optical flow models [14, 41, 42] further improved the optical flow estimation performance.

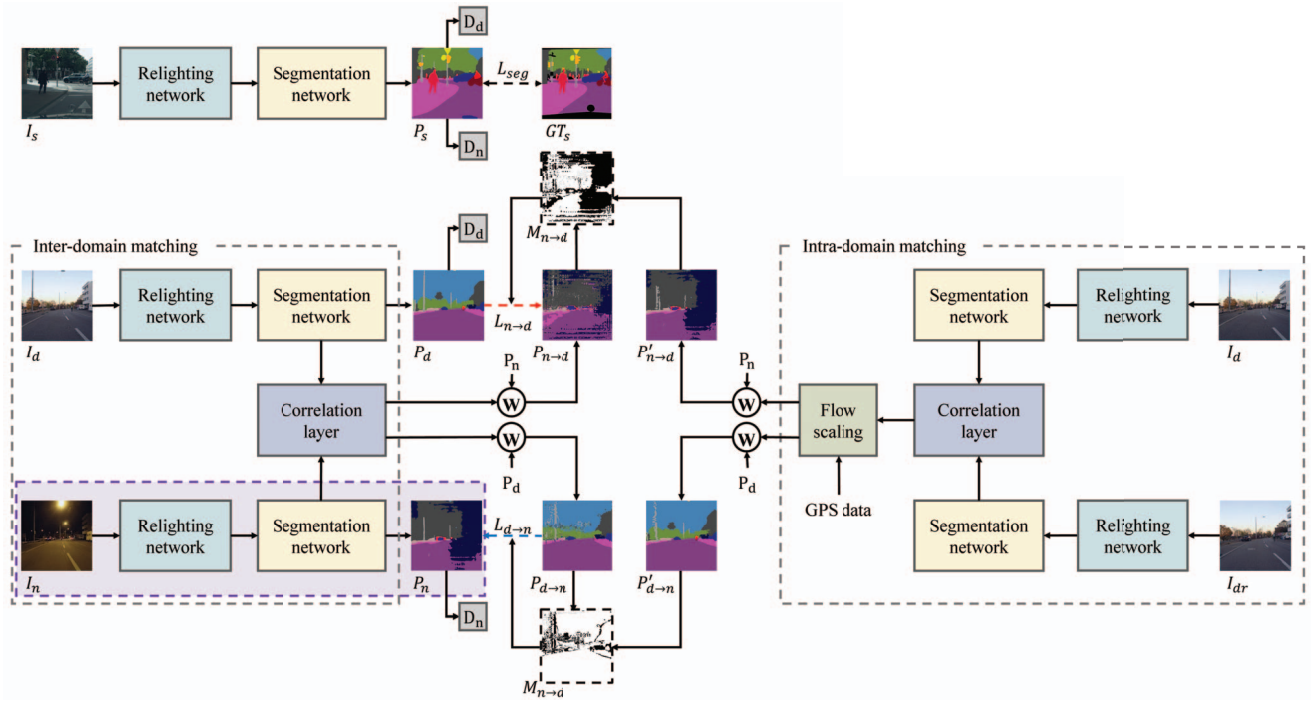

Figure 2: The overview of the proposed GPS-GLASS. The same colored networks share weights, and the correlation layer has no weights that require training. Ⓦ represents the backward warping operator, and the red and blue dotted arrows indicate supervision by the ground-truth and pseudo-labels, respectively. Only the networks inside the purple box are used at the inference stage.

Although the above methods have contributed to the development of optical flow estimation technology, they are still suffering from correspondence matching between different domains. Recently, Zhou *et al.* [48] estimated optical flow to match and align two images captured under different weather conditions. Zhang *et al.* [46] proposed a method to train an image translation network by warping images from different domains. Lee *et al.* [17, 18] introduced a model called SFNet that predicts bidirectional correspondence between different instances of the same object or scene category. Inspired by the superior performance of SFNet, we propose a dense corresponding matching method in different domains for nighttime semantic segmentation. A simple correlation layer trains the segmentation network to extract features that are invariant to the domain gap, e.g., between daytime and nighttime.

3. Proposed Methods

3.1. Framework Overview

Our method involves a source domain S and two target domains T_d and T_n, where S, T_d, and T_n correspond to Cityscapes (daytime) [2], Dark Zurich-D (daytime), and Dark Zurich-N (nighttime) [28] datasets in our case study, respectively. Note that only the source domain has ground-truth segmentation labels, and the two target domains are coarsely paired according to GPS locations. As shown in Fig. 2, our GPS-GLASS consists of a single weight-sharing relighting network (G_R), a single weight-sharing semantic segmentation network (G_S), and two discriminators (D_d and D_n), where we used the same architecture of DANNet [39] for these network components. Let I_s, I_n, and I_d denote image samples corresponding to S, T_d, and T_n, respectively. These images are fed to G_R to make G_S less sensitive to illumination changes [39]. The segmentation results are obtained as $P_s = G_S(G_R(I_s))$, $P_d = G_S(G_R(I_d))$, and $P_n = G_S(G_R(I_n))$. Only P_s has its corresponding ground-truth segmentation labels P_s^*, and the other two results P_d and P_n are supervised by the pseudo-label.

