DepthSegNet24: A Label-Free Model for Robust Day-Night Depth and Semantics

Phan Thi Huyen Thanh[1⋆[0009−0007−1781−4827]], The Hiep Nguyen[2,3∗[0009−0007−5765−0468]], Minh Huy Vu Nguyen[2,3[0009−0003−4599−7182]], Trung Thai Tran[2,3[0009−0002−1422−9685]], Tran Vu Pham[2,3], Duc Dung Nguyen[2,3[0000−0001−7321−7401]], Truong Vinh Truong Duy[1[0000−0001−8000−2214]], and Natori Naotake[1[0000−0002−6128−4145]]

[1] Tokyo Research Center, Aisin Corporation, Japan
[2] AITech Lab., Ho Chi Minh City University of Technology (HCMUT)
[3] Vietnam National University Ho Chi Minh City (VNUHCM)
{thanh.phan,duy.truong,naotake.natori}@aisin.co.jp
{hiep.nguyena113872,huy.vu.cse.9,thai.tran241002,ptvu,nddung}@hcmut.edu.vn

Abstract. This paper presents a novel multi-task model combining self-supervised monocular depth estimation and knowledge-distilled semantic segmentation that can perform both tasks simultaneously and consistently in both daytime and nighttime conditions. By leveraging the joint self-supervised and supervised knowledge distillation learning, the model can learn consistent and complementary representations of the two tasks to improve the generalization ability without relying on annotated ground-truth data. To address the extremely varying lighting conditions between day and night, we first synthesize night and day images from their corresponding real day and night images, and then train the model with the day-night image pairs to provide explicit correspondences between the two lighting conditions for capturing the contextual and detailed information in both scenarios. We also augment the model with a light enhancement module and a daytime depth pseudo-labels for achieving more accurate and robust depth and segmentation. Experimental results on Oxford RobotCar and nuScenes demonstrate the robustness of our model in diverse challenging lighting conditions.

Keywords: Monocular depth estimation · Semantic segmentation · Multi-task learning · Day-night/All-day prediction · Label-free approach

1 Introduction

Monocular depth estimation (MDE) and semantic segmentation (SS) are two fundamental tasks in computer vision that have a wide range of applications in numerous fields, most notably autonomous driving and navigation, 3D reconstruction, augmented reality, object recognition, and robotics. MDE aims to predict the depth of a scene from a single image, while SS focuses on assigning

⋆ Equal contributors

semantic labels to each pixel in the image. These tasks have been addressed independently, with dedicated models designed for each task.

Several recent studies have explored the combination of MDE and SS tasks into a multi-task learning model [1,24,29,30,40]. However, existing works focused primarily on daytime scenarios and did not specifically address the challenges of robustness in both daytime and nighttime scenarios. Real-world deployments require such models to operate reliably in an all-day setting, where extreme variations in lighting conditions, shadows, reflections, and other factors between daytime and nighttime can significantly impact the visual information and performance. Although a multi-task model of MDE and SS that can perform consistently and accurately under different conditions is desirable, to the best of our knowledge, there have been few to no reports on a model of this nature.

In this paper, we propose a novel multi-task model of MDE and SS that is robust in both daytime and nighttime combining self-supervised depth supervision and SS knowledge distillation. By leveraging self-supervised depth supervision, our model can learn depth in a self-supervised manner from consecutive frames only, eliminating the reliance on labeled depth data. Also, by incorporating SS knowledge distillation, our model can benefit from the rich semantic information captured by state-of-the-art pre-trained SS models. Another key feature is the synthesizing and utilization of real and synthetic day-night pairs of images to capture contextual and detailed information in both scenarios of day and night. This combination allows our model to achieve accurate MDE and SS simultaneously while being robust in diverse challenging lighting conditions.

The rest of the paper is organized as follows. Section 2 provides a review of related work in MDE, SS, multi-task learning, domain adaptation and knowledge distillation. The principles and architecture of our proposed model are presented in Section 3. Section 4 describes the experimental setup and presents the results and analysis. Finally, the paper is concluded in Section 5 with a summary of our contributions and future outlook.

2 Related Works

Monocular Depth Estimation (MDE). Supervised deep learning methods for MDE based on direct supervision and fine-grained control over the learning process are still the best performers in the field [2, 13, 27, 36], although they require time-consuming and costly pixel-level manual depth annotations. On the other hand, relying on only consecutive frames, the attractive self-supervised approach has been shown to close in on the performance gap with the supervised counterpart [17,21,46], and recently been extended to all-day MDE for handling different lighting conditions in daytime and nighttime [16,28,35]. Our multi-task model extends the self-supervised MDE to cover SS in a multi-task model for both domains of day and night.

Semantic Segmentation (SS). While there have been a lot of efforts in weakly supervised and unsupervised approaches leveraging techniques such as cluster-

ing, self-training, and domain adaptation [5, 9, 26, 48], the supervised SS trained using labeled data, where each pixel in the image is annotated with its corresponding class label, is still the standard approach [6, 7, 14]. Recently all-day SS methods have been proposed to segment objects and regions in images of both daytime and nighttime [3, 10, 44], which we exploit as a teacher model for enabling our label-free approach.

Multi-task Model of MDE and SS. Multi-task models that combine MDE and SS aim to jointly predict both depth maps and pixel-wise semantic labels from a single input image [1, 24, 29, 30, 40], with recent works adding more downstream tasks such as optical flow [19, 20, 42]. They showcase the benefits of jointly training these tasks, such as improved performance and reduced computational cost by taking advantage of the shared representation. In this work, we carefully design the network and loss functions to handle task interference and increased complexity of our model to unlock its generalization ability in both day and night.

Domain Adaptation. Domain adaptation plays a crucial role in bridging the gap between clear daytime and challenging nighttime. Recent advances have focused on leveraging models trained on well-labeled or clear conditions like daytime to adapt effectively in the more complex nighttime domain across various tasks, such as depth estimation [16, 28, 35], SS [41], object detection [12], etc. Here we synthesize the night domain from the real day domain, and the day domain from the real night domain to train the model with the synthetic-real pairs of day and night images for learning the two contrasting lighting conditions.

