DASP: Self-Supervised Nighttime Monocular Depth Estimation With Domain Adaptation of Spatiotemporal Priors

Peer-reviewed author version

Huang, Yiheng; CHEN, Junhong; Ning, Anqi; Liang, Zhanhong; MICHIELS, Nick; CLAESEN, Luc & Liu, Wenyin (2025) DASP: Self-Supervised Nighttime Monocular Depth Estimation With Domain Adaptation of Spatiotemporal Priors. In: Ieee Robotics and Automation Letters, 11 (2) , p. 2074 -2081.

DOI: 10.1109/LRA.2025.3644148
Handle: http://hdl.handle.net/1942/48125

DASP: Self-supervised Nighttime Monocular Depth Estimation with Domain Adaptation of Spatiotemporal Priors

Yiheng Huang, Junhong Chen, *Member, IEEE,* Anqi Ning, Zhanhong Liang, Nick Michiels, Luc Claesen, *Life Senior Member, IEEE,* and Wenyin Liu, *Senior Member, IEEE*

***Abstract*—Self-supervised monocular depth estimation has achieved notable success under daytime conditions. However, its performance deteriorates markedly at night due to low visibility and varying illumination, e.g., insufficient light causes textureless areas, and moving objects bring blurry regions. To this end, we propose a self-supervised framework named DASP that leverages spatiotemporal priors for nighttime depth estimation. Specifically, DASP consists of an adversarial branch for extracting spatiotemporal priors and a self-supervised branch for learning. In the adversarial branch, we first design an adversarial network where the discriminator is composed of four devised spatiotemporal priors learning blocks (SPLB) to exploit the daytime priors. In particular, the SPLB contains a spatial-based temporal learning module (STLM) that uses orthogonal differencing to extract motion-related variations along the time axis and an axial spatial learning module (ASLM) that adopts local asymmetric convolutions with global axial attention to capture the multiscale structural information. By combining STLM and ASLM, our model can acquire sufficient spatiotemporal features to restore textureless areas and estimate the blurry regions caused by dynamic objects. In the self-supervised branch, we propose a 3D consistency projection loss to bilaterally project the target frame and source frame into a shared 3D space, and calculate the 3D discrepancy between the two projected frames as a loss to optimize the 3D structural consistency and daytime priors. Extensive experiments on the Oxford RobotCar and nuScenes datasets demonstrate that our approach achieves state-of-the-art performance for nighttime depth estimation. Ablation studies further validate the effectiveness of each component.**

Manuscript received: June, 17, 2025; Revised September, 29, 2025; Accepted December, 4, 2025. This paper was recommended for publication by Editor Editor M. Vincze upon evaluation of the Associate Editor and Reviewers' comments. The work of Chen Junhong was supported by China Scholarship Council under Grant 202208440309. This work was supported in part by the National Natural Science Foundation of China under Grant 91748107, in part by the Special Research Fund (BOF) of Hasselt University under Grant BOF23DOCBL11, and in part by the Guangdong Innovative Research Team Program under Grant 2014ZT05G157. *(Corresponding author: Junhong Chen and Wenyin Liu.)*

Yiheng Huang, Zhanhong Liang, and Wenyin Liu are with the College of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China (e-mail: huangyiheng.gdut@gmail.com; cw252128385@gmail.com; liuwy@gdut.edu.cn).

Junhong Chen is with the College of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China, and also with the Digital Future Lab, Flanders Make, Hasselt University, 3590 Diepenbeek, Belgium (e-mail: CSChenjunhong@hotmail.com).

Anqi Ning is with the College of Engineering, Shantou University, Shantou 515063, China (e-mail: ninganqi.stu@gmail.com).

Nick Michiels is with the Digital Future Lab, Flanders Make, Hasselt University, 3590 Diepenbeek, Belgium.

Luc Claesen is with Hasselt University, 3530 Diepenbeek, Belgium.

Digital Object Identifier (DOI): see top of this page.

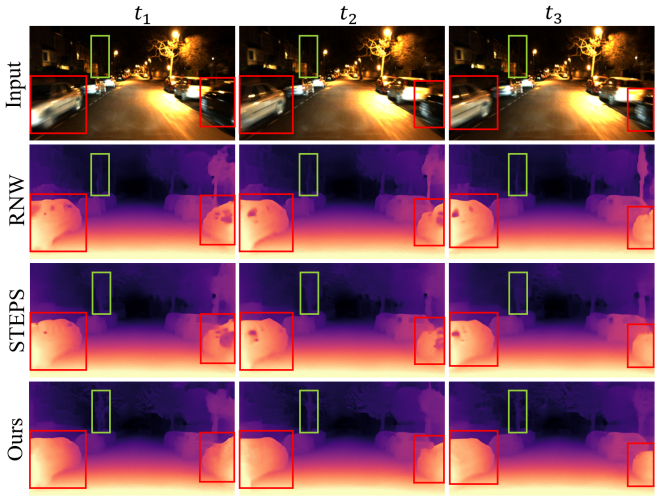

Fig. 1. The first row shows a set of consecutive image frames from the RobotCar dataset. The next three rows show the depth maps predicted by RNW [9], STEPS [10], and our method. The green boxes mark a tree, while the red boxes indicate a moving vehicle. From the figures, we can observe that our method effectively captures spatial structure and maintains consistency in dynamic scenes.

***Index Terms*—Deep Learning for Visual Perception; Deep Learning Methods; Semantic Scene Understanding.**

I. INTRODUCTION

MONOCULAR depth estimation aims to predict dense depth maps from RGB images, which has been widely deployed in various applications, such as 3D scene understanding [1], augmented reality [2], and autonomous driving [3], etc. However, to predict accurate dense depth maps, a large amount of high-quality paired images and depth maps is required, which is tricky to collect from a real-world environment. In this regard, self-supervised methods [4]–[6] have drawn more attention since they do not require costly ground-truth depth labels and estimate the depth through the inference of geometric cues extracted from monocular videos. Moreover, with the efforts of [7], [8], the performance of self-supervised depth estimation is comparable to supervised methods in multiple scenarios, e.g., KITTI, Cityscapes, etc. Unfortunately, these studies mainly focus on daytime depth estimation, with limited performance when facing challenging nighttime scenes.

