All-day Depth Completion

Vadim Ezhov[1*], Hyoungseob Park[1*], Zhaoyang Zhang[1*], Rishi Upadhyay[2], Howard Zhang[2], Chethan Chinder Chandrappa[2], Achuta Kadambi[2], Yunhao Ba[2], Julie Dorsey[1], Alex Wong[1]

***Abstract*— We propose a method for depth estimation under different illumination conditions, i.e., day and night time. As photometry is uninformative in regions under low-illumination, we tackle the problem through a multi-sensor fusion approach, where we take as input an additional synchronized sparse point cloud (i.e., from a LiDAR) projected onto the image plane as a sparse depth map, along with a camera image. The crux of our method lies in the use of the abundantly available synthetic data to first approximate the 3D scene structure by learning a mapping from sparse to (coarse) dense depth maps along with their predictive uncertainty – we term this, SpaDe. In poorly illuminated regions where photometric intensities do not afford the inference of local shape, the coarse approximation of scene depth serves as a prior; the uncertainty map is then used with the image to guide refinement through an uncertainty-driven residual learning (URL) scheme. The resulting depth completion network leverages complementary strengths from both modalities – depth is sparse but insensitive to illumination and in metric scale, and image is dense but sensitive with scale ambiguity. SpaDe can be used in a plug-and-play fashion, which allows for 24% improvement when augmented onto existing methods to preprocess sparse depth. We demonstrate URL on the nuScenes dataset where we improve over all baselines by an average 12.39% in all-day scenarios, 12.02% when tested specifically for daytime, and 14.95% for nighttime scenes. Code available at : https://github.com/ezhovv/all-day-depth**

I. Introduction

Three-dimensional (3D) reconstruction, i.e., depth estimation, facilitates spatial tasks such as virtual and augmented reality, and autonomous navigation and manipulation. Existing works, from monocular to multi-view depth estimation, are largely trained and tested on well-illuminated environments. But when transferred to low-illumination scenarios, i.e. nighttime, the performance of these methods drops drastically due to a domain gap – a covariate shift in the photometric intensities induced by the change in lighting conditions – and in the absence of light, depth cannot be estimated from solely image-based methods. Efforts to reduce the performance gap mainly focus on re-balancing the training dataset by introducing additional images captured in the low-illumination environments. As manual curation of datasets with ground truth depth is expensive, existing training sets are augmented with images synthesized through means including, but not limited to, synthetic rendering or image-to-image translation using a generative model. However, rendering may introduce a synthetic to real domain gap, and image-to-image translation may introduce artifacts.

Counter to current trends, we instead investigate the use of a sparse range sensor, i.e., LiDAR, in addition to a camera, with the aim to robustly reconstruct the 3D scene structure under different lighting conditions, i.e., well-lit daytime and lowly-illuminated nighttime, for all-day depth estimation. Specifically, our approach estimates ego-centric dense depth maps from synchronized images and sparse point clouds projected onto the image plane, e.g., sparse depth maps. Nonetheless, the process of image-guided sparse point cloud (depth) completion is still ill-posed for each pixel without a measured point and susceptible to the photometric covariate shift. But while the point cloud is sparse, we have strong priors about the natural shapes of objects populating the 3D scene based on the configuration of the sparse points. This prior can serve as a form of inductive bias for depth estimation in regions where photometry is uninformative, i.e., poorly lit. Hence, we propose to approximate the 3D scene structure, from the sparse points, as a dense depth map and additionally estimate its predictive uncertainty to gauge the reliability of the approximated dense depth map. To this end, we leverage the abundance of publicly available synthetic data, where high quality ground truth can be used as supervision.

Once learned, our sparse to dense (SpaDe) approximation module can be used in a plug-and-play fashion by preprocessing sparse depth maps for existing methods, pretrained on daytime scenes, to extend them to all-day scenarios. Using plug-and-play with improved versions of SpaDe also improves overall performance. In another mode, existing models can be augmented with SpaDe, where its outputs (depth and uncertainty) can be adaptively fused with those of the downstream model via an uncertainty-driven residual learning (URL) scheme. We evaluate our approach on three recent depth completion methods on the nuScenes [1] dataset and improve by an average of 12.02% in day, 14.95% in night and 12.39% overall.

Our contributions are (i) a light-weight plug-and-play network (SpaDe) to approximate dense depth with predictive uncertainty from sparse points, and (ii) an uncertainty-driven residual learning scheme that alleviates existing models from the need to learn depth from scratch by leveraging SpaDe as an inductive bias. (iii) Plug-and-play with SpaDe is forward-compatible; future (better) versions of SpaDe can further improve results in a seamless integration manner. To the best of our knowledge, this is the first approach to address all-day

*Equal contribution

[1]Vadim Ezhov, Hyoungseob Park, Zhaoyang Zhang, Julie Dorsey and Alex Wong are with the Department of Computer Science, Yale University, CT 06520, USA, {firstname.lastname}@yale.edu

[2]Rishi Upadhyay, Howard Zhang, Chethan Chinder Chandrappa, Yunhao Ba and Achuta Kadambi are with UCLA, CA 90095, USA {rishiu,chinderc, yhba}@ucla.edu,hwdz15508@g.ucla.edu, achuta@ee.ucla.edu

depth estimation from image and sparse range fusion.

II. Related Works

Supervised depth completion learns a map from images and sparse depth maps to dense depth maps using ground truth. Earlier works undertook approaches of compressing sensing [2] and approximating morphological operators [3]. A line of works catered to sparse data by altering network operations [4], [5], [6], and extending architectures [7], [8]. Employing RGB guidance, [9] proposed early fusion after initial convolution. [6] used encoder features of concatenated modalities to upsample the sparse depth map. [10] extended this approach by two-stage sequential fusion. [11] used multi-scale cascade hourglass network. [12] implemented non-local spatial propagation, improving over fixed-local methods [13]. [14] incorporates cost volume while [15], [16] utilized transformer blocks. Several works incorporated auxiliary data in form of confidence maps [17] and uncertainty estimations [5], [18], [19], [20] for lidar and [21] for radar.

