

Towards Cross-View-Consistent Self-Supervised Surround Depth Estimation

Laiyan Ding¹, Hualie Jiang², Jie Li³, Yongquan Chen⁴, and Rui Huang^{1*}

Abstract—Depth estimation is a cornerstone for autonomous driving, yet acquiring per-pixel depth ground truth for supervised learning is challenging. Self-Supervised Surround Depth Estimation (SSSDE) from consecutive images offers an economical alternative. While previous SSSDE methods have proposed different mechanisms to fuse information across images, few of them explicitly consider the cross-view constraints, leading to inferior performance, particularly in overlapping regions. This paper proposes an efficient and consistent pose estimation design and two loss functions to enhance cross-view consistency for SSSDE. For pose estimation, we propose to use only front-view images to reduce training memory and sustain pose estimation consistency. The first loss function is the dense depth consistency loss, which penalizes the difference between predicted depths in overlapping regions. The second one is the multi-view reconstruction consistency loss, which aims to maintain consistency between reconstruction from spatial and spatial-temporal contexts. Additionally, we introduce a novel flipping augmentation to improve the performance further. Our techniques enable a simple neural model to achieve state-of-the-art performance on the DDAD and nuScenes datasets. Last but not least, our proposed techniques can be easily applied to other methods. The code is available at <https://github.com/denyngmx/CVCDepth>.

I. INTRODUCTION

Depth perception is a crucial component of reliable autonomous driving and robotics. Nevertheless, due to the high cost of deploying depth sensors, e.g., LiDAR, acquiring high-quality depth from images becomes an attractive alternative. Recent years have witnessed remarkable development of image-based depth estimation [1], [2], [3], [4], [5] and its applications in various scenarios, including 3D object detection [6], [7] and BEV segmentation [8], [9], etc.

In the field of image-based depth estimation, self-supervised depth estimation from images is of particular interest since it eliminates the need for depth supervision or stereo rectification. It utilizes the image reconstruction from

This work was supported by Shenzhen Science and Technology Program under Grants JCYJ20220818103006012 and 20231128093642002, Guangdong Basic and Applied Basic Research Foundation under Grants 2023A1515011347 and 2023A1515110729, Longgang District supporting funds for Shenzhen “Ten Action Plans” under Grant LGKCSOPT2024002, and Research Foundation of Shenzhen Polytechnic University under Grant 6023312007K.

*Corresponding author

¹Laiyan Ding and Rui Huang are with School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen 518172, China (e-mail: laiyingding@link.cuhk.edu.cn; ruihuang@cuhk.edu.cn).

²Hualie Jiang is with Insta360 Research, Shenzhen 518000, China (e-mail: jianghualie@insta360.com).

³Jie Li is with school of Artificial Intelligence, Shenzhen Polytechnic University, Shenzhen 518055, China (e-mail: jieli1@szpu.edu.cn).

⁴Yongquan Chen is with Shenzhen Institute of Artificial Intelligence and Robotics for Society, The Chinese University of Hong Kong, Shenzhen 518172, China (e-mail: yqchen@cuhk.edu.cn).

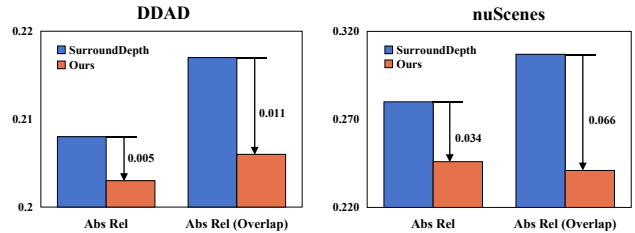


Fig. 1. Performance comparison between ours and previous SOTA SurroundDepth [14] in all regions and overlapping regions.

temporal frames as supervision to train the depth and pose network jointly [3], [10], [11], [12], [13]. However, these methods can only infer scale-ambiguous depth [1].

Recently, self-supervised surround depth estimation has been proposed [15]. This surround-view perception task takes advantage of the multi-camera setup in modern autonomous driving scenarios [16], [10]. The scale-aware poses between cameras and overlapping regions among spatially neighbouring views can help recover scale-aware depth in self-supervised learning. Subsequently, various methods for fusing information among different views have been proposed to improve depth estimation accuracy [14], [17], [18], [19].

In this work, we propose an architectural design and novel losses to enhance cross-view consistency. Firstly, we only use front-view images for pose estimation and the relative camera poses to get the poses of other views. This design is motivated by the fact that front-view depth estimation is better than other views by large margins. Secondly, we introduce a novel dense depth consistency loss to penalize depth-prediction difference in overlapping regions, providing more dense and thus more effective supervision than the loss from MCDP [19]. Thirdly, we propose a loss function to penalize differences in image reconstruction from spatial and spatial-temporal contexts. To further boost the performance, a novel flipping technique for SSSDE is introduced. Current methods typically turn off the widely-used horizontal flipping augmentation since the flipping violates the camera relations. Nevertheless, we manage to apply flipping by modifying the training process carefully.