Knowledge Distillation. Knowledge distillation involves transferring knowledge from a complex teacher model to a simpler student model [18, 23, 25]. This process aims to improve the performance, efficiency, and transferability of the student model by leveraging the knowledge learned by the teacher model. Our model incorporates knowledge-distilled SS and day depth losses for elimination of annotations and performance enhancement.

3 Method

3.1 Principles of our Proposed Method

Our method focuses on a multi-task model of MDE and SS aimed at delivering consistent performance in extremely varying environments in both day and night. The model is designed to capture both shared common representations between the two cross-tasks of MDE and SS and representations between the two lighting conditions of day and night. The model is also trained in a label-free fashion to eliminate the need for expensive and time-consuming annotations. Our overall approach compared with traditional approaches is depicted in Figure 1.

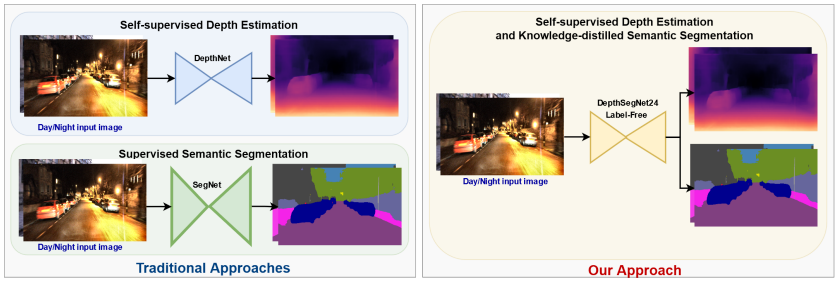


Fig. 1: Overview of our approach. While traditional approaches address either multi-task learning or multi-domain challenges individually, our approach integrates both aspects, i.e. multi-task learning of depth and semantic segmentation in two domains of day and night, providing a holistic solution to complex visual processing tasks.

Multi-task Model of Depth and Semantic Segmentation. Our method combines MDE and SS into a single multi-task learning framework. By employing a shared common encoder and separate decoders for depth and segmentation, the model can leverage the shared features to learn representations that are consistent between the two tasks. This allows the model to take advantage of complementary information from both tasks, ultimately improving their performance. This multi-task approach enhances the overall efficiency and accuracy of the model.

Robustness in Both Daytime and Nighttime. To ensure robustness across varying lighting conditions, we first synthesize night and day images from the corresponding real-world day and night counterparts. We then utilize these day-night image pairs, i.e. real-day-synthetic-night pairs and synthetic-day-real-night pairs, during training. By providing the model with explicit correspondences between the two lighting conditions using these day-night pairs, the model can learn and capture contextual and detailed information in both the day and night scenarios, where lighting can vary significantly.

Label-Free: Combination of Self-Supervised MDE and Knowledge-Distilled SS. Our method employs a label-free training approach by combining self-supervised MDE with knowledge-distilled SS. The self-supervised MDE does not require ground-truth depth labels, while the knowledge-distilled SS allows the model to leverage the knowledge learned by a teacher model. This approach enables the model to benefit from both unsupervised and supervised learning, which boosts its generalization ability in unseen and different scenarios. Additionally, it reduces the dependency on annotations, thereby enhancing the scalability and applicability of the model.

3.2 Model

The proposed model architecture, as illustrated in Figure 2, includes a light enhancement module, a shared encoder, two separate decoders for MDE and SS, as well as a PoseNet for predicting the relative poses of consecutive frames used in the self-supervised MDE approach.

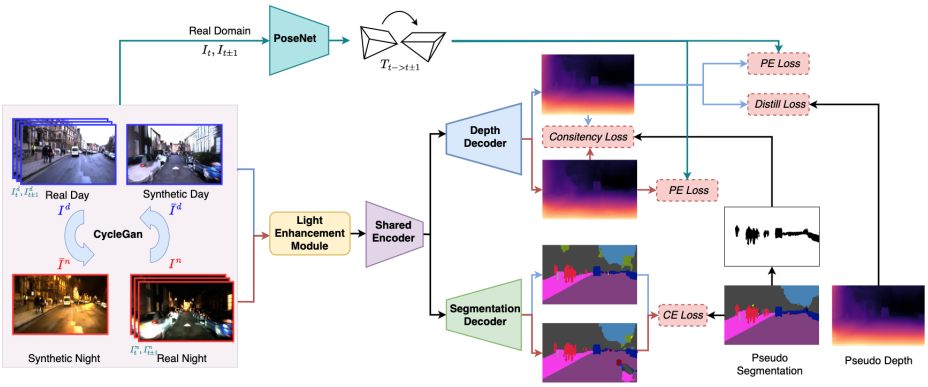


Fig. 2: Our model involves feeding pairs of real-day-synthetic-night and real-night-synthetic-day images into the light enhancement module. These are then passed through the shared encoder and into two separate decoders for depth estimation and semantic segmentation. The real day and night consecutive images are passed through the PoseNet to predict their relative poses, which are then used in training for image reconstruction and the self-supervised depth loss.

Light Enhancement Module. Inspired by [41], the light enhancement module is the initial stage of our model, designed to improve the visibility of nighttime images. This module adjusts the brightness and contrast of nighttime images to approximate daytime conditions, facilitating more accurate feature extraction in subsequent stages. Additionally, the light enhancement module includes a light loss component to optimize the light enhancement process.

Shared Encoder and Separate Decoders. The light-enhanced images are then passed through a shared encoder. This encoder extracts high-level features that are common to both the depth estimation and SS tasks, as well as to both daytime and nighttime conditions. The encoded features are then fed into two separate decoders: the depth decoder and the segmentation decoder. The depth decoder generates depth maps, ensuring that the model understands the spatial structure of the input images, while the segmentation decoder produces segmentation maps for identifying and classifying different objects within the images. This architecture allows our model to benefit from multi-task learning, where

improvements in one task can positively affect the other, ultimately leading to a more comprehensive understanding of the scenes.