Low visibility and varying illuminations are the main challenges for nighttime depth estimation, which bring a series of problems. For example, low visibility often results in textureless regions that are difficult to recognize, leading to

depth missing. As shown in Fig. 1, the green box demonstrates a dark region containing a tree that cannot be accurately captured in the depth map. Although the study in [9] proposed to leverage low-light image enhancement to restore details in low-visibility areas, it still cannot generate accurate depth maps for these areas. Varying illuminations usually happen in the moving objects and streetlights, causing a large area of blur. As shown in Fig. 1, the red box indicates motion blur caused by a moving vehicle, where the car windows are inaccurately estimated. Although several studies [10], [11] adopted masking mechanisms to bypass dynamic regions, their methods often suffer from inconsistency and instability in these areas. For example, in consecutive frames, the depth of car windows is estimated in the previous frame, but it is lost in the subsequent sequence.

To address these problems, we propose a self-supervised framework named DASP that transfers the spatiotemporal priors from daytime to nighttime for monocular depth estimation. Specifically, the DASP contains an adversarial branch and a self-supervised branch. In the adversarial branch, we first develop an adversarial network where the discriminator is composed of four spatiotemporal priors learning blocks (SPLB). Particularly, the SPLB includes a spatial-based temporal learning module (STLM) to capture motion-related changes along the time axis, and an axial spatial learning module (ASLM) to extract spatial information along orthogonal axes. The integration of STLM and ASLM provides sufficient spatiotemporal features to restore the textureless and blurry regions. In the self-supervised branch, we propose a 3D consistency projection loss that projects the pixels from both the target and source frames to the same 3D coordinate, and computes the discrepancy between frames to enhance spatial consistency and daytime priors. Extensive experiments on the Oxford RobotCar and nuScenes datasets validate the effectiveness and stability of our approach.

In summary, our main contributions are as follows:

- We propose a self-supervised framework that exploits the spatiotemporal representations from daytime priors for guiding nighttime depth estimation.
- We devise a spatiotemporal priors learning block (SPLB) which consists of two modules: Spatial-based Temporal Learning Module (STLM) and Axial Spatial Learning Module (ASLM). Through the integration of two modules, our model can obtain sufficient spatiotemporal features to restore textureless and blurry regions.
- We design a 3D projection consistency loss which strengthens the geometric consistency and daytime priors.
- Extensive experiments on the Oxford RobotCar and nuScenes datasets demonstrate that our method achieves state-of-the-art performance across multiple metrics.

II. Related Work

A. Self-supervised Depth Learning from Videos

To alleviate the reliance on labeled data, Zhou et al. [4] first proposed self-supervised monocular depth estimation by jointly learning depth and pose. This method is designed based on static scenes and cannot deal with dynamic scenes, leading to multi-view ambiguity. To solve this problem, a number of strategies have been proposed, such as optical flow [5], instance segmentation [12], uncertainty map [13], and stationary pixel mask [8] to recognize moving objects and mask motion regions. Besides, authors in [14], [15] proposed to model 3D object motion, and authors in [16], [17] present to disentangle object motion to construct cost volumes, but these object-level methods lack precise supervision and still remain inherently ambiguous. Recently, Sun et al. [18] leveraged pseudo-depth as depth priors to estimate depth maps and achieved better performance, which verified the effectiveness of geometry priors. Based on it, Mono et al. [19] introduced a ground-contacting prior to handle ambiguous moving objects. Although this approach achieves promising results in daytime scenarios, its performance degrades significantly under nighttime conditions.

B. Nighttime Self-supervised Learning Methods

Considering daytime and nighttime environments have the same structural information, Spencer et al. [20] proposed to learn depth-invariant representations from daytime and nighttime. However, their method performed worse in the low-visibility and illumination variability environments. To address these challenges, quite a few methods have been proposed, which can be divided into two categories: domain adaptation and self-distillation. In domain adaptation, Vankadari et al. [21] and Liu et al. [22] extracted features from daytime and nighttime domains separately, and applied domain adaptation and separation to alleviate the negative effects of poor visibility and uneven lighting. To fully utilize daytime visual cues, Wang et al. [9] introduced a prior derived from daytime depth distributions to enhance nighttime depth prediction. Zheng et al. [10] adopted a different approach using image enhancement to adjust exposure and reduce photometric inconsistency between daytime and nighttime. However, their methods still suffer from the smoothness of depth and texture recognition. In this regard, Cong et al. [11] introduced a composite structure regularization strategy that aligns feature and depth output space to ensure multiscale consistency in structural and textural predictions. In self-distillation, Gasperini et al. [23] adopted GAN to generate adverse samples from daytime images as input, and devised a distillation loss to improve photometric consistency under nighttime conditions. Based on it, Wang et al. [24] proposed learnable visual prompts that capture domain-specific knowledge to enhance cross-domain adaptation. Although these methods utilize daytime data as spatial priors to guide nighttime depth estimation, they overlook the photometric consistency along the temporal dimension. In this work, we pretrain a daytime depth model to produce depth sequences as spatiotemporal priors and leverage adversarial learning to enhance temporal and spatial consistency.

III. Method

A. Self-supervised Training

The task of self-supervised monocular depth estimation is to reproject the pixel p_t in target frame I_t from the pixel p_s in source frame I_s through a depth network Φ_d