Unsupervised depth completion [22], [23], [24], [25] learn depth by minimizing: sparse depth reconstruction and photometric error between the original image and its reconstruction from other views of the same scene [26], [27]. [28] used Perspective-n-Point [29] and RANSAC [30] to obtain camera pose. [31] applied the losses to test-time adaptation. [24] proposed a calibrated backprojection layer and [32] monitored distillation. [33] expanded the set of augmentations. [34] used line feature from visual SLAM. [35] decouples structure and scale. [8], [36] also learned to approximate dense depth from sparse depth maps, but does not consider uncertainty, nor low illumination scenarios.

All of the above are designed for well-illuminated scenarios. Specifically, unsupervised methods rely on the photometric reconstruction loss, which requires temporal consistency with minimal occlusions in consecutive frames without specular reflections; current unsupervised methods cannot be trained for nighttime scenes. Thus, we explore supervised learning paradigm for all-day depth estimation.

All-day and nighttime depth estimation remain challenging due to a loss of photometric information (low signal to noise) from low illumination and inconsistent exposure. [37] used image enhancement and adaptive masking nighttime scenes. [38] extended the approach to all-day estimation by jointly learning enhancement module. Other works bridged domain gap using image translation [39] and discriminative learning [40] models. [41] instead proposed extracting view-invariant and variant features with an encoder for each domain. [42] demonstrated illumination-invariant photometric loss, compensating for various exposure and motion by image denoising and predicting per-pixel residual flow map. [43], [44] also relied on alternative modality less affected by illumination – thermal images. [44] estimated depth directly from one thermal image while training with RGB images.

Unlike single-modality (monocular) depth estimation, we fuse RGB camera images and synchronized sparse depth maps from LiDAR, which is invariant to illumination changes. We leverage the complementary strengths of these modalities to perform all-day depth estimation without the need for enhancement or image-to-image translation during training.

III. Method

Motivation. Daytime and nighttime images exhibit significant difference in illumination, posing a challenge for depth estimation. To address this we investigate the efficacy of multi-sensor fusion: *Image sensors* (i.e. CMOS sensors in camera) capture dense 2D projections of the 3D scene – photometry is naturally sensitive to illumination. While daytime images typically present distinct object appearances, which allows one to infer object shapes, nighttime images are often presented with low illumination and photometric disturbances (e.g. low signal to noise). Conversely, *range sensors* (e.g. LiDAR, radar) capture a sparse point cloud of the 3D scene. The inherent robustness of range sensors under different lighting conditions motivates us to exploit it as an additional modality for depth estimation. While range sensors can resolve depth, their measurements are often sparse – leading to trade-offs between the image (dense, but sensitive to illumination) and sparse range modalities. To this end, we present a method to cohesively integrate different sensor modalities to estimate depth under "all-day" (day and nighttime) scenarios.

Formulation. Given datasets $\mathcal{D}_{\text{day}}$ and $\mathcal{D}_{\text{night}}$ comprised of daytime and nighttime scenes, respectively, we aim to estimate depth under all-day scenarios i.e. $\mathcal{D}_{\text{all-day}} = (\mathcal{D}_{\text{day}} \cup \mathcal{D}_{\text{night}})$. We assume data samples $(I, z, d^*) \in \mathcal{D}_{\text{all-day}}$, where $I \in \mathbb{R}^{3\times H\times W}$ denotes an RGB image, $z \in \mathbb{R}_+^{H\times W}$ a synchronized point cloud projected onto the image plane as a sparse depth map, and $d^* \in \mathbb{R}_+^{H\times W}$ the ground truth depth. Note: in all-day depth estimation, aside from the variation in illumination between daytime and nighttime images, which degrades performance, there also exists an imbalance in the data for each lighting condition to further exacerbate the errors.

We propose first to learn a deep neural network f_{SpaDe} from synthetic data to approximate the coarse 3D scene structure as a dense depth map and its predictive uncertainty: $[\hat{z}, \hat{\sigma}] = f_{\text{SpaDe}}(z)$ where $\hat{z} \in \mathbb{R}_+^{H\times W}$ refers to the predicted depth map and $\hat{\sigma} \in \mathbb{R}^{H\times W}$ to the uncertainty. Once trained, f_{SpaDe} can be frozen (see Fig. 1-(a)) and augmented onto existing pretrained depth completion networks f_{DC} to enable all-day depth estimation. We hypothesize that the dense depth predicted by f_{SpaDe} will serve as a strong prior to bias predictions in image regions that are uninformative (Fig. 4).

In its "plug-and-play" mode, f_{SpaDe} is used as a preprocessing step to densify the input sparse depth map to a pretrained depth completion network f_{DC}. Sparse depth z and predicted depth $\hat{z}$ are combined into a single input depth map $\tilde{z} \in \mathbb{R}_+^{H\times W}$; any location in $\hat{z}$ where there exists z is substituted with its value: $\tilde{z} = \mathbb{1}_z \cdot z + (1 - \mathbb{1}_z) \cdot \hat{z}$, where $\mathbb{1}_z \in \{0, 1\}^{H\times W}$ is an indicator for positions where positive z exists. The output depth $d \in \mathbb{R}_+^{H\times W}$ is thus $d = f_{\text{DC}}(I, \tilde{z})$ when using SpaDe in plug-and-play.

In a more performant mode, one can train f_{DC} while freezing f_{SpaDe}. Here, f_{SpaDe} serves again as a preprocessing step, where sparse and predicted depth are again combined into a single input $\tilde{z} \in \mathbb{R}_+^{H\times W}$ and concatenated with uncertainty.