Real-Day-Synthetic-Night Pairs and Synthetic-Day-Real-Night Pairs. For the real-day-synthetic-night pairs, the real day images I^d are converted to synthetic night images $\bar{I}^n$ using CycleGan [47]. Conversely, the real night images I^n are transformed to synthetic day images $\bar{I}^d$ for the synthetic-day-real-night pairs. This helps the model learn how to handle images captured under real low-light conditions, rather than relying solely on the synthetic night images. The pairs are then processed by the light enhancement module, which adjusts the visibility of both the real and synthetic night images before feature extraction.

3.3 Losses and Training

Light Enhancement Loss. Based on [41], the light enhancement loss is a combination of three different loss functions:

- The variation loss $\mathcal{L}_{\text{tv}}$, widely used in many previous works [34, 39, 43], to smooth the images,
- The exposure control loss $\mathcal{L}_{\text{exp}}$ to force the lighting effects in the day and night scenarios to be as consistent as possible, and
- The structural similarity loss $\mathcal{L}_{\text{ssim}}$ to ensure that the generated lighting maps are consistent with the original images.

Unlike [41] which uses the same light enhancement loss $\mathcal{L}_{\text{light}}$ for both daytime and nighttime images, thereby altering the intensity distributions of both, we only aim to adjust the light intensity distributions of the nighttime images to closely match those of the daytime images without dramatically changing the daytime images. Therefore, we remove the loss term $\mathcal{L}_{\text{exp}}$ for the daytime images and apply it to the nighttime ones exclusively. The light enhancement loss $\mathcal{L}_{\text{light}}$ in our model is defined as

$$\mathcal{L}_{\text{light-day}} = \alpha_{\text{tv}}\mathcal{L}_{\text{tv}} + \alpha_{\text{ssim}}\mathcal{L}_{\text{ssim}}, \tag{1}$$

$$\mathcal{L}_{\text{light-night}} = \alpha_{\text{tv}}\mathcal{L}_{\text{tv}} + \alpha_{\text{exp}}\mathcal{L}_{\text{exp}} + \alpha_{\text{ssim}}\mathcal{L}_{\text{ssim}}, \tag{2}$$

$$\mathcal{L}_{\text{light}} = \mathcal{L}_{\text{light-day}} + \mathcal{L}_{\text{light-night}}, \tag{3}$$

where α_{tv}, α_{exp}, and α_{ssim} are set to 10, 1, and 1.

Self-supervised Depth Loss. We adopt a self-supervised learning approach [17, 46] to reconstruct the target image $I_{s\to t}$ from the source image $I_{s\in\{t-1,t+1\}}$ and the depth D_t of the target image I_t through the following equation.

$$I_{s\to t} = I_s(\text{Proj}(D_t, T_{t\to s}, K)), \tag{4}$$

where K is the camera intrinsic, and $T_{t\to s}$ is the transformation matrix representing the camera transition from the target image to the source image estimated by the PoseNet as

$$T_{t\to s} = \text{PoseNet}(I_s, I_t). \tag{5}$$

To evaluate the quality of the reconstructed image, we utilize the pixel-wise photometric error defined as

$$\mathcal{L}_{\text{pe}}(I_t, I_{s\to t}) = \frac{\alpha}{2}(1 - \text{SSIM}(I_t, I_{s\to t})) + (1-\alpha)\|I_s - I_{s\to t}\|_1. \tag{6}$$

We also add a smoothness function to impose constraints on the depth map and facilitate the blurring of regions with similar pixels.

$$\mathcal{L}_{\text{smooth}}(D_t, I_t) = |\partial_x D_t| e^{-|\partial_x I_t|} + |\partial_y D_t| e^{-|\partial_y I_t|}. \tag{7}$$

The terms $\mathcal{L}_{\text{pe}}$ and $\mathcal{L}_{\text{smooth}}$ form the real-domain self-supervised loss function as follows.

$$\mathcal{L}_{\text{sf-real}}(I_t, I_{s\to t}, D_t) = \mathcal{L}_{\text{pe}}(I_t, I_{s\to t}) + 0.001\mathcal{L}_{\text{smooth}}(D_t, I_t). \tag{8}$$

The term $\mathcal{L}_{\text{sf-real}}$ performs reliably in the real domain where the lighting across consecutive frames is consistent. However, in the synthetic domain, it suffers from the problem that the reconstructed image is very dissimilar from the target image $\bar{I}_t$ as there may be changes between frames that can not be accounted for. Following previous works [16, 35], we improve depth prediction from the synthetic domain $\bar{D}_t$ by using this depth map to warp the image in the real domain correspondingly as follows.

$$\bar{I}_{s\to t} = I_s(\text{Proj}(\bar{D}_t, T_{t\to s}, K)), \tag{9}$$

where the depth $\bar{D}_t$ is predicted from the synthetic image $\bar{I}_t$, and the transformation matrix $T_{t\to s}$ is predicted through the corresponding consecutive frames I_s, I_t in the real domain. The self-supervised loss for the synthetic domain is defined as

$$\mathcal{L}_{\text{sf-syn}}(I_t, \bar{I}_{s\to t}, \bar{D}_t) = \mathcal{L}_{\text{pe}}(I_t, \bar{I}_{s\to t}) + 0.001\mathcal{L}_{\text{smooth}}(\bar{D}_t, I_t). \tag{10}$$

Finally, our self-supervised depth loss is a combination of three losses comprised of real day, synthetic night, and real night defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{sf}} = &\mathcal{L}_{\text{sf-real}}(I_t^d, I_{s\to t}^d, D_t^d) + \\ &\mathcal{L}_{\text{sf-syn}}(I_t^d, \bar{I}_{s\to t}^n, \bar{D}_t^n) + \\ &0.1\mathcal{L}_{\text{sf-real}}(I_t^n, I_{s\to t}^n, D_t^n).\end{aligned} \tag{11}$$

In Equation (11), we incorporate the loss term $\mathcal{L}_{\text{sf-real}}(I_t^n, I_{s\to t}^n, D_t^n)$ to enhance the robustness of our model's depth predictions in nighttime scenes. However, the effectiveness of $\mathcal{L}_{\text{sf-real}}$ may be adversely affected by significant variations in lighting conditions across consecutive frames in the real night domain. To mitigate this issue, we adjust the weight of this loss term in practice, reducing it to 0.1 to ensure more stable training outcomes under variable lighting conditions.