Specifically, the proposed training framework called GPS-GLASS obtains the pseudo-label by estimating dense correspondence between the daytime and nighttime images, where the correlation layer is applied to the intermediate features of the segmentation network. In addition, since the dense correspondence between daytime and nighttime images can be inaccurate, GPS-GLASS obtains another pseudo-label by estimating dense correspondence between the daytime images and applying GPS-based flow scaling. By using the two different sources for acquiring the pseudo-

label, GPS-GLASS trains the nighttime semantic segmentation network without any annotation from nighttime images. The details of GPS-GLASS will be explained in the following subsections.

3.2. GPS-guided Learning Approach

3.2.1 Correspondence matching using inter-domain

Our key idea is to align P_d and P_n such that the aligned segmentation result can be used as the pseudo-label. To this end, inspired by SFNet [18], the correlation layer is adopted to compute the dense correspondence of the image features between two different domains. A simple correlation layer without trainable parameters allows the segmentation network to extract features that are robust to domain changes, such as daytime and nighttime. Let $f^d = \left\{f_l^d, f_g^d\right\}$ be the set of the local and global features extracted from the semantic segmentation network for the input I_d. In the case of PSPNet [47], which is our chosen architecture for semantic segmentation, f_l^d and f_g^d are extracted before and after passing through the PSPmodule of PSPNet, respectively, and have the same dimension of $H \times W \times D$. $f^n = \left\{f_l^n, f_g^n\right\}$ is extracted similarly from the semantic segmentation network for the input I_n. Then, the correlation layer computes the correlation between f^d and f^n as follows:

$$c_x\left(\mathbf{p}, \mathbf{q}\right) = \left(\frac{f_x^d\left(\mathbf{p}\right)}{\left\|f_x^d\left(\mathbf{p}\right)\right\|}\right)^\top \left(\frac{f_x^n\left(\mathbf{q}\right)}{\left\|f_x^n\left(\mathbf{q}\right)\right\|}\right), x \in \left\{l, g\right\}, \tag{1}$$

where $\top$ is the transpose operator, $\|\cdot\|$ measures L2 norm, $\mathbf{p}$ and $\mathbf{q}$ represent 2D coordinates, and $f_x^d\left(\mathbf{p}\right)$ and $f_x^n\left(\mathbf{q}\right)$ are the D-dimensional vectors at $\mathbf{p}$ and $\mathbf{q}$, respectively. We combine the correlation volumes obtained from the local and global features by $c = c_l \odot c_g$, where $\odot$ represents element-wise multiplication. Instead of using the standard argmax function to obtain the correspondence from c, we use soft-argmax [9, 15] to allow backpropagation through the correlation layer as follows:

$$c'\left(\mathbf{p}, \mathbf{q}\right) = \frac{\exp\left(\alpha \cdot c\left(\mathbf{p}, \mathbf{q}\right)\right)}{\sum\limits_{\mathbf{q}' \in \mathbf{Q}} \exp\left(\alpha \cdot c\left(\mathbf{p}, \mathbf{q}'\right)\right)}, \tag{2}$$

where $\mathbf{Q}$ is the set of 2D positions in f^n, and α is the temperature parameter. Note that soft-argmax converges to argmax as α increases, but an excessively high value of α can lead to unstable gradient flow during training. From the grid search of α, we chose $\alpha = 10^4$ in our experiments. The optical flow field from the daytime to nighttime $F_{d\to n}$ is obtained as

$$F_{d\to n}\left(\mathbf{p}\right) = \sum_{\mathbf{q} \in \mathbf{Q}} c'\left(\mathbf{p}, \mathbf{q}\right) \cdot \mathbf{q}. \tag{3}$$

The optical flow field from the nighttime to daytime $F_{n\to d}$ is obtained in a similar manner by switching $\mathbf{p}$ and $\mathbf{q}$ in Eqs. (1)-(3). Finally, the semantic segmentation map warped from nighttime to daytime, denoted as $P_{n\to d}$, is obtained using $F_{d\to n}$ and P_n by backward warping. Similarly, the semantic segmentation map warped from daytime to nighttime, denoted as $P_{d\to n}$, is obtained using $F_{n\to d}$ and P_d. Fig 3 shows some examples of $P_{n\to d}$ and $P_{d\to n}$. Although these warped predictions are imperfect, $P_{n\to d}$ ($P_{d\to n}$) is expected to be close to P_d (P_n). Therefore, we can use $P_{n\to d}$ and $P_{d\to n}$ for the nighttime semantic segmentation network training.