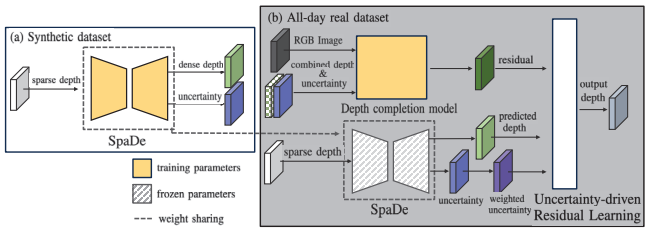

Fig. 1: **An Illustration of Uncertainty-driven Residual Learning**. (a) Our Sparse-to-Dense module (SpaDe) is trained on synthetic data to approximate dense depth from sparse points. (b) SpaDe is used as an inductive bias in Uncertainty-driven Residual Learning (URL), and the depth estimation model is trained to adaptively refine the approximated depth based on the estimated log uncertainty. SpaDe can also be used in a plug-and-play manner to enable all-day depth estimation for a pretrained depth estimator without training.

To indicate original sparse points, uncertainty is imputed with low-uncertainty constant γ: $\hat{\sigma} = \mathbb{1}_z \cdot \gamma + (1 - \mathbb{1}_z) \cdot \hat{\sigma}$, where γ was empirically set to -2. The input now reads $\bar{z} = [\tilde{z}; \hat{\sigma}]$. To alleviate the burden of learning depth from scratch, one may leverage f_{DC} to correct more uncertain regions, allowing one to dedicate model capacity to learning the residual $\hat{d}$ by minimizing typical supervised loss w.r.t. ground truth d^* on $\mathcal{D}_{\text{all-day}}$. The final output is attained with uncertainty-weighted scheme $d = \lambda(\hat{\sigma})\hat{z} + (1 - \lambda(\hat{\sigma}))\hat{d}$, where balancing factor $\lambda(\hat{\sigma})$ grants greater contribution to $\hat{d}$ in higher uncertainty regions. This lends to an uncertainty-driven residual learning scheme (see Fig. 1-(b)) where f_{DC} refines $\hat{z}$ with $\hat{d}$ based on predictive uncertainty $\hat{\sigma}$.

A. Learning to approximate dense depth from sparse range

Given a synthetic dataset with training samples $(z, d^*_{\text{syn}}) \in \mathcal{D}_{\text{syn}}$, where d^*_{syn} denotes the rendered ground truth, we propose a light-weight convolutional encoder-decoder, Sparse-to-Dense network (SpaDe), $f_{\text{SpaDe}}(z) = [\hat{z}, \hat{\sigma}]$, which not only approximates the dense depth $\hat{z}$, but also estimates its log uncertainty $\hat{\sigma}$, from sparse depth map z. To learn f_{SpaDe}, we leverage the high quality, dense ground truth that can be readily obtained in synthetic datasets, sourced from depth buffers in 3D rendering engines, and minimize an L2 loss:

$$\mathcal{L}_{\text{SpaDe-z}}(\hat{z}, d^*_{\text{syn}}) = ||\hat{z} - d^*_{\text{syn}}||^2_2. \tag{1}$$

While the range sensor is robust to illumination changes, the approximated depth may contain erroneous regions, given that it is solely predicted from sparse points. To quantify the reliability of each predicted point, we additionally predict the log-Gaussian uncertainty $\hat{\sigma}$ as a dense map $\hat{\sigma} \in \mathbb{R}^{H \times W}$.

The log uncertainty loss function aims to learn the predictive uncertainty $\hat{\sigma} \in \mathbb{R}^{H \times W}$ of the approximated depth, assuming the uncertainty of $\hat{z}$ follows a Gaussian distribution,

$$\mathcal{L}_{\text{SpaDe}-\sigma}(\hat{z}, d^*_{\text{syn}}) = \frac{1}{2}\left(\frac{\hat{z} - d^*_{\text{syn}}}{e^{\hat{\sigma}}}\right)^2 + \hat{\sigma}. \tag{2}$$

In practice, we train for depth first, and then train the uncertainty decoder while freezing encoder and depth decoder. To show the efficacy of SpaDe as a inductive bias, we visualize the prediction, uncertainty and error on Waymo (Fig. 2).

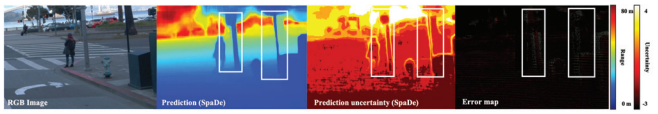

Fig. 2: **Sparse-to-Dense module (SpaDe) on real dataset**. The boxes highlight the alignment between SpaDe's predictive uncertainty and depth discontinuity regions that often erroneous. The uncertainty aligns well with error when compared to ground truth.

B. Uncertainty-driven residual learning (URL)

Given available datasets, one may also augment a depth completion network f_{DC} with SpaDe f_{SpaDe} and train it to learn the residual $\hat{d}$ of the predicted depth map $\hat{z}$, which re-purposes the downstream f_{DC} as a refinement module for $\hat{z}$. To leverage f_{SpaDe} as an inductive bias, the refinement can be conducted adaptively, where the weighting function $\lambda(\hat{\sigma})$ assigns residuals $\hat{d}$ with larger weight for higher $\hat{\sigma}$ (more freedom to modify $\hat{z}$) and likewise lower weight for lower $\hat{\sigma}$:

$$\lambda(\hat{\sigma}) = \frac{1}{1 + e^{\alpha(\hat{\sigma} - \beta)}}, \tag{3}$$

where the $\alpha = 0.8$ and $\beta = 0$ are hyperparameters to calibrate the reliability of the respective models. Our uncertainty-driven residual learning (URL) scheme manifests as a linear combination balanced by $\lambda(\hat{\sigma})$. The output depth reads

$$d = \lambda(\hat{\sigma})\hat{z} + (1 - \lambda(\hat{\sigma}))\hat{d}, \tag{4}$$

where $\lambda(\cdot)$ is the weighting function based on log uncertainty value. Intuitively, $\lambda(\cdot)$ guides the downstream f_{DC} to prioritize the reduction of errors in regions exhibiting high uncertainty.