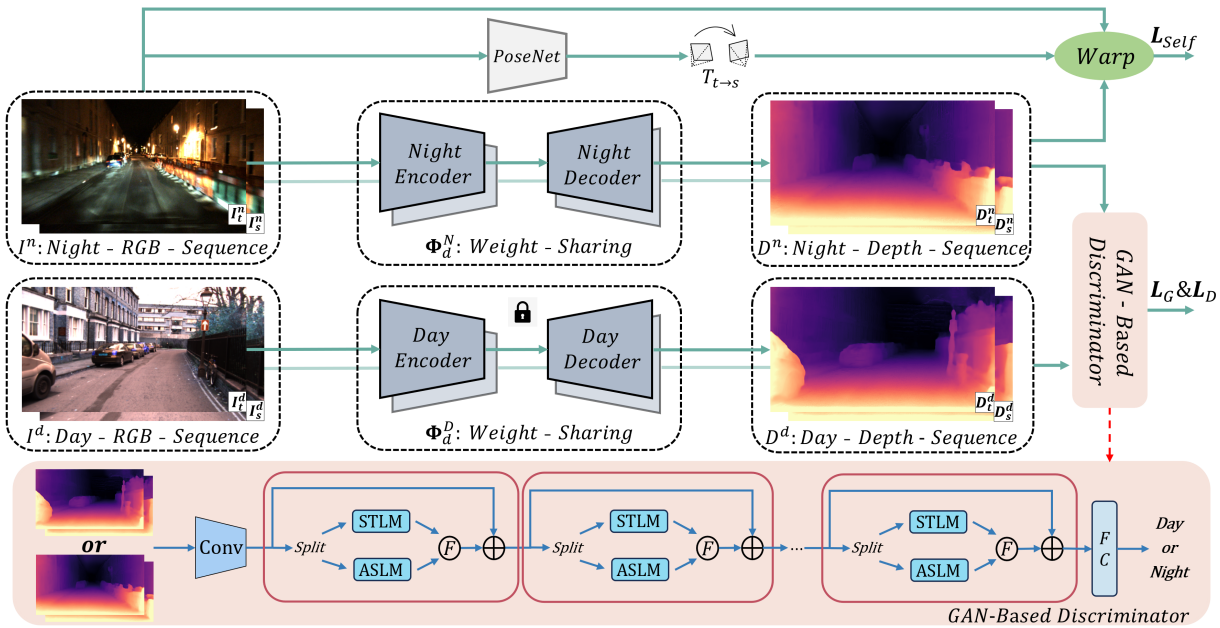

Fig. 2. Overview of our proposed framework. When given a nighttime sequence (I_t^n, I_s^n), we first use a PoseNet to predict relative pose $T_{t\to s}$ and an encoder-decoder network to predict depth maps (D_t^n, D_s^n) respectively, and then warp them to construct a geometric mapping for self-supervised learning. While given a daytime sequence (I_t^d, I_s^d), we adopt a pretrained and fixed model to extract depth priors (D_t^d, D_s^d), and together with nighttime depth maps to feed into a GAN-based discriminator with STLM and ASLM to extract spatiotemporal representation and distinguish day and night. Finally, the framework is jointly optimized with self-supervised and adversarial loss.

and a pose network Φ_p, where the depth network predicts the correspondence depth map $D_t = \Phi_d(I_t)$ and the pose network generates the relative pose $T_{t\to s} = \Phi_p(I_t, I_s)$. The transformation between source pixel p_s and target pixel p_t can be formulated as follows:

$$p_s \sim K T_{t\to s} D_t(p_t) K^{-1} p_t \tag{1}$$

where $\sim$ denotes the homogeneous equivalence and K represents the camera intrinsic matrix. Based on the transformation, the target frame $\hat{I}_t$ can be recovered from I_s by:

$$\hat{I}_t = \langle I_s, p_s \rangle \tag{2}$$

where $\langle\cdot\rangle$ denotes the differentiable bilinear sampling. Intuitively, to reduce the reprojection error, we first follow [25] to introduce a photometric consistency loss that combines a weighted SSIM and weighted ℓ_1 error to calculate the difference between I_t and $\hat{I}_t$:

$$\mathcal{L}_{\text{p}} = \alpha \cdot \frac{1 - \text{SSIM}(I_t, \hat{I}_t)}{2} + (1-\alpha) \cdot \|I_t - \hat{I}_t\|_1 \tag{3}$$

where the weight α is set to 0.85. After that, we follow [26] to apply a disparity smoothness loss to facilitate the smoothness of generated depth and avoid depth ambiguity:

$$\mathcal{L}_{\text{ds}} = |\partial_x D_t| e^{-|\partial_x I_t|} + |\partial_y D_t| e^{-|\partial_y I_t|} \tag{4}$$

Furthermore, we adopt the geometric consistency loss [27] to penalize depth inconsistencies between adjacent frames:

$$\mathcal{L}_{\text{geom}} = \frac{1}{|V|} \sum_{p \in V} D_{\text{diff}}(p) \tag{5}$$

where V is the set of valid projections within image boundaries, and D_{diff} measures per-pixel depth inconsistency between D_t and D_s. This yields a self-discovered mask:

$$M_s = 1 - D_{\text{diff}} \tag{6}$$

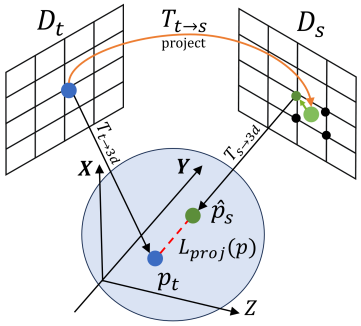

Fig. 3. The computation of 3D projection consistency loss. According to Eq. 1, we first reproject the target point p_t in the target depth map D_t to the source depth map D_s to obtain the interpolated point $\hat{p}_s$. And then two pose networks are deployed to project the points into a shared 3D coordinate. Finally, we calculate the Euclidean distance between two points as a 3D projection consistency loss.

where $M_s \in [0, 1]$ highlights the view-consistent regions while suppressing inconsistent parts.

Although previous works take pixel-wise error and smoothness into consideration, they only focus on frame-level error, i.e., they only compute the unidirectional reprojection from the target frame to the source frame without considering the spatial diversity during reprojection, which makes the model extremely unstable when dealing with occluded or blurred areas. To this end, we design a 3D projection consistency loss that projects the pixel from both the source and target frames into the shared 3D space, and then computes the discrepancy between the two projected pixels, thus optimizing the depth network and pose network. Fig. 3 illustrates the projection process, and the loss is formulated as follows:

$$\mathcal{L}_{\text{proj}} = \left\| D_s(\hat{p}_s) K^{-1} \hat{p}_s - T_{t\to s} D_t(p_t) K^{-1} p_t \right\|_2 \tag{7}$$

where D_s and D_t represent the depth map predicted by depth network, $\hat{p}_s$ is the reprojected pixel after differentiable bilinear sampling.Through the 3D projection consistency loss,