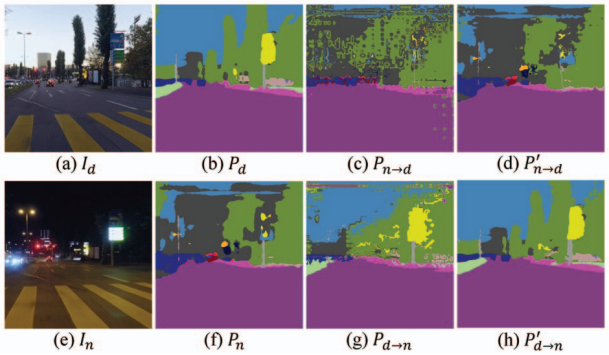

Figure 3: Examples of the segmentation results and warped pseudo-labels obtained during training. (a) and (e) are the input images, and (b) and (f) are the corresponding segmentation results. (c), (g) and (d), (h) are the results using the corresponding matching in the inter and intra domains, respectively.

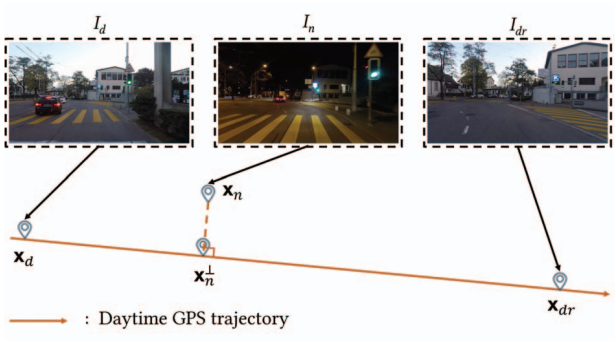

Figure 4: Illustration of the nighttime, daytime, and daytime reference images with their corresponding GPS positions.

3.2.2 Pseudo-supervision using intra-domain matching

Due to the suboptimal performance of the relighting network, dense correspondence matching between daytime and nighttime is still challenging. Observing that most existing semantic image segmentation datasets [2, 7, 29] provide video frames, we propose to use another daytime reference image, denoted as I_{dr}, for generating an additional

pseudo-label. In the Dark Zurich dataset, I_n is the nearest nighttime image of I_d, but the neighboring frames of I_d along the forward and backward directions, denoted as I_d^+ and I_d^-, are also available. From the GPS positions of these images, we can determine I_{dr} as either I_d^+ or I_d^-.

Specifically, let $\mathbf{x}_d$, $\mathbf{x}_d^+$, $\mathbf{x}_d^-$, and $\mathbf{x}_n$ denote the GPS positions of I_d, I_d^+, I_d^-, and I_n, respectively. Here, each GPS position is given as a 2D vector containing the latitude and longitude. Then, I_{dr} is determined as follows:

$$I_{dr} = \begin{cases} I_d^+, \text{ if } CS\left(\mathbf{x}_d - \mathbf{x}_n, \mathbf{x}_d^+ - \mathbf{x}_n\right) \\ \qquad < CS\left(\mathbf{x}_d - \mathbf{x}_n, \mathbf{x}_d^- - \mathbf{x}_n\right), \\ I_d^-, \text{ otherwise,} \end{cases} \tag{4}$$

where CS measures the cosine similarity. In other words, as illustrated in Fig. 4, if I_n is located along the forward direction of I_d, I_d^+ is chosen as I_{dr}. Otherwise, I_d^- is chosen as I_{dr}.

Given I_d and I_{dr}, we obtain the optical flow field from the daytime to daytime reference, denoted as $F_{d\to dr}$, by following the same procedure of Eqs. (1)-(3) but with the features f^d and f^{dr}, where $f^{dr} = \left\{f_l^{dr}, f_g^{dr}\right\}$ is the feature extracted from I_{dr}. The optical flow field from the daytime reference to daytime $F_{dr\to d}$ is obtained similarly. For the generation of the pseudo-label for the nighttime semantic segmentation network training, another pair of the optical flow fields are obtained as $F'_{d\to n} = \lambda F_{d\to dr}$ and $F'_{n\to d} = \lambda F_{dr\to d}$. The scale factor λ is chosen as

$$\lambda = \frac{HD\left(\mathbf{x}_d, \mathbf{x}_n^{\perp}\right)}{HD\left(\mathbf{x}_d, \mathbf{x}_{dr}\right)}, \tag{5}$$

where HD measures the Haversine distance of two positions [11], and $\mathbf{x}_n^{\perp}$ represents the position projected onto the line joining $\mathbf{x}_d$ and $\mathbf{x}_{dr}$, as illustrated in Fig. 4. Finally, the semantic segmentation map warped from nighttime to daytime, denoted as $P'_{n\to d}$, is obtained using $F'_{d\to n}$ and P_n by backward warping. Similarly, the semantic segmentation map warped from daytime to nighttime, denoted as $P'_{d\to n}$, is obtained using $F'_{n\to d}$ and P_d. Fig 3 shows some examples of $P'_{n\to d}$ and $P'_{d\to n}$. We now have four warped segmentation maps, i.e., $P_{n\to d}$, $P_{d\to n}$, $P'_{n\to d}$, and $P'_{d\to n}$, which are used for the nighttime semantic segmentation network training.