Consistency Loss Between Real and Synthetic Domains. To achieve uniform depth across both day and night images, our model incorporates a consistency loss $\mathcal{L}_{\text{cst}}$ between the depth maps generated from daytime images and those derived from their corresponding nighttime image counterparts. While the transition from real night to synthetic day enhances model training by providing real low-light conditions, it also introduces challenges in calculating consistency loss. Synthetic daytime images can still exhibit imperfections, such as overexposed areas or inaccurate color reproduction, particularly in the 'sky' region. These imperfections can lead to discrepancies in depth estimation, where those areas are misrepresented.

To address the discrepancies in depth estimation, particularly in the 'sky' region of synthetic daytime images, we implement a masking strategy before calculating the consistency loss. By excluding the 'sky' class from the loss computation, we mitigate the impact of these errors on the overall depth estimation accuracy. This masking approach refines the consistency loss calculation by concentrating on more reliably reproduced areas. This ensures that our depth estimation remains robust and consistent, minimizing the influence of synthetic artifacts on the model's performance. As a result, our consistency loss is defined as follows.

$$\mathcal{L}_{\text{cst-real}} = \frac{1}{N}\sum_{i=1}^{N}\left\|D^{d,i} - \bar{D}^{n,i}\right\|^2, \tag{12}$$

$$\mathcal{L}_{\text{cst-syn}} = \frac{1}{N}\sum_{i=1}^{N}\left\|M_i \cdot (\bar{D}^{d,i} - D^{n,i})\right\|^2, \tag{13}$$

$$\mathcal{L}_{\text{cst}} = \mathcal{L}_{\text{cst-real}} + \mathcal{L}_{\text{cst-syn}}, \tag{14}$$

where M_i is the mask generated from the pseudo-segmentation map, and each pair $D^d - \bar{D}^n$, $\bar{D}^d - D^n$ is the depth prediction from each real-day-synthetic-night and synthetic-day-real-night pair respectively. Since depth predictions in the daytime domain (D^d, $\bar{D}^d$) are typically more accurate than those in the nighttime domain, we exploit them as pseudo-labels to align the depth predictions of the nighttime domain.

Day Depth Distillation Loss. To address the problem of detrimental effects on daytime domain performance when training models across daytime and nighttime domains, we make use of pseudo-labels generated from a daytime MDE model that is exclusively trained on daytime images. These daytime pseudo-labels are aimed at refining the depth estimation network to align more precisely with the unique features of the daytime domain.

The day depth distillation loss $\mathcal{L}_{\text{ds}}$ based on the daytime pseudo-labels $\hat{D}^d$ is as follows.

$$\mathcal{L}_{\text{ds}} = \frac{1}{N}\sum_{i=1}^{N}\left\|D^{d,i} - \hat{D}^{d,i}\right\|^2 + (1 - SSIM(D^d, \hat{D}^d)). \tag{15}$$

Knowledge-Distilled SS Loss. To train the segmentation network, we leverage pseudo-labels generated by an off-the-shelf SS model, rather than relying on ground-truth pixel-level segmentation masks. However, pseudo-labels also inherently carry the risk of incorporating errors from the off-the-shelf model. This risk is particularly noticeable under challenging lighting conditions such as nighttime scenarios. In order to deal with this problem, we incorporate a confidence thresholding mechanism for selectively including only pixels whose pseudo-labels have the predicted probability, i.e. confidence level, above a specific threshold in the cross-entropy loss as below.

$$\mathcal{L}_{\mathrm{CE}}(y, p) = \sum_{i=1}^{N} \sum_{c=1}^{C} \mathbb{1}_{(y_{ic} > \tau)} \cdot y_{ic} \cdot \log(p_{ic}), \tag{16}$$

where N is the total number of pixels, C is the number of semantic classes, p_{ic} is the prediction for pixel i belong to class c, y_{ic} is the generated pseudo-label, and τ is the threshold set to 0.1 in our experiments. Our segmentation loss is then defined as

$$\mathcal{L}_{\mathrm{seg}} = \mathcal{L}_{\mathrm{CE}}(y^d, p^d) + \mathcal{L}_{\mathrm{CE}}(y^d, \bar{p}^n) + \mathcal{L}_{\mathrm{CE}}(y^n, p^n), \tag{17}$$

where y^d, and y^n are the SS pseudo-labels for real day and real night respectively, p^d, $\bar{p}^n$ are our model's predictions for real-day-synthetic-night pairs. Similar to the self-supervised depth loss, we also expose our model to real night SS conditions by prediction on real night domain p^n.

Total Loss. Our total loss is defined as

$$\mathcal{L} = \alpha_{\mathrm{light}}\mathcal{L}_{\mathrm{light}} + \alpha_{\mathrm{sf}}\mathcal{L}_{\mathrm{sf}} + \alpha_{\mathrm{cst}}\mathcal{L}_{\mathrm{cst}} + \alpha_{\mathrm{ds}}\mathcal{L}_{\mathrm{ds}} + \alpha_{\mathrm{seg}}\mathcal{L}_{\mathrm{seg}}, \tag{18}$$

where α_{light}, α_{sf}, α_{cst}, α_{ds}, and α_{set} are set to 0.01, 1.0, 1.0, 1.0, and 1.0.

4 Experiments

4.1 Datasets

In line with previous practice [15,16,28,37,45], we use large-scale Oxford RobotCar [32] and NuScenes [4] driving datasets in this paper (see Appendix for details).

4.2 Implementations

Our method is implemented in the PyTorch [33] framework. We follow the network architecture of [17] with a standard shared ResNet-18 [22] encoder and two decoders for depth and SS with skip connections. The Adam optimizer with an initial learning rate of $2.5e^{-4}$ is used to train our model through 40 epochs with a batch size of 4 and with each epoch taking approximately 0.5 hours on a GeForce RTX 4090 24 GB GPU. We utilize the popular Mask2Former [8], in particular, the model trained with Cityscapes [11], and Monodepth2 [17] as the SS and daytime depth teacher models, respectively.