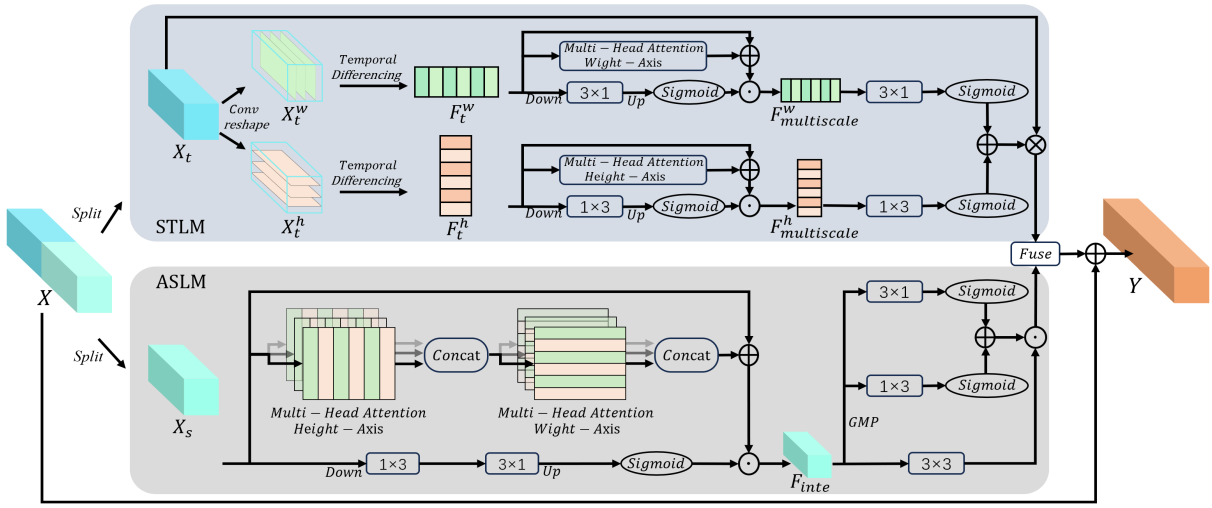

Fig. 4. Overview of spatiotemporal priors learning block (SPLB). First of all, the input X is split into two parts, X_s and X_t, and then they are sent to the Spatial-based temporal learning module (STLM) and the Axial spatial learning module (ASLM), respectively. The STLM captures temporal features via orthogonal differencing of adjacent frames, while the ASLM uses asymmetric convolutions with global axial attention to extract spatial features. Finally, the outputs of the two modules are fused to generate the final spatiotemporal representation Y.

the discrepancy between two projected points can be calculated via depth, which is more direct and intuitive.

Finally, the total self-supervised loss is a weighted combination of the aforementioned losses:

$$\mathcal{L}_{\text{self}} = \lambda_1 \mathcal{L}_{\text{p}} \otimes M_s + \lambda_2 \mathcal{L}_{\text{ds}} + \lambda_3 \mathcal{L}_{\text{geom}} + \lambda_4 \mathcal{L}_{\text{proj}} \quad (8)$$

where $\lambda_1, \lambda_2, \lambda_3$ and λ_4 are empirically set as 0.7, 0.1, 0.5, and 0.5, respectively.

B. Adversarial Adaptation to Learn Spatiotemporal Priors

Considering that GANs have been proven to learn the underlying patterns from data, we adopt a GAN-based network to extract spatiotemporal features from daytime priors. Specifically, for the generator, we use Monodepth2 [8] to generate indistinguishable nighttime samples. Additionally, we pretrain another Monodepth2 on daytime data to provide daytime priors. For the discriminator, we stack four elaborated Spatiotemporal Priors Learning Blocks (SPLB), each of which consists of two branches, STLM and ASLM, to extract appropriate temporal and spatial depth representation patterns from input X. Fig. 4 shows the overview of the SPLB.

1) Spatial-based temporal learning module (STLM). In video sequences, temporal priors could directly guide spatial structure. For example, in the dynamic scenes, previous frames could provide more accurate spatial priors for the subsequent frames. To exploit this spatiotemporal consistency, we devise STLM, which decomposes the entire sequence along the time axis and applies convolutional differences to capture motion-related changes between consecutive frames.

Specifically, given a temporal input $X_t \in \mathbb{R}^{T \times \frac{C}{2} \times H \times W}$, we first compressed it by a factor of r in the channel dimension and reshaped it into a temporal sequence along the horizontal and vertical axes. Because the horizontal branch and vertical branch adopt the same process, we take the horizontal axis as an example here:

$$X_t^w \in \mathbb{R}^{H \times \frac{C}{2r} \times W \times T}, \quad X_t^w = \{x_1^w, x_2^w, \ldots, x_T^w\} \quad (9)$$

where each $x_t^w \in \mathbb{R}^{H \times \frac{C}{2r} \times W}$ denotes the horizontal features at time t. And then we deploy a convolution network with zero-padding to compute directional differences along the time dimension to obtain inter-frame motion features:

$$F_t^w = \zeta_{3\times3}(x_{t+1}^w) - x_t^w, \quad t = 1, \ldots, T-1 \quad (10)$$

where ζ denotes the convolutions, and the subscript indicates the kernel sizes. After that, a three-branch structure with axis-specific asymmetric convolution, axial attention, and a residual connection is used to capture multiscale inter-frame variations:

$$F_{local}^w = \sigma(Up(\zeta_{3\times1}(Down(F_t^w)))) \quad (11)$$

$$F_{multiscale}^w = F_{local}^w \odot (F_t^w + MHA_W(F_t^w)) \quad (12)$$

where σ is the sigmoid function, F_{local}^w and MHA_W denote the local features and global attention [28] along the width axes. To further refine the direction-aware temporal, we apply an activated asymmetric convolution with a sigmoid function:

$$F_{\text{refine}}^w = \sigma(\zeta_{1\times3}(F_{\text{multiscale}}^w)) \quad (13)$$

Finally, we combine two refined axes features and the original time input X_t to generate the final temporal features:

$$F_t = (F_{refine}^w + F_{refine}^h) \odot X_t \quad (14)$$

where F_{refine}^w and F_{refine}^h represent the refined horizontal and vertical features, respectively.

2) Axial spatial learning module (ASLM). Since the street scene images are captured by the cameras and LiDARs on the car, the view of street scenes extends vertically from near to far, while the depth decreases horizontally from near to far. Besides, street scenes often contain many structural objects, such as streetlights, buildings, and cars, which generally follow the distribution of vertical and horizontal axes. Based on these observations, we proposed ASLM, which uses local asymmetric convolutions with global axial attention to extract multiscale structural depth representations and leverage them to guide depth estimation at nighttime.