3.2.3 Confidence map

The first pair of the warped predictions, i.e., $P_{n\to d}$ and $P_{d\to n}$, can be inaccurate due to imperfect relighting and flow estimation. The second pair of the warped predictions, i.e., $P'_{n\to d}$ and $P'_{d\to n}$, can also be inaccurate because $\mathbf{x}_n$ is generally not lying on the line joining $\mathbf{x}_d$ and $\mathbf{x}_{dr}$, and thus the simple scaling by λ can lead to imprecise flow fields. Moreover, GPS positions are not always precise. We thus define a 2D confidence map such that only consistent predictions are used for pseudo-supervision. Specifically, the confidence map for the nighttime to daytime warping, denoted as $M_{n\to d}$, is defined as follows:

$$\mathbf{I}(\mathbf{p}) = \left\{\mathbf{i} \middle| \operatorname{argmax}\left(P_{n\to d}\left(\mathbf{p}\right)\right) = \operatorname{argmax}\left(P'_{n\to d}\left(\mathbf{p}+\mathbf{i}\right)\right)\right\}, \tag{6}$$

$$M_{n\to d}\left(\mathbf{p}\right) = \begin{cases} 1, & \text{if } \exists \mathbf{I}(\mathbf{p}) \in \Omega, \\ 0, & \text{otherwise,} \end{cases} \tag{7}$$

where Ω is a set of positions in the 3×3 kernel. $P_{n\to d}\left(\mathbf{p}\right)$ extracts the C-dimensional vector at $\mathbf{p}$, where C is the number of semantic classes. The confidence map for the daytime to nighttime warping, denoted as $M_{d\to n}$, can be defined in a similar manner. These binary confidence maps are used when training the nighttime semantic network.

3.3. Objective Functions

We use five loss terms for GPS-GLASS: light loss L_{light}, semantic segmentation loss L_{seg}, adversarial loss L_{adv}, discriminator loss L_{dis}, and warping loss. Because we use the same loss functions defined in DANNet for the first four terms [39], we only detail the warping loss in this subsection.

We now have $P_{n\to d}$ and $P_{d\to n}$ and their confidence maps $M_{n\to d}$ and $M_{d\to n}$, which can be used to supervise the training of the nighttime semantic segmentation network. Note that $P'_{n\to d}$ and $P'_{d\to n}$ are integrated to $P_{n\to d}$ and $P_{d\to n}$ since only consistent predictions are used by the confidence maps. First, we use $P_{d\to n}$ for the pseudo-supervision of P_n. Specifically, the first warping loss term $L_{d\to n}$ is defined as follows:

$$H(P_{d\to n}(\mathbf{q}), P_n(\mathbf{q})) = \sum_{k\in\mathbb{C}} E_o(P_{d\to n}\left(\mathbf{q}; k\right))\log P_n\left(\mathbf{q}; k\right), \tag{8}$$

$$L_{d\to n} = -\frac{1}{N_p \cdot C} \sum_{\mathbf{q}\in\mathbf{Q}^-} M_{d\to n}(\mathbf{q}) H(P_{d\to n}(\mathbf{q}), P_n(\mathbf{q})), \tag{9}$$

where H measures the cross entropy, E_o denotes the one-hot encoding [39], $\mathbb{C}$ is a set of all semantic segmentation classes, N_p is the number of pixels. $P_n\left(\mathbf{q}; k\right)$ represents the probability of the k-th object class at the position $\mathbf{q}$ of P_n. Note that the cross-entropy loss is measured only for the reliable prediction with $M_{d\to n}\left(\mathbf{q}\right) = 1$. Here, we define a set of ignore indexes, $\tilde{\mathbf{Q}}$, as follows:

$$\tilde{\mathbf{Q}} = \left\{\mathbf{q} \middle| \begin{array}{l} \operatorname{argmax}\left(P_n\left(\mathbf{q}\right)\right) \in C_{dyn}, \\ \operatorname{argmax}\left(P_n\left(\mathbf{q}\right)\right) \neq \operatorname{argmax}\left(P_{d\to n}\left(\mathbf{q}\right)\right) \end{array}\right\}, \tag{10}$$

where $\mathbb{C}_{dyn}$ is a set of dynamic semantic classes, including cars, people, etc. Then, $\mathbf{Q}^-$ in Eq. (10) is defined as $\mathbf{Q}^- = \mathbf{Q} \cap \tilde{\mathbf{Q}}^c$. We found this special handling is necessary to prevent undesirable pseudo-supervision of dynamic object classes.