Table 1: Depth on Oxford RobotCar (depth = 40 m). Best: **bold**, second best: underlined. Our multi-task model achieves the best or second-best performance compared to depth-specifically-optimized dedicated SOTA methods.

Night - Oxford RobotCar Method	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	$\delta < 1.25$↑	$\delta < 1.25^2$↑	$\delta < 1.25^3$↑
Monodepth2 [17] (day)	0.477	5.389	9.163	0.466	0.351	0.635	0.826
Monodepth2 [17] (night)	0.661	25.213	12.187	0.553	0.551	0.849	0.914
HR-Depth [31]	0.512	5.800	8.726	0.484	0.388	0.666	0.827
ADDS-DepthNet [28]	0.233	2.344	6.859	0.270	0.631	0.908	0.962
Ours	**0.195**	**1.714**	**5.936**	**0.240**	**0.734**	**0.917**	**0.970**
Day - Oxford RobotCar Method	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	$\delta < 1.25$↑	$\delta < 1.25^2$↑	$\delta < 1.25^3$↑
Monodepth2 [17] (day)	0.117	0.673	3.747	0.161	0.867	0.973	0.991
Monodepth2 [17] (night)	0.306	2.313	5.468	0.325	0.545	0.842	0.937
HR-Depth [31]	0.121	0.732	3.947	0.166	0.848	0.970	0.991
ADDS-DepthNet [28]	**0.109**	**0.584**	3.578	**0.153**	**0.880**	**0.976**	**0.992**
Ours	0.113	0.604	**3.490**	0.161	0.865	0.974	0.991

4.3 Results and Discussions

Depth on Oxford RobotCar. Table 1 shows evaluation scores for different depth estimation models on both nighttime and daytime scenes of this Oxford RobotCar dataset. It is worth mentioning that the Monodepth2, HR-Depth [31], and ADDS-DepthNet [28] models are single-task models designed and optimized for this specific MDE task. In the nighttime scene up to 40 m, Monodepth2 trained with the daytime and nighttime splits achieved an absolute relative score of 0.477 and 0.661, respectively. HR-Depth achieved a score of 0.512, ADDS-DepthNet achieved a score of 0.233, and our model achieved a score of 0.195. In the daytime scene up to 40 m, Monodepth2 achieved a score of 0.117 and 0.306 respectively for the trained daytime and nighttime splits. HR-Depth achieved a score of 0.121, ADDS-DepthNet achieved a score of 0.109, and our model achieved a score of 0.113. The results also emphasize the increased difficulty of the nighttime scene compared to the daytime scene. Overall, our model performed relatively well in both nighttime and daytime scenes, with the lowest score achieved for the nighttime scene and on par with HR-Depth for the daytime scene. It is important to highlight that our model, unlike the other methods, is not a single-task model tailored solely for this task. This suggests that our model may have a more robust performance across different scenarios.

Depth on nuScenes. Table 2 provides a comparison of the models' performance in estimating depth in different lighting conditions on two scenarios: "day-clear" and "night" in the nuScenes dataset. Monodepth2 consistently achieved relatively good accuracy in both scenarios. RNW [38] and STEPS [45] performed slightly lower than Monodepth2. PackNet-SfM showed better performance than Monodepth2, RNW, and STEPS. md4all achieved the lowest scores, partially thanks to the additional velocity supervision. To provide further insights into multi-task models, we integrated a semantic encoder into the Monodepth2 base-

Table 2: Depth on nuScenes (depth = 80 m). Best: **bold**, second best: underlined, M: monocular self-supervised, v: velocity supervision on PoseNet based on odometry. Our multi-task model performs relatively well in both night and day.

Method	Train	day-clear - nuScenes			night - nuScenes		
		Abs Rel↓	RMSE↓	$\delta < 1.25$↑	Abs Rel↓	RMSE↓	$\delta < 1.25$↑
Monodepth2 [17]	M	**0.137**	6.692	0.850	0.283	9.729	0.518
Monodepth2-naive	M	0.215	9.256	0.660	0.256	11.018	0.565
RNW [38]	M	0.287	9.185	0.562	0.333	10.098	0.437
PackNet-SfM [21]	M	0.157	7.230	0.826	0.262	11.063	0.566
STEPS [45]	M	0.258	9.864	**0.858**	0.287	9.120	0.572
md4all [16]	M+v	**0.137**	**6.452**	0.846	**0.192**	**8.507**	**0.710**
Ours	M	0.151	6.712	0.816	0.211	8.899	0.661

line to construct a basic multi-task model for depth estimation and SS. The results from this multi-task Monodepth2-naive model indicate that the naive multi-task approach decreased the performance of Monodepth2 itself in both day and night conditions, compared to the comprehensive mechanisms for multi-task learning employed in our model. Our model performed relatively well, with scores similar to PackNet-SfM and Monodepth2 in both scenarios. Nevertheless, the models' performance provides insights into their relative strengths and weaknesses in different lighting conditions. In general, compared with Oxford RobotCar, nuScenes is much more challenging for all the models. Once again, it is worth mentioning that except for our model and Monodepth2-naive, all other models are single-task models customized for this specific task.

Semantic Segmentation on nuScenes and Oxford RobotCar. Table 3 provides the mean Intersection over Union (mIoU) and the accuracy of various semantic classes of two models, the optimized SS teacher model Mask2Former and our multi-task depth and SS student model, on day and night scenes of nuScenes and Oxford RobotCar.

For the nuScenes dataset, Mask2Former achieved mIoU scores of 68.31 and 39.71 during the day and night, respectively. In terms of individual class accuracy, Mask2Former performed well in road, vegetation, car, and truck classes. Meanwhile, we achieved lower mIoU scores of 57.74 during the day and 35.05 in nighttime, performing relatively well in road, vegetation, car and terrain classes.