Specifically, given an input $X_s \in \mathbb{R}^{T \times \frac{C}{2} \times H \times W}$, we apply multi-head self-attention along the height and width axes to extract global structural features:

TABLE I
COMPARISON WITH STATE-OF-THE-ART METHODS, WHERE LOWER ERROR AND HIGHER ACCURACY INDICATE BETTER PERFORMANCE.

Methods	Max Depth	Error ↓				Accuracy ↑		
		Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
RobotCar-Night								
Monodepth2 [8]	40m	0.661	25.213	12.187	0.553	0.551	0.849	0.914
ADDS [22]	40m	0.233	2.344	6.859	0.270	0.631	0.908	0.962
md4all-DD [23]	40m	0.202	1.882	7.929	0.264	0.642	0.921	0.970
ACDepth [29]	40m	0.187	1.633	6.843	0.242	0.703	0.925	0.971
PromptMono [24]	40m	0.206	2.057	6.497	0.246	0.736	0.917	0.966
RNW [9]	40m	0.176	1.323	4.922	0.225	0.772	0.933	0.975
STEPS [10]	40m	0.154	1.108	4.682	0.213	0.803	0.937	0.974
SRNSD [11]	40m	0.136	0.799	4.257	0.194	0.836	0.951	0.983
Ours	40m	**0.132**	**0.786**	**4.125**	**0.187**	**0.849**	**0.958**	**0.988**
Monodepth2 [8]	60m	0.580	21.446	12.771	0.521	0.552	0.840	0.920
ADDS [22]	60m	0.231	2.674	8.800	0.286	0.620	0.892	0.956
md4all-DD [23]	60m	0.206	2.066	7.790	0.262	0.669	0.910	0.967
ACDepth [29]	60m	0.198	1.921	7.372	0.255	0.681	0.913	0.963
RNW [9]	60m	0.185	1.894	7.319	0.246	0.735	0.910	0.965
STEPS [10]	60m	0.170	1.686	6.797	0.234	0.758	0.923	0.968
PromptMono [24]	60m	0.172	1.540	6.567	0.233	0.763	0.924	0.972
SRNSD [11]	60m	0.169	1.450	6.439	0.226	0.768	0.926	0.975
Ours	60m	**0.164**	**1.442**	**6.315**	**0.218**	**0.777**	**0.930**	**0.981**
NuScense-Night								
Monodepth2 [8]	60m	1.185	42.306	21.613	1.567	0.184	0.360	0.504
RNW [9]	60m	0.326	3.999	9.932	0.417	0.492	0.765	0.870
Light-Dark [30]	60m	0.340	4.838	10.136	0.414	0.526	0.772	0.889
STEPS [10]	60m	0.292	3.363	9.120	0.390	0.572	0.805	0.908
Ours	60m	**0.276**	**3.072**	**8.819**	**0.367**	**0.584**	**0.809**	**0.916**

$$F_{global} = MHA_W(MHA_H(X_s)) \tag{15}$$

where MHA_H denotes attention along the height axes. Additionally, to preserve original spatial information, a residual connection is adopted. Considering attention mechanism pays more attention to the global contextual features, we apply two asymmetric convolutions in input X_s to obtain local features and multiply them with the attention features to obtain the integrated features F_{inte}:

$$F_{local} = \sigma(\mathrm{Up}(\zeta_{3\times1}(\zeta_{1\times3}(\mathrm{Down}(X_s))))) \tag{16}$$

$$F_{inte} = F_{local} \odot (F_{global} + X_s) \tag{17}$$

To further refine direction-aware features, we apply global max pooling on F_{inte} and deploy two asymmetric convolutions to compute horizontal and vertical attention maps and add them together:

$$F_{dire} = \sigma(\zeta_{3\times1}(GMP(F_{inte}))) + \sigma(\zeta_{1\times3}(GMP(F_{inte}))) \tag{18}$$

where F_{dire} is direction-aware features. Finally, we use a 3×3 convolution to extract refined features, and multiply with the F_{dire} to obtain the final spatial features F_s:

$$F_s = \zeta_{3\times3}(F_{inte}) \odot F_{dire} \tag{19}$$

3) Integration of spatiotemporal features. To integrate temporal and spatial features, we adopt a 3×3 convolution to fuse the F_t and F_s:

$$Y_{inte} = \zeta_{3\times3}(F_t + F_s) \tag{20}$$

where Y_{inte} is the integrated features. We add it to two branches, and then concatenate them along the channel dimension. The final output Y is formed by concatenating the features and the original input X via a residual connection:

$$Y = \mathrm{Concat}(F_t + Y_{inte},\ F_s + Y_{inte}) + X \tag{21}$$

C. Final loss

In order to optimize the depth maps D^n generated by the nighttime generator Φ_d^N, we introduce a pretrained daytime model Φ_d^D to generate accurate depth maps D^d as priors to confuse the discriminator Φ_A and force the nighttime generator to mimic. The loss function is as follows:

$$\mathcal{L}_{\mathrm{D}} = \frac{1}{2N^d}\sum_{D^d}(\Phi_A(D^d)-1)^2 + \frac{1}{2N^n}\sum_{D^n}(\Phi_A(D^n))^2 \tag{22}$$

$$\mathcal{L}_{\mathrm{G}} = \frac{1}{2N^n}\sum_{D^n}(\Phi_A(D^n)-1)^2 \tag{23}$$

where N^d and N^n are the number of daytime and nighttime training images. Note that the depth maps here are not fixed to two frames, but refer to a sequence of depth maps.

In summary, the final loss is composed of Self-supervised loss, generator loss, and discriminator loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{self}} + \mathcal{L}_{\mathrm{G}} + \mathcal{L}_{\mathrm{D}} \tag{24}$$

IV. EXPERIMENT

A. Dataset

RobotCar. Oxford RobotCar [31] is a large-scale urban driving dataset under diverse conditions. We build RobotCar-Night using left images from the front stereo camera in sequence 2014-12-16-18-44-24, cropped to 1152×672. The training set has 19k frames from the first five splits (excluding stationary frames), and the test set has 411 frames from the fifth and sixth splits. Depth GT for testing is generated using the official toolbox with front LMS LiDAR and INS data.

nuScenes. nuScenes [3] contains 1000 driving scenes in Boston and Singapore. We select 60 nighttime scenes, crop images to 1536×768, and use over 10k frames for training and 500 for testing. Test depth GT is obtained from top LiDAR via the official toolbox.