The second warping loss $L_{n\to d}$ is defined as follows:

$$L_{n\to d} = -\frac{1}{N_p \cdot C} \sum_{\mathbf{p}\in\mathbf{P}^-} M_{n\to d}(\mathbf{p}) H(P_d(\mathbf{p}), P_{n\to d}(\mathbf{p})), \tag{11}$$

where $\mathbf{P}^-$ is defined in a similar manner as $\mathbf{Q}^-$. In other words, P_d is used as the pseudo-supervision of $P_{n\to d}$ for the nighttime segmentation network training.

The objective functions for the target daytime and nighttime domains, L_{T_d} and L_{T_n}, and the source domain L_S are defined as:

$$L_{T_d} = \mu_1 L_{light} + \mu_2 L_{adv}, \tag{12}$$

$$L_{T_n} = \mu_1 L_{light} + L_{n\to d} + L_{d\to n} + \mu_2 L_{adv}, \tag{13}$$

$$L_S = \mu_1 L_{light} + \mu_3 L_{seg} + \mu_4 L_{dis}, \tag{14}$$

where μ_1, μ_2, μ_3, and μ_4 are empirically chosen as 0.01, 0.01, 1, and 1, respectively, following the baseline method DANNet [39]. In every training iteration of GPS-GLASS, we sequentially optimize L_{T_d}, L_{T_n}, and L_S for daytime, nighttime, and source domains, respectively.

4. Experimental Results

4.1. Datasets

The **Cityscapes dataset [2]** includes 5,000 images taken in street scenes with pixel-level annotations for a total of 19 categories. We used the training set (2,975 images) as the labeled source domain S in the GPS-GLASS training stage. The **Dark Zurich dataset [28]** includes 2,416 nighttime images, 2,920 twilight images and 3,041 daytime images for training, which are all unlabeled with the size of 1,920×1,080. The images across different domains are coarsely paired according to the GPS distance-based nearest neighbor assignment. Consequently, most of these images share many image contents that are valuable for domain adaptation in semantic segmentation. Following the previous works [39, 43], we only used 2,416 day-night image pairs in the training stage of GPS-GLASS as the unlabeled target domains, T_d and T_n. For the quantitative performance evaluation, the Dark Zurich dataset provides 50 finely annotated nighttime images, which are also used for our ablation study.

The **ACDC-night dataset [30]** is an extended version of the Dark Zurich dataset, including 1006 nighttime images (400, 106, and 500 images for training, validation, and test). The dataset also provides finely annotated nighttime images as the Dark Zurich dataset. The performance evaluation on the ACDC-night dataset was conducted on a test set using an online evaluation website [27].

The **NightCity+ dataset [4]** is an extended version of the NightCity dataset [33] that re-annotates incorrectly labeled regions of the validation set. We used the NightCity+ validation dataset (1,299 images) only for the performance evaluation.

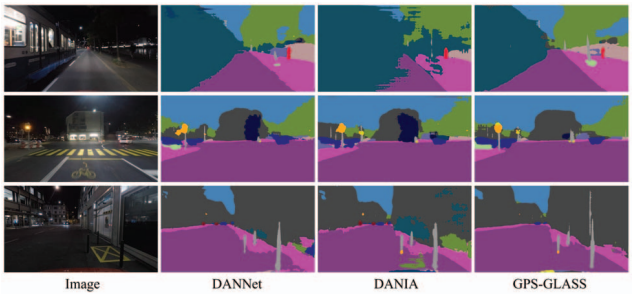

Figure 5: Visual comparison of our GPS-GLASS with other state-of-the-art methods on ACDC-night.

4.2. Implementation Details

We implemented GPS-GLASS using PyTorch. The training was performed with a single Nvidia Titan RTX GPU. Following [1], we trained our network using the stochastic gradient descent optimizer with a momentum of 0.9 and a weight decay of 5×10^{-4}. We used Adam optimizer [16] for training the discriminators with β of 0.9 and 0.99. The initial learning rate of the generator and discriminators was set to 2.5×10^{-4} and then reduced to the power of 0.9 using the poly learning rate policy. For data augmentation, random cropping of the size 512×512 was applied with a scale factor between 0.5 and 1.0 for the Cityscapes dataset, and random cropping of the size 960×960 was applied with a scale factor between 0.9 and 1.1 for the Dark Zurich dataset. In addition, we applied random horizontal flips for training. We used PSPNet [47] as the segmentation network model, which has shown state-of-the-art performance in nighttime semantic segmentation. We pre-trained PSPNet on the Cityscapes dataset for 150K iterations using L_{seg}. Then, we set the batch size to 2 and trained the model for 100K iterations.