As for the Oxford RobotCar dataset, we recorded mIoU scores of 64.31 during the day and 49.73 during the night, and delivered well in road, car, building, and sidewalk classes. Due to the unavailability of ground-truth labels, the results were calculated using Mask2former results as the ground-truth.

Overall, Mask2Former generally achieved higher accuracy in both datasets as expected for a teacher model. Our method showed competitive performance in various semantic classes, especially in the nuScenes dataset. However, it is important to consider the limitations of using Mask2Former results as the ground-truth for the Oxford RobotCar dataset. The multi-task nature of our method, combining depth and SS, suggests its potential for more comprehensive understanding of the scenes.

Table 3: Semantic segmentation. Our multi-task model delivers competitive performance in various semantic classes against the semantic-segmentation-dedicated teacher model.

(a) nuScenes (based on ground-truth labels).

Method	Split	mIoU	road	sidewalk	fence	vegetation	terrain	person	car	truck
Mask2Former [8]	*Day*	**68.31**	95.97	58.52	28.41	86.59	62.34	60.76	77.29	76.62
Ours	*Day*	57.74	93.15	41.02	25.09	82.45	57.44	40.47	68.61	54.18
Mask2Former [8]	*Night*	**39.71**	87.14	32.14	24.68	67.34	27.62	7.42	58.36	12.94
Ours	*Night*	35.05	86.37	37.24	5.93	67.94	12.93	1.91	61.23	6.81

(b) Oxford RobotCar (based on pseudo-labels by Mask2Former).

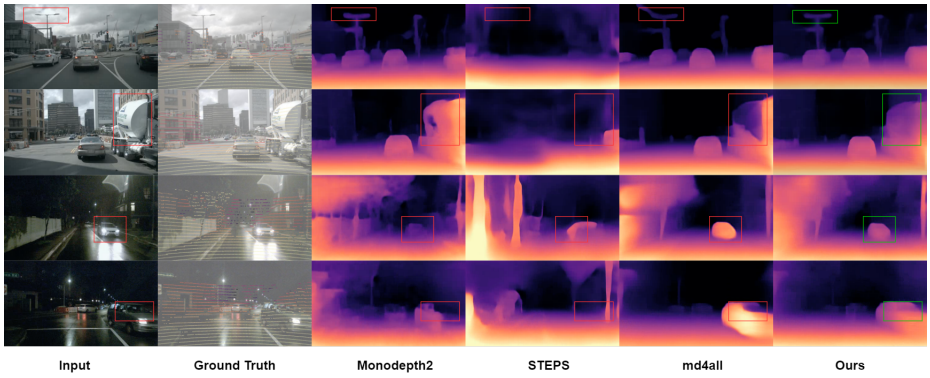
Method	Split	mIoU	road	sidewalk	building	wall	fence	pole	vegetation	sky	car
Ours	*Day*	64.31	97.52	70.12	88.36	24.05	8.35	37.70	70.19	92.99	89.5
Ours	*Night*	49.73	91.36	49.93	73.54	43.85	11.3	13.88	62.19	19.67	81.86

Qualitative Results. Figure 3 visualizes depth and SS predictions for some representative day and night scenes on nuScenes and Oxford RobotCar. The depth maps generated by our model show clear distinctions and relative distances between objects such as cars and traffic lights even at night, while the segmentation maps also correctly identify and classify the objects and boundaries of the cars and road with appropriate semantic labels. The results highlight the ability of our multi-task model in seamless integration of MDE and SS for producing visually consistent and coherent perception.

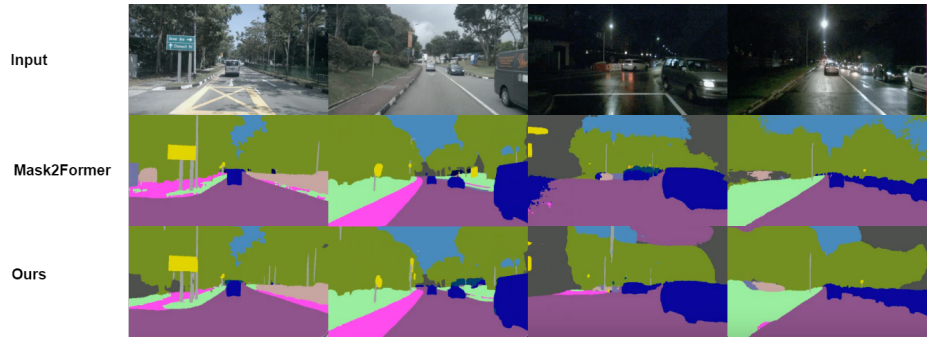
Computational Efficiency. By comparing the GMACs, number of parameters, and inference time of these models (Table 4), it is evident that our multi-task model achieves a 10x higher computational efficiency with fewer computations and parameters compared to a naive combination of 2 single-task models while maintaining a similar level of performance in terms of depth estimation and SS. This highlights the advantage of combining multiple tasks into a single model, resulting in improved computational efficiency without sacrificing performance.

4.4 Ablation Study

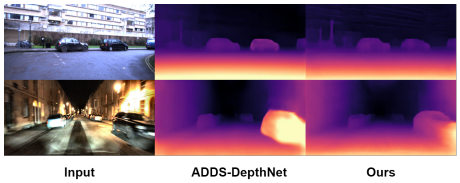
We analyze the performance of our method by examining different variants with the core components of the light enhancement module, the day depth distillation, and the SS distillation of the real and synthetic domains on nuScenes. As



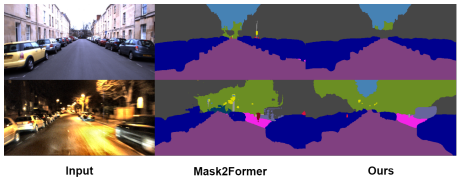
(a) Depth on nuScenes.



(b) Semantic segmentation on nuScenes.



(c) Depth on Oxford RobotCar.



(d) Semantic segmentation on RobotCar.