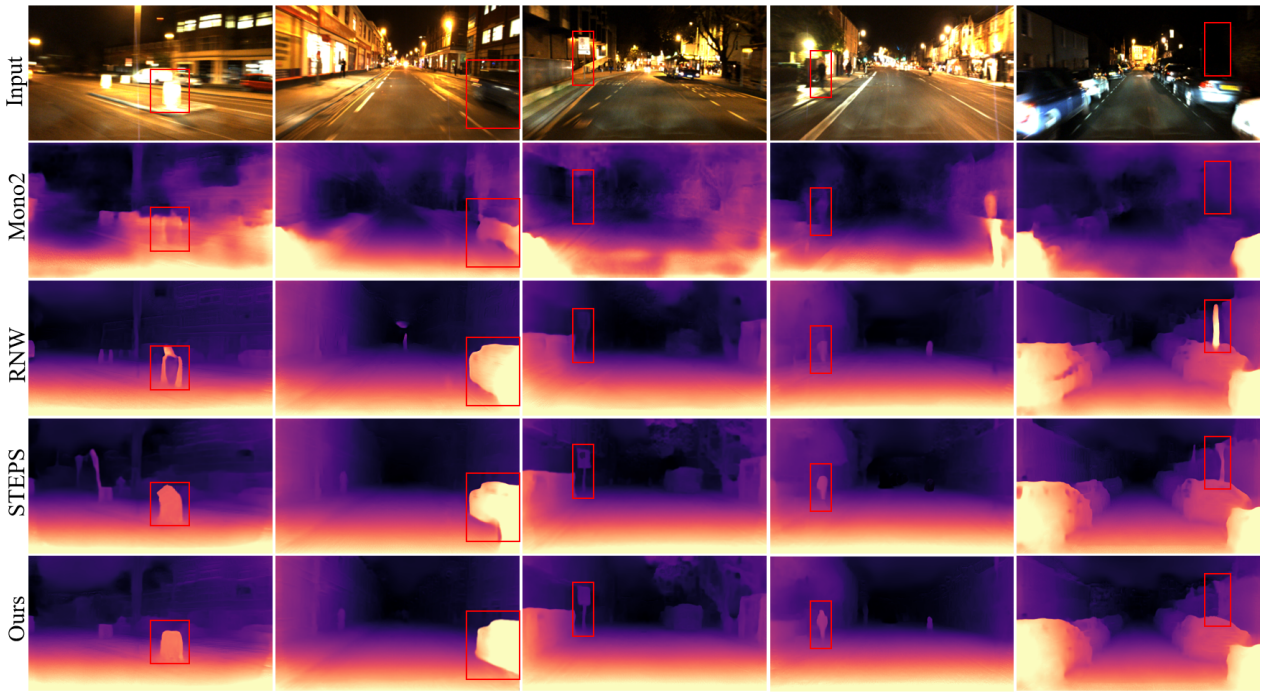

Fig. 5. Qualitative comparison of state-of-the-art methods on the RobotCar dataset, where the key differences are highlighted with red boxes.

B. Implementation Details

The daytime depth estimation network is pretrained based on Monodepth2 [8] to generate spatiotemporal priors, and the nighttime network is optimized by self-supervised training and adversarial learning. The network was trained for 50 epochs on an RTX 3090 GPU using the Adam optimizer with a batch size of 8. We set the initial learning rate as $3e^{-5}$, linearly warmed up to 1×10^{-4} after 500 iterations, and halved at the 15th epoch. To verify the efficiency of our method, we also perform the experiments on an embedded platform NVIDIA Jetson AGX Orin. In terms of model evaluation, we choose seven standard metrics, including: Abs Rel, Sq Rel, RMSE, RMSE log, and accuracy with the thresholds δ of 1.25, 1.25^2, and 1.25^3. Notably, although multi-frame information is utilized during training, only a single frame is required at inference time. For comparison, we compare our method with state-of-the-art monocular nighttime depth estimation approaches, including ADDS [22], RNW [9], STEPS [10], md4all-DD [23], Light-Dark [30], SRNSD [11], PromptMono [24], and ACDepth [29]. All the results are reported under the depth ranges of 40m and 60m, and all comparison approaches are trained and tested on the same dataset.

C. Compare with State-of-the-art Methods

Table I summarizes a quantitative comparison of the state-of-the-art approaches on the Oxford RobotCar dataset [31] and the nuScenes dataset [3]. From the table, we can observe that daytime-oriented methods such as Monodepth2 [8], which is retrained on nighttime data, perform poorly at night. While recent domain adaptation-based methods and self-distillation approaches achieve notable performance improvements. Compared with state-of-the-art methods, our method outperforms other methods on all metrics in the range of 40m and 60m, demonstrating its effectiveness and robustness. In particular, on the RobotCar dataset, our network achieves 2.94% and 2.96% improvements in Abs Rel over the SRNSD in the range of 40m and 60m; while on the nuScenes dataset, our method achieves a 5.48% improvement in Abs Rel compared with STEPS in the range of 60m. We believe it is beneficial to the proposed SPLB that captures the spatiotemporal priors to guide depth estimation and the 3D projection consistency loss to maintain consistency in 3D space. In terms of efficiency, the inference times based on RTX 3090 and NVIDIA Jetson AGX Orin are 163.9 fps and 36.6 fps, respectively.

Intuitively, we visualize several depth estimation examples from RobotCar and nuScenes datasets, where the key differences are highlighted with red boxes. Fig. 5 reveals the examples from RobotCar, from which we can observe that all the methods can well estimate the road surface, but when handling the moving objects and photometrically inconsistent areas, other methods estimate the wrong depth maps or produce inconsistent depth values. In contrast, our proposed method is able to generate clear and smooth depth maps thanks to the SPLB that extracts the spatial features of structural objects (such as guideposts, tree trunks, etc.) and temporal priors of the moving objects (such as cars, pedestrians, etc.). Fig. 6 demonstrates the examples from the more challenging nuScenes dataset, from which we can notice that under low illuminated and noise-corrupted environments, our model is still able to produce accurate depth maps with clear contours and well-preserved structures, e.g., the clear moving car and the distinct guidepost. We believe it is because the 3D projection consistency loss bridges the objects from the target and source frames in a shared 3D space, facilitating the estimation of the objects.