4.3. Performance Comparisons

4.3.1 Comparison on ACDC-night and Dark Zurich

We compared GPS-GLASS with several state-of-the-art domain adaptation-based nighttime semantic segmentation methods, including DANNet [39], DANIA [40], MGCDA [29], GCMA [28], DMAda [3], and CCDistill [6]. For the comparison with the other techniques, BDL, AdaptSegNet, ADVENT [20, 35, 38] were also evaluated, where they were trained to adapt from Cityscapes to Dark Zurich. We report the mean intersection over union (mIoU) as the evaluation metric. For accurate performance comparison, we used ACDC-night, an extended version of Dark Zurich, which provides a large number of images with difficult object classes to be segmented. Table 1 reports the mIoU results on ACDC-night. All of these compared methods used the common ResNet-101 [8] as a backbone. We used DANNet and DANIA with PSPNet [47] for a fair comparison

Table 1: Performance comparison on ACDC-night. The best and second-best results are boldfaced and underlined, respectively.

Method	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motocycle	bicycle	**mIoU**
PSPNet	75.5	16.3	47.3	14.5	10.4	23.2	29.0	22.8	40.5	10.8	12.0	39.2	15.3	44.3	2.6	23.0	37.5	13.8	27.9	26.6
DMAda	74.7	29.5	49.4	17.1	12.6	31.0	38.2	30.0	48.0	22.8	0.2	47.0	25.4	63.8	12.8	46.1	23.1	24.7	24.6	32.7
GCMA	78.6	45.9	58.5	17.7	18.6	37.5	**43.6**	**43.5**	58.7	39.2	22.5	**57.9**	29.9	**72.1**	21.5	56.3	41.8	35.7	35.4	42.9
MGCDA	74.5	52.5	69.4	7.7	10.8	**38.4**	40.2	43.3	61.5	36.3	37.6	55.3	25.6	71.2	10.9	46.4	32.6	27.3	33.8	40.8
DANNet	90.7	61.2	75.6	35.9	28.8	26.6	31.4	30.6	**70.8**	39.4	78.7	49.9	28.8	65.9	24.7	44.1	**61.1**	25.9	34.5	47.6
DANIA	91.0	60.9	**77.7**	40.3	**30.7**	34.3	37.9	34.5	70.0	37.2	79.6	45.7	**32.6**	66.4	11.1	37.0	60.7	32.6	**37.9**	48.3
CCDistill	90.0	60.7	75.6	**42.0**	28.3	27.5	29.2	32.2	67.7	36.0	77.4	46.7	24.2	69.7	**48.2**	45.4	53.9	**40.5**	36.0	49.0
GPS-GLASS	**91.8**	**65.0**	76.4	38.1	30.0	35.8	38.5	37.6	69.2	**41.4**	**79.8**	45.8	31.2	69.6	38.0	**59.9**	45.7	24.9	37.2	**50.3**

Table 2: Performance comparison on Dark Zurich-val and NightCity+.

Method	mIoU	
	Dark Zurich-val	NightCity+
PSPNet	12.28	19.04
GCMA	26.65	-
MGCDA	26.10	-
DANNet	36.76	29.93
DANIA	38.14	28.92
GPS-GLASS	**38.19**	**31.81**

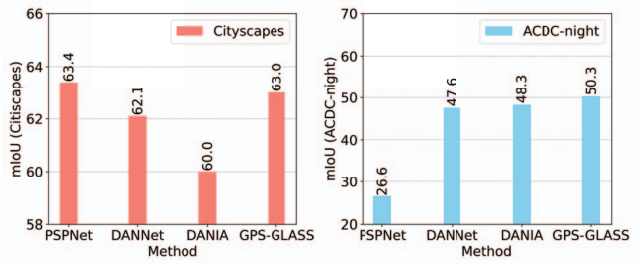

Figure 6: Performance comparison of daytime and nighttime semantic segmentation results for our GPS-GLASS and other state-of-the-art methods.

with our GPS-GLASS.

GPS-GLASS achieved a 1.3% performance improvement in terms of the mIoU over the second-best method, CCDistill. Note that GPS-GLASS does not increase the number of network parameters or processing time compared to DANNet because the same architecture of PSPNet is used in the inference stage. The performance improvements are significant in several categories, such as road, sidewalk, terrain, and sky, which are difficult to identify in nighttime scenes. Meanwhile, due to pixel-level aligned pseudo-supervision, improvements are also noticeable in small-scale classes such as poles, lights, and sign, compared to the baseline method, DANNet. Consistent results were obtained from Dark Zurich-val as shown in Table 2. These results indicate that our approach effectively performed the domain adaptation from the daytime to nighttime. Fig. 5 shows several results for visual comparison.

4.3.2 Generalization Ability for Nighttime

To show the generalization ability of our proposed method, we tested our model trained on Dark Zurich to NightCity+. As shown in Table 2, GPS-GLASS achieved a 1.88% performance improvement in terms of the mIoU over the second-best method, DANNet. This result demonstrates that the proposed GPS-GLASS trained on Dark Zurich generalizes well to another challenging nighttime dataset.