To learn f_{DC}, we minimize a supervised loss on $\mathcal{D}_{\text{all-day}}$. While any depth completion base model can be seamlessly integrated into our framework, we consider three models for f_{DC} in URL, where each model minimizes the loss function specified in their respective paper, which follows the form:

$$\mathcal{L}_{\text{sup}} = ||d - d^*||_p, \tag{5}$$

where p denotes the L-p norm for the loss, either L1 or L2.

Additionally, as ground truth in real datasets are at most semi-dense, we also include a local smoothness regularizer on d as part of the training objective. Specifically, local smoothness loss is computed on gradients of depth prediction in both horizontal and vertical directions, ∂_X and ∂_Y, respectively. We denote the smoothness loss $\mathcal{L}_{\text{sm}}$ as follows:

$$\mathcal{L}_{\text{sm}} = \frac{1}{|\Omega|}\sum_{x \in \Omega} I_X(x)|\partial_X d(x)| + I_Y(x)|\partial_Y d(x)|, \tag{6}$$

where $I_X(x) = e^{-|\partial_X I(x)|}$, $I_Y(x) = e^{-|\partial_Y I(x)|}$ and Ω denotes image domain. The total loss function for f_{DC} reads:

$$\mathcal{L}_{\text{total}} = w_{\text{sm}}\mathcal{L}_{\text{sm}} + w_{\text{sup}}\mathcal{L}_{\text{sup}}, \tag{7}$$

where w_{sup} and w_{sm} are the weights for each loss term.

IV. Experiments and Results

We consider three recent depth completion models: ENet [10], MSG-CHN[11] and CostDCNet [14]. Each is evaluated using KITTI [45] metrics (MAE, RMSE, iMAE, iRMSE) under daytime, nighttime, and all-day scenarios.

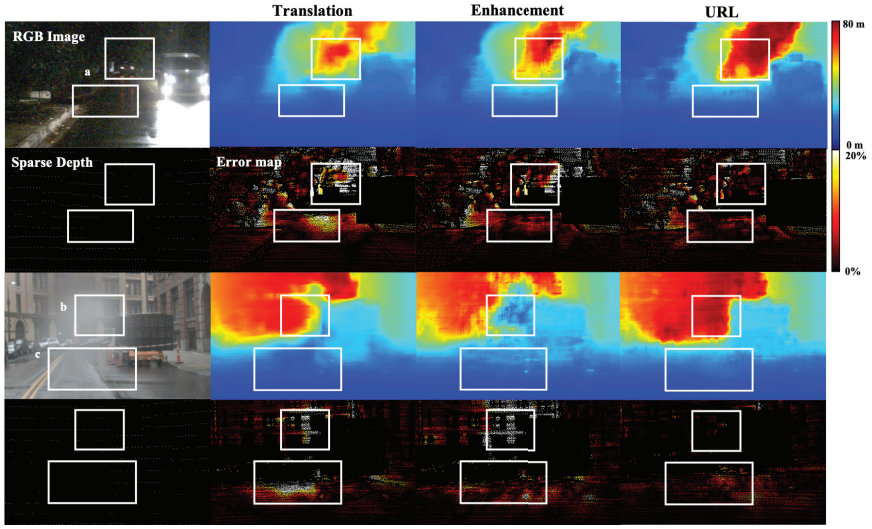

Fig. 3: **Representative results of all-day depth estimation on nuScenes day and night images.** The region for detailed comparisons are highlighted by boxes. URL performs better on (a) low-illumination conditions, (b) depth discontinuity regions and (c) missing sparse points.

A. Datasets

nuScenes [1] is an outdoor dataset comprised of 1000 scenes from Boston and Singapore. The dataset contains around 40,000 annotated keyframes (around 40 samples per scene). We use the original nuScenes train/val split (700 scenes for train, 150 for val). Note: we use the 544x1600 bottom crop to validate the models.

Waymo Open Dataset [46] consists of 1150 scenes from different illuminations (day/night/dawn), Each scene includes average 197 frames with high-quality synchronized LIDAR and RGB image. We used Waymo to train a second set of baselines to test the applicability of SpaDe and a "future" update to SpaDe (trained with more data) for plug-and-play. We show that SpaDe can even improve methods trained with images captured under different illumination conditions.

Virtual KITTI (VKITTI) [47] consists of 35 synthetic videos (5 cloned from the KITTI [4], each with 7 variations in weather, lighting or camera angle) for a total of 1242×375 sized $\approx$17K frames. We only use the dense depth maps of VKITTI to train SpaDe. To acquire the sparse points, we imitate the sparse depth measurement of nuScenes.

SYNTHIA [48] is a synthetic collection of urban scenes, rendered in a virtual city under all four seasons and dynamic illumination conditions. We only utilize dense maps in conjunction with VKITTI to train SpaDe-V2.

B. Implementation details

All models were trained using four NVIDIA RTX 3090 GPUs. MSG-CHN, ENet and CostDCNet used 1, 2, and 4 GPUs respectively. SpaDe was trained on two NVIDIA RTX 3080 Ti GPUs. MSG-CHN took 32hrs with a batch size of 16. ENet and CostDCNet took 55 hours and 35 hours, respectively, with corresponding batch sizes of 16 and 24.

Note: the sparse depth map z, was passed to the geometric convolutional layer of ENet and to the 3-D encoder of CostDCNet as required by their methods. All models take in $\bar{z} = [\tilde{z}, \hat{\sigma}]$ as input to their respective 2-D depth encoder.