Fig. 3: Qualitative results on nuScenes and Oxford RobotCar. Our results display clear depth and correct classification of the objects and their sharp boundaries in both day and night scenes.

shown in Table 5, the light enhancement module helped lead to an improvement in both depth and SS in both day and night, with a few exceptions of performance drop in RMSE and δ in the nighttime depth, highly likely due to synthetic artifacts. Similarly, the day depth distillation proved its substantial role in boosting the model's overall performance in all the metrics, again except for night depth's RMSE. Domain selection in the SS distillation confirmed that the combination of 3 domains of real day, synthetic night, and real night offered better representations for more stable performance.

Table 4: Computational efficiency on a GeForce RTX 4090 24 GB GPU. A ten-fold computational speedup can be achieved with our multi-task model compared to a naive combination of two single-task models.

	GMACs					Number of parameters (M)					Time (s)
Model	Light Enhance Module	Encoder	Depth Decoder	Seg Decoder	Total	Light Enhance Module	Encoder	Depth Decoder	Seg Decoder	Total	
Ours (depth+semantics)	20.573	6.699	5.358	6.254	38.884	0.167	11.177	3.153	3.192	17.688	0.0020
Monodepth2 [17] (depth)	-	6.699	5.358	-	12.057	-	11.177	3.153	-	14.330	0.0016
Mask2Former [8] (semantics)	-	-	-	-	152.639	-	-	-	-	209.660	0.0267

Table 5: Ablation study on nuScenes. LE: light enhancement, PD: day depth distillation, SN: synthetic night, RN: real night (real day images always included).

Variant				day-clear - nuScenes				night - nuScenes			
RN	SN	LE	PD	mIoU	Abs Rel↓	RMSE↓	$\delta < 1.25$↑	mIoU	Abs Rel↓	RMSE↓	$\delta < 1.25$↑
	✓	✓	✓	<u>56.26</u>	0.157	6.897	0.811	34.95	**0.210**	8.803	**0.668**
✓		✓	✓	55.39	<u>0.155</u>	6.761	0.811	**35.43**	<u>0.211</u>	**8.747**	<u>0.666</u>
✓	✓		✓	54.14	0.156	<u>6.760</u>	<u>0.812</u>	34.42	0.214	8.794	0.655
✓	✓	✓		56.13	0.166	7.115	0.803	33.87	0.214	<u>8.765</u>	0.651
✓	✓	✓	✓	**57.74**	**0.151**	**6.712**	**0.816**	<u>35.05</u>	<u>0.211</u>	8.899	0.661

4.5 Limitations

The use of synthetic day or night images instead of all-real images in the day-night input image pairs in training poses certain limitations for our model, as synthetic images may not accurately capture the complexities and variations present in real-world day and night scenes. Relying on a daytime depth teacher model is also another limitation. Additionally, using a simple ResNet-18-based network architecture in line with previous methods for ease of comparison may result in limited capability compared to complex networks with attention and adaptive convolutions.

5 Conclusion

In this work, we presented DepthSegNet24, a unified approach for robust around-the-clock MDE and SS. Our model delivers consistent performance in diverse lighting conditions by leveraging self-supervised learning, knowledge distillation, synthetic and real day-night image pairs, and light enhancement. This work helps contribute to applications such as autonomous driving and robotics, where a comprehensive understanding of the scene is crucial. We are now improving the model's architecture and investigating additional datasets and scenarios to enhance the model's generalization ability.

Acknowledgments. We acknowledge Ho Chi Minh City University of Technology (HCMUT), VNU-HCM for supporting this study.

References

1. Bansal, N., Ji, P., Yuan, J., Xu, Y.: Semantics-depth-symbiosis: Deeply coupled semi-supervised learning of semantics and depth. 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) pp. 5817–5828 (2022), https://api.semanticscholar.org/CorpusID:249889717
2. Bhat, S.F., Alhashim, I., Wonka, P.: Adabins: Depth estimation using adaptive bins. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 4009–4018. Computer Vision Foundation / IEEE (2021)
3. Bi, Q., You, S., Gevers, T.: Interactive learning of intrinsic and extrinsic properties for all-day semantic segmentation. IEEE Transactions on Image Processing **32**, 3821–3835 (2023), https://api.semanticscholar.org/CorpusID:259369199
4. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. pp. 11618–11628. Computer Vision Foundation / IEEE (2020)
5. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: European Conference on Computer Vision (2018), https://api.semanticscholar.org/CorpusID:263891125
6. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: European Conference on Computer Vision (2018), https://api.semanticscholar.org/CorpusID:3638670
7. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1280–1289 (2021), https://api.semanticscholar.org/CorpusID:244799297
8. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
9. Cho, J.H., Mall, U., Bala, K., Hariharan, B.: Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 16789–16799 (2021), https://api.semanticscholar.org/CorpusID:232427835
10. Choi, S., Jung, S., Yun, H., Kim, J.T., Kim, S., Choo, J.: Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 11575–11585 (2021), https://api.semanticscholar.org/CorpusID:232404762
11. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3213–3223 (2016), https://api.semanticscholar.org/CorpusID:502946
12. Du, Z., Shi, M., Deng, J.: Boosting object detection with zero-shot day-night domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12666–12676 (2024)

13. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada. pp. 2366–2374 (2014)
14. Erişen, S.: Sernet-former: Semantic segmentation by efficient residual network with attention-boosting gates and attention-fusion networks. ArXiv **abs/2401.15741** (2024), https://api.semanticscholar.org/CorpusID:267312119
15. Gasperini, S., Koch, P.N., Dallabetta, V., Navab, N., Busam, B., Tombari, F.: R4dyn: Exploring radar for self-supervised monocular depth estimation of dynamic scenes. 2021 International Conference on 3D Vision (3DV) pp. 751–760 (2021), https://api.semanticscholar.org/CorpusID:236965838
16. Gasperini, S., Morbitzer, N., Jung, H., Navab, N., Tombari, F.: Robust monocular depth estimation under challenging conditions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8177–8186 (2023)
17. Godard, C., Aodha, O.M., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. pp. 3827–3837. IEEE (2019). https://doi.org/10.1109/ICCV.2019.00393
18. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. International Journal of Computer Vision **129**, 1789 – 1819 (2020), https://api.semanticscholar.org/CorpusID:219559263
19. Guizilini, V., Ambruş, R., Chen, D., Zakharov, S., Gaidon, A.: Multi-frame self-supervised depth with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 160–170 (2022)
20. Guizilini, V., Lee, K.H., Ambruş, R., Gaidon, A.: Learning optical flow, depth, and scene flow without real-world labels. IEEE Robotics and Automation Letters **7**(2), 3491–3498 (2022)
21. Guizilini, V.C., Ambrus, R., Pillai, S., Gaidon, A.: Packnet-sfm: 3d packing for self-supervised monocular depth estimation. ArXiv **abs/1905.02693** (2019), https://api.semanticscholar.org/CorpusID:263858372
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 770–778. IEEE Computer Society (2016)
23. Hinton, G.E., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. ArXiv **abs/1503.02531** (2015), https://api.semanticscholar.org/CorpusID:7200347
24. Hoang, N.H.T., Nguyen, T.H., Nguyen, M.H.V., Tran, T.T., Le, X.H., Nguyen, D.D.: Mldse: Multi-task learning for depth and segmentation estimation. 2023 RIVF International Conference on Computing and Communication Technologies (RIVF) pp. 248–253 (2023), https://api.semanticscholar.org/CorpusID:268611696
25. Hu, C., Li, X., Liu, D., Wu, H., Chen, X., Wang, J., Liu, X.: Teacher-student architecture for knowledge distillation: A survey. ArXiv **abs/2308.04268** (2023), https://api.semanticscholar.org/CorpusID:260704230
26. Kweon, H., Yoon, K.J.: From sam to cams: Exploring segment anything model for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19499–19509 (June 2024)
27. Lee, J.H., Han, M., Ko, D.W., Suh, I.H.: From big to small: Multi-scale local planar guidance for monocular depth estimation. CoRR **abs/1907.10326** (2019)

28. Liu, L., Song, X., Wang, M., Liu, Y., Zhang, L.: Self-supervised monocular depth estimation for all day images using domain separation. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 12717–12726 (2021), https://api.semanticscholar.org/CorpusID:237142261
29. Liu, M., Wang, S., Guo, Y., He, Y., Xue, H.: Pano-sfmlearner: Self-supervised multi-task learning of depth and semantics in panoramic videos. IEEE Signal Processing Letters **28**, 832–836 (2021), https://api.semanticscholar.org/CorpusID:233990511
30. Lu, Y., Sarkis, M., Lu, G.: Multi-task learning for single image depth estimation and segmentation based on unsupervised network. 2020 IEEE International Conference on Robotics and Automation (ICRA) pp. 10788–10794 (2020), https://api.semanticscholar.org/CorpusID:221845335
31. Lyu, X., Liu, L., Wang, M., Kong, X., Liu, L., Liu, Y., Chen, X., Yuan, Y.: Hr-depth: High resolution self-supervised monocular depth estimation. ArXiv **abs/2012.07356** (2020), https://api.semanticscholar.org/CorpusID:229152988
32. Maddern, W.P., Pascoe, G., Linegar, C., Newman, P.: 1 year, 1000 km: The oxford robotcar dataset. The International Journal of Robotics Research **36**, 15 – 3 (2017), https://api.semanticscholar.org/CorpusID:22556995
33. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
34. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D: nonlinear phenomena **60**(1-4), 259–268 (1992)
35. Saunders, K., Vogiatzis, G., Manso, L.J.: Self-supervised monocular depth estimation: Let's talk about the weather. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8907–8917 (2023)
36. Saxena, A., Sun, M., Ng, A.Y.: Make3d: Learning 3d scene structure from a single still image. IEEE Trans. Pattern Anal. Mach. Intell. **31**(5), 824–840 (2009). https://doi.org/10.1109/TPAMI.2008.132
37. Vankadari, M.B., Garg, S., Majumder, A., Kumar, S., Behera, A.: Unsupervised monocular depth estimation for night-time images using adversarial domain feature adaptation. In: European Conference on Computer Vision (2020), https://api.semanticscholar.org/CorpusID:222133092
38. Wang, K., Zhang, Z., Yan, Z., Li, X., Xu, B., Li, J., Yang, J.: Regularizing nighttime weirdness: Efficient self-supervised monocular depth estimation in the dark. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 16035–16044 (2021), https://api.semanticscholar.org/CorpusID:236956395
39. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8798–8807 (2018)
40. Wang, Y., Tsai, Y.H., Hung, W.C., Ding, W., Liu, S., Yang, M.H.: Semi-supervised multi-task learning for semantics and depth. 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) pp. 2663–2672 (2021), https://api.semanticscholar.org/CorpusID:238857255
41. Wu, X., Wu, Z., Guo, H., Ju, L., Wang, S.: 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. pp. 15769–15778 (2021)
42. Xu, H., Zhang, J., Cai, J., Rezatofighi, H., Yu, F., Tao, D., Geiger, A.: Unifying flow, stereo and depth estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)

43. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE transactions on image processing **26**(7), 3142–3155 (2017)
44. Zhang, Y., Chen, H., He, Y., Ye, M., Cai, X., Zhang, D.: Road segmentation for all-day outdoor robot navigation. Neurocomputing **314**, 316–325 (2018), https://api.semanticscholar.org/CorpusID:52175530
45. Zheng, Y., Zhong, C., Li, P., Gao, H., Zheng, Y., Jin, B., Wang, L., Zhao, H., Zhou, G., Zhang, Q., Zhao, D.: Steps: Joint self-supervised nighttime image enhancement and depth estimation. 2023 IEEE International Conference on Robotics and Automation (ICRA) pp. 4916–4923 (2023), https://api.semanticscholar.org/CorpusID:256503497
46. Zhou, T., Brown, M.A., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 6612–6619 (2017), https://api.semanticscholar.org/CorpusID:11977588
47. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 2242–2251 (2017), https://api.semanticscholar.org/CorpusID:206770979
48. Zou, Y., Yu, Z., Kumar, B.V.K.V., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: European Conference on Computer Vision (2018), https://api.semanticscholar.org/CorpusID:52954862