4.3.3 Generalization Ability for Daytime

One of the technical challenges of nighttime domain adaptation is the generalization ability for daytime. To this end, we compared the proposed GPS-GLASS with PSPNet [47] (trained on Cityscapes) as well as DANNet and DANIA (pre-trained on Cityscapes and then trained with Dark-Zurich via UDA). As shown in Fig. 6, DANNet and DANIA showed noticeable performance drops from PSPNet on Cityscapes. However, our proposed GPS-GLASS achieved state-of-the-art performance at nighttime at the sacrifice of a 0.4% performance reduction at daytime. We consider that the correlation layer of GPS-GLASS enabled the segmentation network to extract domain-invariant features through correspondence matching, resulting in high performance across two domains.

We conducted additional analysis for dense correspon-

Figure 7: Comparisons of the correspondence matching results obtained by the proposed GPS-GLASS, DANNet, and DANIA.

Table 3: Ablation study on several model variants of our method on Dark Zurich-val.

Method		mIoU	Gain
w/o pseudo-supervision		24.68	-
DANNet		36.76	+12.08
inter-intra mixing	avg	35.92	+11.24
	max	35.70	+11.02
warping loss	$L_{n\to d}$	33.90	+9.22
	$L_{d\to n}$	36.06	+11.38
feature matching	local	37.45	+12.77
	global	36.67	+11.99
matching domain	inter	34.86	+10.18
	intra	36.21	+11.53
GPS-GLASS		**38.19**	**+13.51**

dence matching between nighttime and daytime images. To this end, we computed the cosine similarity between the feature of the nighttime image at the red cross position, as shown in Fig. 7, and the features of the daytime image, where the features were obtained before the last layer of the segmentation network. The pixels with similarity score higher than the threshold are colored in red, where the threshold was set to 0.25 for all methods and scenes. As shown in Fig. 7, GPS-GLASS assigned high similarities for the pixels with the semantic class. However, DANNet and DANIA resulted in high similarities for many pixels belonging to different semantic classes (e.g., building, sky, sidewalk, and road). These results indicate that the correlation layer drives the segmentation network to extract features with high similarities to the pixels with the same semantic classes between daytime and nighttime images.

4.4. Ablation Study

In order to demonstrate the effectiveness of individual components of GPS-GLASS, several modified models of GPS-GLASS were trained, and the best performances in Dark Zurich-val are reported in Table 3. GPS-GLASS without any pseudo-supervision serves as a naive baseline, which leads to the lowest mIoU of 24.68. Due to the static loss [39], DANNet achieved a 12.08% mIoU increase compared to the baseline. We applied other inter-intra pseudo-label mixing methods: taking the average of two pseudo-labels or taking the label with the higher probability for each pixel. For Dark Zurich-val, these two methods, denoted as avg and max in Table 3, increased the mIoU by 11.24% and 11.02%, respectively, which are worse than the performance improvement obtained using the confidence map (13.51%). Both warping loss terms, $L_{n\to d}$ and $L_{d\to n}$, were found to be essential compared to their single-use. In addition, because we obtained the integrated correlation volume by element-wise multiplication of the correlation volumes from the local and global features, we evaluated the performance obtained without using the local or global feature. The use of both features resulted in 0.74% or 1.52% higher mIoU compared to the single-use of the local or global feature, respectively. Last, because GPS-GLASS obtains pseudo-supervision from both intra-matching and inter-matching, we evaluated the performance without applying intra-matching or inter-matching and obtained 3.33% or 1.98% lower mIoU compared to GPS-GLASS, respectively.

5. Conclusions

In this paper, we proposed GPS-GLASS, a novel training methodology for nighttime semantic segmentation based on unlabeled daytime-nighttime image pairs and their GPS data. GPS-GLASS obtains pixel-level aligned pseudo-supervision through bidirectional correspondence matching between the daytime and nighttime. To address the difficulty of correspondence matching between different domains, GPS-GLASS also acquires another pseudo-supervision through correspondence matching in the same daytime domain using the GPS data. The confidence map is used to exclude pseudo-supervision of less reliable predictions. Our GPS-GLASS does not increase the number of network parameters or inference time compared to the adopted baseline model. Experimental results on the ACDC-night, Dark Zurich-val, and NightCity+ datasets demonstrate the effectiveness of the proposed method.

Acknowlegement

This work is supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A2C2002810).

References

[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017.

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016.

[3] Dengxin Dai and Luc Van Gool. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *ITSC*, pages 3819–3824, 2018.

[4] Xueqing Deng, Peng Wang, Xiaochen Lian, and Shawn Newsam. NightLab: A dual-level architecture with hardness detection for segmentation at night. In *CVPR*, pages 16938–16948, 2022.

[5] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. In *ICCV*, pages 2758–2766, 2015.

[6] Huan Gao, Jichang Guo, Guoli Wang, and Qian Zhang. Cross-domain correlation distillation for unsupervised domain adaptation in nighttime semantic segmentation. In *CVPR*, pages 9913–9923, 2022.

[7] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *Int. J. Robot. Res.*, 2013.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[9] Sina Honari, Pavlo Molchanov, Stephen Tyree, Pascal Vincent, Christopher Pal, and Jan Kautz. Improving landmark localization with semi-supervised learning. In *CVPR*, pages 1546–1555, 2018.