SpaDe was trained on VKITTI and later frozen for URL. We trained it for 30 epochs with a learning rate of 2e-4 to learn depth, and later froze the encoder and depth decoder for training uncertainty decoder for a total of 55 epochs with an initial learning rate of 2e-4, and reduce to 1e-4 and 5e-5 at epoch 25 and 40. Random translation, crop, flip, and patch removal augmentations were used. For the baselines, MSG-CHN was trained for 90 epochs with initial learning rate of 1e-3, reduced to 5e-4, 2e-4, 1e-4 at epoch 10, 20 and 25. ENet was trained for 50 epochs under similar schedule, while CostDCNet was trained for 85 epochs using 2e-4 learning rate. We use color jitter, random resize and crop, horizontal flip for augmentations and the crop size of 544x704.

Baselines. We consider KITTI [4] pretrained models of ENet [10], MSG-CHN [11] and CostDCNet [14] as baselines i.e. zero-shot generalization. Note that KITTI does not contain nighttime scenes. We also train baselines on Waymo to show that SpaDe can yield improvements even if a model was pretrained with nighttime imagery. Additionally, we train each method on the original nuScenes dataset, which contains an imbalanced day and night time scenes. We further consider two adaptations of the previously mentioned baselines with image translation and image enhancement respectively.

Image-to-image translation. Following [40], [39], we train depth completion models on the data with extra images translated by day-to-night translation network (Table II, marked with "+Translation"). Due to limited number of night-time images in nuScenes [1], the translation network is first pretrained on BDD [49], then finetuned on nuScenes.

Enhancement baseline. Following [38], we train our baselines with the SCI image enhancement module [50]. SCI estimates the illumination stage by stage and generate a corresponding calibrated residual map, with which the image is enhanced. Although in [38] they jointly train the depth estimation network and the image enhancement module, we directly use the pretrained image enhancement module and freeze it during our training of baselines.

TABLE I: **Plug-and-play evaluation on nuScenes for daytime, nighttime, and all-day depth completion.** SpaDe improves models with plug-and-play and is also forward-compatible with SpaDe-V2. Best results are in **bold**, and second place results are underlined.

		nuScenes-Daytime				nuScenes-Nighttime				nuScenes-All			
Method		MAE	RMSE	iMAE	iRMSE	MAE	RMSE	iMAE	iRMSE	MAE	RMSE	iMAE	iRMSE
SpaDe SpaDe-V2	VKITTI VKITTI+Synthia	1704.850 1621.320	4280.270 4060.945	4.321 4.230	8.765 8.582	1755.366 1693.932	4335.351 4198.802	4.457 4.507	10.391 10.244	1709.207 1628.582	4282.599 4074.733	4.453 4.258	8.951 8.748
MSG-CHN [11]	KITTI Pretrained	2995.773	5708.828	10.973	15.774	2786.535	5231.965	10.209	16.19	2974.845	5661.133	10.897	15.815
	+SpaDe	1697.195	4268.740	**4.380**	8.811	1748.527	4324.559	**4.568**	10.343	1702.329	4274.323	**4.399**	8.964
	+SpaDe-V2	**1636.521**	**4069.945**	4.415	**8.751**	**1706.236**	**4193.282**	4.634	**10.294**	**1643.494**	**4082.281**	4.437	**8.906**
	Waymo Pretrained	1834.319	4241.108	7.279	14.610	1802.856	4195.771	7.425	16.218	1829.338	4232.333	7.286	14.756
	+SpaDe	1674.598	4094.884	6.125	12.204	1766.666	4235.392	6.376	13.725	1683.807	4108.937	6.150	12.356
	+SpaDe-V2	**1628.546**	**3977.431**	**6.082**	**12.144**	**1740.155**	**4172.814**	**6.356**	**13.682**	**1639.709**	**3996.972**	**6.109**	**12.298**
ENet [10]	KITTI Pretrained	7218.710	11207.972	35.189	43.661	8699.404	12980.821	37.604	47.89	7366.805	11385.288	35.431	44.084
	+SpaDe	1765.993	4360.448	4.481	8.832	1884.250	4476.141	4.739	10.435	1777.820	4372.019	4.507	8.993
	+SpaDe-V2	**1748.441**	**4214.698**	**4.381**	**8.662**	**1875.675**	**4374.090**	**4.673**	**10.302**	**1761.166**	**4230.640**	**4.410**	**8.826**
	Waymo Pretrained	1909.895	4337.763	8.585	16.568	2073.584	4549.566	8.822	17.827	1924.354	4354.606	8.600	16.677
	+SpaDe	1821.182	4216.877	7.683	14.775	1981.966	4423.804	7.925	16.004	1835.439	4233.353	7.700	14.883
	+SpaDe-V2	**1776.277**	**4102.502**	**7.649**	**14.715**	**1956.100**	**4364.737**	**7.900**	**15.955**	**1792.483**	**4124.623**	**7.666**	**14.824**
CostDCNet [14]	KITTI Pretrained	2296.108	4962.61	7.342	13.074	2409.223	5249.862	7.235	13.533	2307.421	4991.34	7.331	13.12
	+SpaDe	1796.846	4319.800	**5.178**	9.743	1870.850	4460.740	**5.081**	10.953	1804.248	4333.896	**5.168**	9.864
	+SpaDe-V2	**1745.836**	**4136.099**	5.230	**9.699**	**1838.623**	**4351.694**	5.146	**10.870**	**1755.116**	**4157.662**	5.222	**9.816**
	Waymo Pretrained	1955.058	4415.014	6.957	13.001	2061.947	4681.097	7.099	14.500	1963.792	4437.207	6.964	13.138
	+SpaDe	1797.379	4211.605	6.345	12.129	1870.006	4379.434	6.490	13.597	1804.643	4228.390	6.360	12.275
	+SpaDe-V2	**1743.765**	**4092.958**	**6.300**	**12.060**	**1823.228**	**4284.455**	**6.461**	**13.557**	**1751.713**	**4112.110**	**6.316**	**12.209**