D. Ablation Study

Importance of the 3D Projection Consistency Loss. By comparing #1 and #2 in Table II, we can find that the 3D

Fig. 6. Qualitative comparison of state-of-the-art methods on the nuScenes dataset, where the key differences are highlighted with red boxes.

TABLE II
QUANTITATIVE RESULTS ON ROBOTCAR DATASET. THE DEPTH RANGE IS SET TO 60M, AND THE BEST RESULTS ARE MARKED IN BOLD.

#	Proj	SPLB		Error ↓				Accuracy ↑		
		STLM	ASLM	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
1	×	×	×	0.178	1.758	7.251	0.244	0.742	0.917	0.970
2	✓	×	×	0.176	1.721	7.147	0.239	0.747	0.920	0.971
3	×	✓	×	0.172	1.646	7.083	0.235	0.758	0.922	0.974
4	×	×	✓	0.173	1.653	7.009	0.234	0.757	0.921	0.974
5	×	✓	✓	0.168	1.499	6.830	0.228	0.767	0.924	0.977
6	✓	✓	×	0.167	1.486	6.722	0.225	0.769	0.926	0.977
7	✓	×	✓	0.168	1.493	6.859	0.226	0.767	0.924	0.976
8	✓	✓	✓	**0.164**	**1.442**	**6.315**	**0.218**	**0.777**	**0.930**	**0.981**

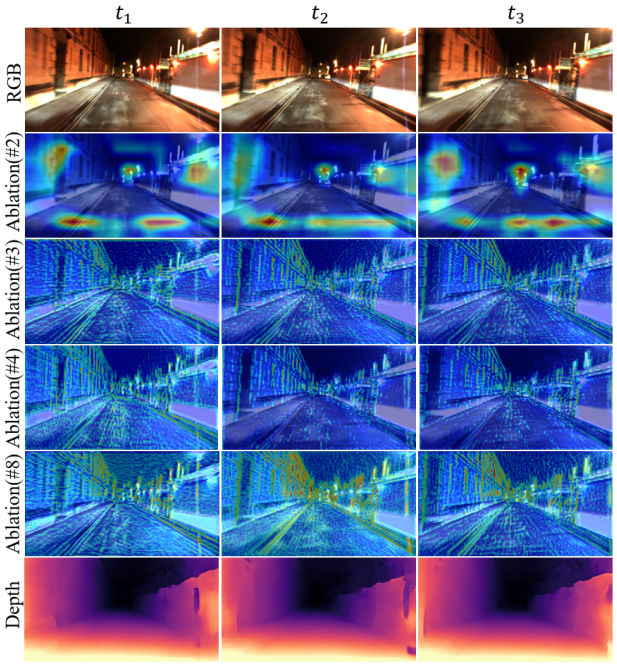

Fig. 7. Visualization of the ablation study.

projection consistency loss can improve certain performance, indicating its effectiveness in enhancing geometric consistency through the alignment of depth predictions in a shared 3D space and optimizing the daytime prior.

Effectiveness of Spatiotemporal Priors Learning Block (SPLB). The SPLB consists of two submodules, the spatial-based temporal learning module (STLM) and the axial spatial learning module (ASLM), so we conduct three ablation studies to verify the effectiveness of these submodules. Specifically, by comparing #2 and #6, as well as #2 and #7 in Table II, we find that both STLM and ASLM have a critical impact on the nighttime depth estimation since STLM captures motion-related variations along the time axis and ASLM extracts the spatial patterns along the orthogonal axis. When combining STLM and ASLM (as shown in #5), our model achieves the best results, effectively demonstrating the advantage of the dual-branch design and the complementary effect of temporal and spatial features in depth estimation.

Effectiveness of the combination of 3D Projection Consistency Loss and Spatiotemporal Priors Learning Block (SPLB). By comparing #2, #3, #4, #5, and #8, we can notice that the improvements are limited when only deploying 3D Projection Consistency Loss or SPLB. However, by combining 3D Projection Consistency Loss and SPLB, the results are significantly improved. We believe this is because the 3D Projection Consistency Loss, as a self-supervised loss, requires sufficient spatial-temporal features and structural information for self-learning, and STLM and ASLM precisely provide the spatial-temporal priors to the 3D Projection Consistency Loss, thereby strengthening the performance of the network.

Visualization and Analysis. To further investigate the effectiveness of each proposed component, we utilize Grad-CAM to visualize the attention maps for each component, including the 3D Projection Consistency Loss (#2), STLM (#3), ASLM (#4), and all components (#8) with three consecutive frames. As shown in Fig. 7, #2 can maintain attention across frames, but lacks structural information. #3 captures the motion-related changes and maintains consistency along the time axis. #4 provides clear spatial information, but cannot maintain temporal consistency. While #8 can produce more stable and concentrated attention along axes and maintain consistent attention on 3D structure over time. These

TABLE III
COMPARISON OF THE NUMBER OF SPATIOTEMPORAL PRIOR FRAMES.

Number of Frames	Error ↓			Accuracy ↑		
	Abs Rel	Sq Rel	RMSE	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Frame(1)	0.179	1.742	7.123	0.744	0.918	0.969
Frame(2)	0.170	1.534	6.681	0.770	0.925	0.975
Frame(3)	**0.164**	**1.442**	**6.315**	**0.777**	**0.930**	**0.981**
Frame(5)	0.168	1.561	6.631	0.769	0.926	0.973

indicate that the 3D projection consistency loss can maintain prior consistency, and the STLM and ASLM can effectively capture accurate spatiotemporal representation.