[10] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, pages 2462–2470, 2017.

[11] James Inman. *Navigation and nautical astronomy, for the use of British seamen*. F. & J. Rivington, 1849.

[12] Somi Jeong, Youngjung Kim, Eungbean Lee, and Kwanghoon Sohn. Memory-guided unsupervised image-to-image translation. In *CVPR*, pages 6558–6567, 2021.

[13] Liming Jiang, Changxu Zhang, Mingyang Huang, Chunxiao Liu, Jianping Shi, and Chen Change Loy. TSIT: A simple and versatile framework for image-to-image translation. In *ECCV*, pages 206–222, 2020.

[14] Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. Learning to estimate hidden motions with global motion aggregation. *arXiv preprint arXiv:2104.02409*, 2021.

[15] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *ICCV*, pages 66–75, 2017.

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[17] Junghyup Lee, Dohyung Kim, Wonkyung Lee, Jean Ponce, and Bumsub Ham. Learning semantic correspondence exploiting an object-level prior. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.

[18] Junghyup Lee, Dohyung Kim, Jean Ponce, and Bumsub Ham. SFNet: Learning object-aware semantic correspondence. In *CVPR*, pages 2278–2287, 2019.

[19] S Lee, J Hyun, H Seong, and E Kim. Unsupervised domain adaptation for semantic segmentation by content transfer. In *AAAI*, pages 8306–831, 2021.

[20] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *CVPR*, pages 6936–6945, 2019.

[21] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NIPS*, pages 700–708, 2017.

[22] Ziyang Ma, Kaiming He, Yichen Wei, Jian Sun, and Enhua Wu. Constant time weighted median filtering for stereo matching and beyond. In *ICCV*, pages 49–56, 2013.

[23] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *CVPR*, pages 4161–4170, 2017.

[24] Christoph Rhemann, Asmaa Hosni, Michael Bleyer, Carsten Rother, and Margrit Gelautz. Fast cost-volume filtering for visual correspondence and beyond. In *CVPR*, pages 3017–3024, 2011.

[25] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, pages 102–118. Springer, 2016.

[26] Eduardo Romera, Luis M Bergasa, Kailun Yang, Jose M Alvarez, and Rafael Barea. Bridging the day and night domain gap for semantic segmentation. In *IV*, pages 1312–1318, 2019.

[27] Christos Sakaridis. ACDC website. `https://acdc.vision.ee.ethz.ch/`, Feb. 2022.

[28] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *ICCV*, pages 7374–7383, 2019.

[29] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.

[30] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *ICCV*, pages 10765–10775, 2021.

[31] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *CVPR*, pages 8934–8943, 2018.

[32] Lei Sun, Kaiwei Wang, Kailun Yang, and Kaite Xiang. See clearer at night: towards robust nighttime semantic segmentation through day-night image conversion. In *Proceedings*

of the Artificial Intelligence and Machine Learning in Defense Applications, volume 11169, page 111690A, 2019.

[33] Xin Tan, Ke Xu, Ying Cao, Yiheng Zhang, Lizhuang Ma, and Rynson WH Lau. Night-time scene parsing with a large real dataset. *IEEE Trans. Image Process.*, 2021.

[34] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, pages 402–419, 2020.

[35] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, pages 7472–7481, 2018.

[36] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. Domain adaptation for structured output via discriminative patch representations. In *ICCV*, pages 1456–1465, 2019.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

[38] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, pages 2517–2526, 2019.

[39] Xinyi Wu, Zhenyao Wu, Hao Guo, Lili Ju, and Song Wang. Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation. In *CVPR*, pages 15769–15778, 2021.

[40] Xinyi Wu, Zhenyao Wu, Lili Ju, and Song Wang. A one-stage domain adaptation network with image alignment for unsupervised nighttime semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[41] Haofei Xu, Jiaolong Yang, Jianfei Cai, Juyong Zhang, and Xin Tong. High-resolution optical flow from 1d attention and correlation. In *ICCV*, pages 10498–10507, 2021.

[42] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. GMFlow: Learning optical flow via global matching. *arXiv preprint arXiv:2111.13680*, 2021.

[43] Qi Xu, Yinan Ma, Jing Wu, Chengnian Long, and Xiaolin Huang. CDAda: A curriculum domain adaptation for nighttime semantic segmentation. In *ICCVW*, pages 2962–2971, 2021.

[44] Qingxiong Yang. Stereo matching using tree filtering. *PAMI*, 37(4):834–846, 2015.

[45] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, pages 2636–2645, 2020.

[46] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *CVPR*, pages 5143–5153, 2020.

[47] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017.

[48] Huabing Zhou, Jiayi Ma, Chiu C Tan, Yanduo Zhang, and Haibin Ling. Cross-weather image alignment via latent generative model with intensity consistency. *IEEE Trans. Image Process.*, 2020.