TABLE II: **Evaluation on nuScenes for daytime, nighttime, and all-day depth completion.** Best results are in **bold**, and second best are underlined. Despite not having extra training data or image enhancement, SpaDe and URL improve all three baselines by 12.39%.

		nuScenes-Daytime				nuScenes-Nighttime				nuScenes-All			
Architecture	Method	MAE	RMSE	iMAE	iRMSE	MAE	RMSE	iMAE	iRMSE	MAE	RMSE	iMAE	iRMSE
MSG-CHN [11]	Baseline	1674.465	3926.396	5.749	10.842	2068.051	4470.071	6.801	13.079	1713.830	3980.773	5.854	11.066
	+Translation	1612.436	3837.921	5.650	13.066	1993.668	4564.014	6.066	12.290	1650.566	3910.543	5.692	12.988
	+Enhancement	1723.677	4345.944	4.519	8.747	2223.441	5118.473	6.049	12.37	1773.619	4423.110	4.672	9.110
	+SpaDe	1370.595	3550.412	4.086	8.102	1650.553	3875.170	4.838	**10.211**	1387.848	3563.635	4.073	8.187
	+URL	**1334.965**	**3478.742**	**3.859**	**7.818**	**1522.108**	**3685.869**	**4.703**	10.340	**1353.683**	**3499.458**	**3.943**	**8.070**
ENet [10]	Baseline	1338.449	3389.611	3.968	7.713	1611.057	3706.047	5.322	10.945	1358.298	3406.133	4.081	8.046
	+Translation	1378.031	3449.752	4.582	8.740	1581.408	3644.441	5.773	11.716	1397.238	3473.109	4.711	9.052
	+Enhancement	1314.127	3351.362	3.997	7.831	1771.770	3913.751	6.616	13.592	1359.865	3407.534	4.259	8.407
	+SpaDe	1280.334	**3297.927**	3.932	7.656	1578.806	3634.643	5.176	10.579	1310.143	**3331.935**	4.057	7.951
	+URL	**1215.368**	3337.389	**3.558**	**7.407**	**1439.183**	**3576.976**	**4.507**	**9.994**	**1237.750**	3361.348	**3.653**	**7.666**
CostDCNet [14]	Baseline	1221.396	3326.633	3.896	7.879	1451.260	3608.133	4.857	10.793	1243.161	3351.456	3.988	8.163
	+Translation	1202.946	3334.581	3.666	7.608	1404.093	3623.874	4.239	**9.671**	1221.858	3360.176	3.72	7.807
	+Enhancement	1202.613	3344.348	3.765	7.882	1413.974	3571.865	4.934	11.138	1222.546	3363.755	3.878	8.2
	+SpaDe	**1099.036**	**3312.056**	**3.272**	**7.289**	**1264.984**	3554.725	**3.946**	9.696	**1115.634**	**3336.327**	**3.339**	**7.530**
	+URL	1142.437	3324.694	3.320	7.309	1304.025	**3538.857**	4.036	9.746	1158.598	3346.114	3.392	7.553

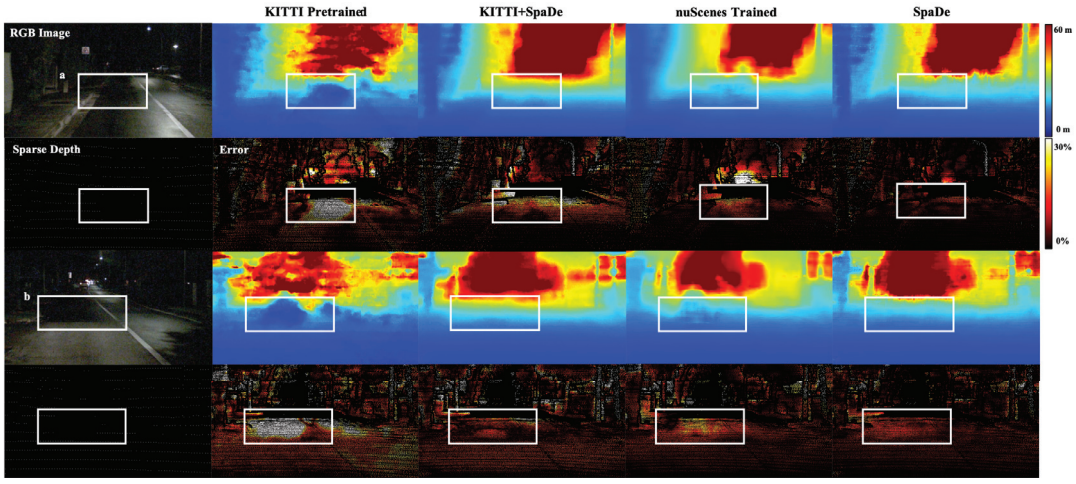

Fig. 4: **Ablation Study on nuScenes** Predictions from SpaDe serve as strong inductive bias for downstream depth completion models. Augmenting KITTI pretrained models with SpaDe improves estimates in regions where photometry is uninformative (highlighted).

C. Main results

Table I shows our study on the effect of illumination change to existing depth completion methods. As expected, models pretrained on KITTI are unable to generalize to nuScenes because of the photometric domain gap; errors in night time are generally higher than daytime. On its own, SpaDe pretrained on VKITTI (row 1) improves over **all KITTI and Waymo pretrained** models, despite not using the RGB image. This shows the robustness of using sparse range as an additional modality to enable transfer of model across various lighting conditions. Augmenting pretrained models with SpaDe (rows with "+SpaDe") significantly improves

results across all metrics in both day and nighttime splits. We additionally trained baselines with Waymo and show that the results hold. This promising result illustrates the plug-and-play capability of SpaDe, which can improve domain generalization while being agnostic to model architecture.