Impact of the Number of Spatiotemporal Priors Frames. As shown in Table III, we can find that using one or two frames leads to poor performance since too less frames could not provide sufficient temporal features. While using more frames will also lead to a performance drop, since the accumulated inter-frame errors will increase the day-night differences. Using three frames can provide a coherent temporal structure that balances semantic context and temporal consistency, achieving the best performance.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we present DASP, a self-supervised framework that exploits spatiotemporal priors for nighttime monocular depth estimation. Specifically, we first develop an adversarial network where the discriminator consists of four spatiotemporal priors learning blocks (SPLB). Particularly, the SPLB includes a spatial-based temporal learning module (STLM) to capture the motion-related variations along the time axis, and an axial spatial learning module (ASLM) to extract the spatial depth representation. The combination of STLM and ASLM provides sufficient spatiotemporal features for depth estimation. And then we devise a 3D projection consistency loss to strengthen geometric consistency and daytime priors. Extensive experiments conducted on two mainstream datasets demonstrate the effectiveness and stability of our method for nighttime depth estimation. In the future, we will further investigate the robustness of our model, especially in intense lighting environments and heavily blurred scenes.

REFERENCES

[1] J.-B. Weibel, P. Sebeto, S. Thalhammer, and M. Vincze, "Challenges of depth estimation for transparent objects," in *Proc. Int. Symp. Visual. Computing*. Springer, 2023, pp. 277–288.

[2] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "Dtam: Dense tracking and mapping in real-time," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2011, pp. 2320–2327.

[3] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2020, pp. 11 621–11 631.

[4] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017, pp. 1851–1858.

[5] Z. Yin and J. Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018, pp. 1983–1992.

[6] H. Yang, C. Zhao, L. Sheng, and Y. Tang, "Self-supervised monocular depth estimation in the dark: towards data distribution compensation," in *Proc. Int. Joint Conf. Artificial Intell.*, 2024.

[7] Y. Zou, Z. Luo, and J.-B. Huang, "Df-net: Unsupervised joint learning of depth and flow using cross-task consistency," in *Proc. Eur. Conf. Comp. Vis.*, 2018, pp. 36–53.

[8] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2019, pp. 3828–3838.

[9] K. Wang, Z. Zhang, Z. Yan, X. Li, B. Xu, J. Li, and J. Yang, "Regularizing nighttime weirdness: Efficient self-supervised monocular depth estimation in the dark," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2021, pp. 16 055–16 064.

[10] Y. Zheng, C. Zhong, P. Li, H.-a. Gao, Y. Zheng, B. Jin, L. Wang, H. Zhao, G. Zhou, Q. Zhang *et al.*, "Steps: Joint self-supervised nighttime image enhancement and depth estimation," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2023, pp. 4916–4923.

[11] R. Cong, C. Wu, X. Song, W. Zhang, S. Kwong, H. Li, and P. Ji, "Srnsd: Structure-regularized night-time self-supervised monocular depth estimation for outdoor scenes," *IEEE Trans. Image Process.*, 2024.

[12] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova, "Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos," in *Proc. Conf. AAAI*, 2019, pp. 8001–8008.

[13] M. Poggi, F. Aleotti, F. Tosi, and S. Mattoccia, "On the uncertainty of self-supervised monocular depth estimation," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2020, pp. 3227–3237.

[14] H. Li, A. Gordon, H. Zhao, V. Casser, and A. Angelova, "Unsupervised monocular depth learning in dynamic scenes," in *Conference on Robot Learning*, 2021, pp. 1908–1917.

[15] S. Lee, F. Rameau, F. Pan, and I. S. Kweon, "Attentive and contrastive learning for joint depth and motion field estimation," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2021, pp. 4862–4871.

[16] Z. Feng, L. Yang, L. Jing, H. Wang, Y. Tian, and B. Li, "Disentangling object motion and occlusion for unsupervised multi-frame monocular depth," in *Proc. Eur. Conf. Comp. Vis.*, 2022, pp. 228–244.

[17] J. Watson, O. Mac Aodha, V. Prisacariu, G. Brostow, and M. Firman, "The temporal opportunist: Self-supervised multi-frame monocular depth," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2021, pp. 1164–1174.

[18] L. Sun, J.-W. Bian, H. Zhan, W. Yin, I. Reid, and C. Shen, "Sc-depthv3: Robust self-supervised monocular depth estimation for dynamic scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 497–508, 2023.

[19] J. Moon, J. L. G. Bello, B. Kwon, and M. Kim, "From-ground-to-objects: Coarse-to-fine self-supervised monocular depth estimation of dynamic objects with ground contact prior," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2024, pp. 10 519–10 529.

[20] J. Spencer, R. Bowden, and S. Hadfield, "Defeat-net: General monocular depth via simultaneous unsupervised representation learning," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2020, pp. 14 402–14 413.

[21] M. Vankadari, S. Garg, A. Majumder, S. Kumar, and A. Behera, "Unsupervised monocular depth estimation for night-time images using adversarial domain feature adaptation," in *Proc. Eur. Conf. Comp. Vis.*, 2020, pp. 443–459.

[22] L. Liu, X. Song, M. Wang, Y. Liu, and L. Zhang, "Self-supervised monocular depth estimation for all day images using domain separation," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2021, pp. 12 737–12 746.

[23] S. Gasperini, N. Morbitzer, H. Jung, N. Navab, and F. Tombari, "Robust monocular depth estimation under challenging conditions," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2023, pp. 8177–8186.

[24] C. Wang, G. Zhang, Z. Cheng, and W. Zhou, "Promptmono: Cross prompting attention for self-supervised monocular depth estimation in challenging environments," *arXiv preprint arXiv:2501.13796*, 2025.

[25] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[26] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017, pp. 270–279.

[27] J. Bian, Z. Li, N. Wang, H. Zhan, C. Shen, M.-M. Cheng, and I. Reid, "Unsupervised scale-consistent depth and ego-motion learning from monocular video," *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[28] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, "Axial-deeplab: Stand-alone axial-attention for panoptic segmentation," in *Proc. Eur. Conf. Comp. Vis.*, 2020, pp. 108–126.

[29] K. Jiang, J. Cao, Z. Yu, J. Jiang, and J. Zhou, "Always clear depth: Robust monocular depth estimation under adverse weather," in *Proc. Int. Joint Conf. Artificial Intell.*, 2025.

[30] Q. Liang, L. Wang, L. Wang, X. Liu, and G. Wang, "Light-dark: A novel lightweight self-supervised monocular depth estimation in the dark," in *Proc. Int. Conf. Intelligent Computing*, 2024, pp. 3–14.

[31] W. P. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The oxford robotcar dataset," *The International Journal of Robotics Research*, vol. 36, pp. 15 – 3, 2017.