Moreover, all models can easily be improved with future iterations of SpaDe. To demonstrate this, we train SpaDe-V2 on VKITTI and Synthia. We observed consistent improvement when augmented both KITTI and Waymo pretrained models with SpaDe. With a better SpaDe model, downstream models also improves to similar degree, implying seamless integration with updates to SpaDe, i.e., forward-compatibility.

In Table II, we compare with the current trend of using image-to-image translation to re-balance day and nighttime distributions (and to increase the data volume). We observe a positive influence from training the baselines on translated images as compared to those on the original nuScenes with consistent improvement in both day and night time results. When compared to our method, the results were surprising.

Even without URL, if one were to train the downstream model with SpaDe frozen and augmented to process the sparse inputs, there are immediate benefits. This is thanks to the inductive bias coming from the dense depth produced by SpaDe. Additionally, as typical convolutions are not suited for processing sparse inputs [23], providing the depth completion model with dense, albeit an approximation, depth allows the network to properly make use of convolution operations. This is highlighted in MSG-CHN, where we improve the nuScenes model from an MAE of 1713.83 to 1387.85. Moreover, when evaluated on nighttime scenes, all networks augmented with SpaDe received improvements. In fact, without even training on the additional translated images, models augmented with SpaDe are competitive and sometimes even improve over those that were. We also note that image translation requires a large computation overhead during training.

Finally, we compare URL against those trained with image-to-image translation (expanding dataset) and nighttime enhancement. URL improves pretrained baselines by 12.3% in all-day scenarios, (12.17% day, 14.38% night) and 12.39% across *all models* on all-day (12.02% day, 14.95% night). We improve over enhancement by 10.9% (9.7% day, 17.2% night) and translation by 13.93% (14.14% day, 11.85% night).

We attribute such performance to the proposed uncertainty-driven residual learning (URL) scheme. While enhancement module attempts to close the domain gap, the possibility of introducing artifacts arises under dynamic illumination conditions. On the contrary, SpaDe operates on the illumination-robust modality, with potential erroneous regions in prediction quantified by predictive uncertainty. URL, in turn, utilizes adaptive weighting to preserve high confidence regions, re-purposing model capacity to learn the residual. Leveraging SpaDe as a strong depth prior or regularizer, URL acts as a validator given the image, correcting more uncertain regions, which improves over the use of translation or enhancement.

Our method improves over low-illumination region in Fig. 3(a), depth discontinuity in Fig. 3(c), and missing sparse depth regions in Fig. 3(b). This demonstrates the inductive bias from SpaDE enables a robust estimation under low-illuminated conditions and homogeneous regions.

V. Discussion

We have proposed a multimodal fusion method for all-day depth completion, which leverages the strength of complementary sensor configuration under diverse illumination conditions. SpaDe, trained on readily available synthetic data, utilizes sparse range to approximate dense depth and their predictive uncertainty. SpaDe can be used plug-and-play without training, and forward compatibility allows seamless integration of improved SpaDe models to boost performance. Given the target dataset, URL offers improved performance compared to existing methods that rely on image translation techniques, which are prone to introducing artifacts in the training images; training on them may backfire.

Nonetheless, our method does have several limitations. Given the case of high uncertainty in approximated depth and low-illumination in nighttime images, model estimates from URL are not informative. This can be partially mitigated by including predictive uncertainty in downstream models. Our work focuses on ease of use for SpaDe and its applicability in plug-and-play; we leave design of downstream models to future works. Another avenue is to enhance image capture, where computational imaging is relevant. This may bridge the gap between daytime and nighttime images, lending to a more sensor driven paradigm. As our approach is the first all-day depth completion method, we will release code including data setup and processing pipelines and models; we hope our promising results will motivate innovations in enabling robust estimation under diverse illumination settings.

Acknowledgements. This work was supported by NSF-2112562, Athena AI Institute.

References

[1] 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 *CVPR*, 2020.

[2] N. Chodosh, C. Wang, and S. Lucey, "Deep convolutional compressed sensing for lidar depth completion," in *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part I 14*. Springer, 2019, pp. 499–513.

[3] M. Dimitrievski, P. Veelaert, and W. Philips, "Learning morphological operators for depth completion," in *Advanced Concepts for Intelligent Vision Systems*, J. Blanc-Talon, D. Helbert, W. Philips, D. Popescu, and P. Scheunders, Eds. Cham: Springer International Publishing, 2018, pp. 450–461.

[4] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, "Sparsity invariant cnns," in *2017 International Conference on 3D Vision (3DV)*. IEEE, 2017, pp. 11–20.

[5] A. Eldesokey, M. Felsberg, and F. S. Khan, "Propagating confidences through cnns for sparse data regression," in *Proceedings of British Machine Vision Conference (BMVC)*, 2018.

[6] Z. Huang, J. Fan, S. Cheng, S. Yi, X. Wang, and H. Li, "Hms-net: Hierarchical multi-scale sparsity-invariant network for sparse depth completion," 2020.

[7] Y. Chen, B. Yang, M. Liang, and R. Urtasun, "Learning joint 2d-3d representations for depth completion," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 10 023–10 032.

[8] Y. Yang, A. Wong, and S. Soatto, "Dense depth posterior (ddp) from single image and sparse range," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3353–3362.

[9] F. Ma, G. V. Cavalheiro, and S. Karaman, "Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera," 2018.

[10] M. Hu, S. Wang, B. Li, S. Ning, L. Fan, and X. Gong, "Towards precise and efficient image guided depth completion," 2021.

[11] A. Li, Z. Yuan, Y. Ling, W. Chi, C. Zhang, *et al.*, "A multi-scale guided cascade hourglass network for depth completion," in *The IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 32–40.

[12] J. Park, K. Joo, Z. Hu, C.-K. Liu, and I.-S. Kweon, "Non-local spatial propagation network for depth completion," in *European Conference on Computer Vision, ECCV 2020*. European Conference on Computer Vision, 2020.

[13] X. Cheng, P. Wang, C. Guan, and R. Yang, "Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 10 615–10 622.

[14] J. Kam, J. Kim, S. Kim, J. Park, and S. Lee, "Costdcnet: Cost volume based depth completion for a single rgb-d image," in *European Conference on Computer Vision*. Springer, 2022, pp. 257–274.

[15] K. Rho, J. Ha, and Y. Kim, "Guideformer: Transformers for image guided depth completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6250–6259.

[16] Z. Youmin, G. Xianda, P. Matteo, Z. Zheng, H. Guan, and M. Stefano, "Completionformer: Depth completion with convolutions and vision transformers," *arXiv preprint arXiv:2304.13030*, 2023.

[17] W. Van Gansbeke, D. Neven, B. De Brabandere, and L. Van Gool, "Sparse and noisy lidar completion with rgb guidance and uncertainty," in *2019 16th International Conference on Machine Vision Applications (MVA)*. IEEE, 2019, pp. 1–6.

[18] A. Eldesokey, M. Felsberg, K. Holmquist, and M. Persson, "Uncertainty-aware cnns for depth completion: Uncertainty from beginning to end," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[19] C. Qu, T. Nguyen, and C. Taylor, "Depth completion via deep basis fitting," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 71–80.

[20] C. Qu, W. Liu, and C. J. Taylor, "Bayesian deep basis fitting for depth completion with uncertainty," *arXiv preprint arXiv:2103.15254*, 2021.

[21] A. D. Singh, Y. Ba, A. Sarker, H. Zhang, A. Kadambi, S. Soatto, M. Srivastava, and A. Wong, "Depth estimation from camera image and mmwave radar point cloud," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9275–9285.

[22] A. Wong, X. Fei, B.-W. Hong, and S. Soatto, "An adaptive framework for learning unsupervised depth completion," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3120–3127, 2021.

[23] A. Wong, X. Fei, S. Tsuei, and S. Soatto, "Unsupervised depth completion from visual inertial odometry," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1899–1906, 2020.

[24] A. Wong and S. Soatto, "Unsupervised depth completion with calibrated backprojection layers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 747–12 756.

[25] S. S. Shivakumar, T. Nguyen, I. D. Miller, S. W. Chen, V. Kumar, and C. J. Taylor, "Dfusenet: Deep fusion of rgb and sparse depth information for image guided dense depth completion," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 13–20.

[26] X. Fei, A. Wong, and S. Soatto, "Geo-supervised visual depth prediction," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1661–1668, 2019.

[27] A. Wong and S. Soatto, "Bilateral cyclic constraint and adaptive regularization for unsupervised monocular depth prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5644–5653.

[28] F. Ma, G. V. Cavalheiro, and S. Karaman, "Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 3288–3295.

[29] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Epnp: An accurate o(n) solution to the pnp problem," *International Journal of Computer Vision*, vol. 81, 02 2009.

[30] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," in *Readings in Computer Vision*, M. A. Fischler and O. Firschein, Eds. San Francisco (CA): Morgan Kaufmann, 1987, pp. 726–740. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780080515816500702

[31] H. Park, A. Gupta, and A. Wong, "Test-time adaptation for depth completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 519–20 529.

[32] T. Y. Liu, P. Agrawal, A. Chen, B.-W. Hong, and A. Wong, "Monitored distillation for positive congruent depth completion," in *European Conference on Computer Vision*. Springer, 2022.

[33] Y. Wu, T. Y. Liu, H. Park, S. Soatto, D. Lao, and A. Wong, "Augundo: Scaling up augmentations for monocular depth completion and estimation," in *European Conference on Computer Vision*. Springer, 2024.

[34] J. Jeon, H. Lim, D.-U. Seo, and H. Myung, "Struct-mdc: Mesh-refined unsupervised depth completion leveraging structural regularities from visual slam," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6391–6398, 2022.

[35] Z. Yan, K. Wang, X. Li, Z. Zhang, J. Li, and J. Yang, "Desnet: Decomposed scale-consistent network for unsupervised depth completion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, 2023, pp. 3109–3117.

[36] A. Wong, S. Cicek, and S. Soatto, "Learning topology from synthetic data for unsupervised depth completion," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1495–1502, 2021.

[37] 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," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16 035–16 044, 2021.

[38] 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," *arXiv preprint arXiv:2302.01334*, 2023.

[39] A. Sharma, L.-F. Cheong, L. Heng, and R. T. Tan, "Nighttime stereo depth estimation using joint translation-stereo learning: Light effects and uninformative regions," in *2020 International Conference on 3D Vision (3DV)*. IEEE, 2020, pp. 23–31.

[40] M. B. Vankadari, S. Garg, A. Majumder, S. Kumar, and A. Behera, "Unsupervised monocular depth estimation for night-time images using adversarial domain feature adaptation," in *European Conference on Computer Vision*, 2020.

[41] L. Liu, X. Song, M. Wang, Y. Liu, and L. Zhang, "Self-supervised monocular depth estimation for all day images using domain separation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 737–12 746.

[42] M. Vankadari, S. Golodetz, S. Garg, S. Shin, A. Markham, and N. Trigoni, "When the sun goes down: Repairing photometric losses for all-day depth estimation," in *Conference on Robot Learning*. PMLR, 2023, pp. 1992–2003.

[43] N. Kim, Y. Choi, S. Hwang, and I.-S. Kweon, "Multispectral transfer network: Unsupervised depth estimation for all-day vision," in *AAAI Conference on Artificial Intelligence*, 2018.

[44] Y. Lu and G. Lu, "An alternative of lidar in nighttime: Unsupervised depth estimation based on single thermal image," *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 3832–3842, 2021.

[45] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.

[46] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2446–2454.

[47] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," 2016.

[48] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3234–3243.

[49] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," 2020.

[50] L. Ma, T. Ma, R. Liu, X. Fan, and Z. Luo, "Toward fast, flexible, and robust low-light image enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 5637–5646.