Together, these contributions yield state-of-the-art SSSDE results on the DDAD [10] and nuScenes [16] datasets with only a simple model. As shown in Figure 1, we achieve lower *Abs Rel* compared to the previous SOTA SurroundDepth [14], especially on the more challenging nuScenes[16] dataset. Moreover, the performance gains in overlapping

regions are almost doubled, indicating that our method can produce more cross-view-consistent results.

II. RELATED WORK

A. Self-Supervised Monocular Depth Estimation

These approaches eliminate the need for ground truth and rectified stereo pairs for depth learning. Instead, they learn depth and motions simultaneously [1], [20], [21], [22], [12], [13]. The supervision signals come from the reconstruction error between the reference image and the reconstructed image from temporally neighbouring frames. Nevertheless, these methods can only produce scale-ambiguous depth [1].

B. Self-Supervised Surround Depth Estimation

FSM [15] is the first work to introduce self-supervised surround depth estimation. It proposes spatial and spatial-temporal reconstruction to recover scale-aware depth. Furthermore, it designs a multi-camera pose consistency loss that penalizes pose-prediction differences among different views. Later, SurroundDepth [14] proposes to predict the pose of the vehicle and transform it back to poses in each view using camera extrinsics. Furthermore, a transformer-based Cross-View Transformer module is utilized to integrate features from different views [23]. VFDepth [17] proposes a canonical pose estimation module that predicts the canonical pose of the front-view camera and distributes it to other views using camera extrinsics. Additionally, feature fusion among different views in 3D voxel space is conducted. MCDP [19] introduces a depth consistency loss and an iterative depth refinement method. More recently, R3D3 [24] utilizes a complex SLAM system [25] and a refinement network to refine depth outputs from the SLAM system. Different from previous works, we propose two general loss functions to enhance the cross-view consistency of the SSSDE outputs.

C. Data Augmentation

Data augmentation is an effective solution to limited data and overfitting. Various augmentation techniques include geometric transformations, color space augmentations, mixing images, adversarial training, and generative adversarial networks [26]. In the field of self-supervised depth estimation, most works [13], [27] follow MonoDepth2 [3] to apply color jittering and horizontal flipping as the training augmentation techniques. However, in the surround depth estimation setup, flipping is non-trivial as it would destroy the geometry relationship between cameras defined by camera extrinsics. In this work, we carefully exploit the widely used horizontal flipping augmentation for its potential in self-supervised surround depth estimation.

III. METHODOLOGY

In this section, we first review the self-supervised depth estimation. Then, we describe our overall architecture, followed by detailed descriptions of our pose estimation design, loss functions, and augmentation methods.

A. Self-Supervised Depth Estimation

Self-supervised monocular depth estimation aims to learn scale-ambiguous depth from a single image to bypass the high cost of collecting depth ground truth [28] for supervision and the need for rectified stereo image pairs [2]. Most of these approaches follow the pioneering work from Zhou *et al.* [1]. The fundamental idea is to reconstruct the reference image with the source image, predicted depth and poses, and differentiable bilinear sampling [29]. The commonly used photometric loss [3] is a weighted combination of Structural Similarity Index Measure (SSIM) [30] and L1 loss. With this photometric supervision, depth and poses can be trained jointly in an end-to-end manner.

Self-Supervised Surround Depth Estimation (SSSDE) is an extension of self-supervised monocular depth estimation, first introduced by FSM [15]. In addition to the temporal learning as in self-supervised monocular depth estimation, FSM [15] also leverages the overlapping region in spatially neighbouring views. Using predicted depth and known camera extrinsics, neighbouring views can be partially reconstructed. This allows for the recovery of scale-aware depth since the extrinsics used here are in absolute scale.

B. Overview of Proposed Architecture

Figure 2 shows the entire network architecture, which includes depth and pose networks. Surround views are fed into the depth network to obtain depth predictions. The pose network takes only the front-view images and outputs the pose in the front-view. By utilizing known camera extrinsics, poses in other views can be recovered. Furthermore, the dense depth consistency loss is applied to predicted depth maps, and the multi-view reconstruction consistency loss is added in image reconstruction. Finally, we apply a novel augmentation method during training.

C. Front-View Pose Only Design

Pose estimation is a key component in self-supervised depth estimation. In the field of SSSDE, previous works have taken advantage of the fact that all cameras are attached to the vehicle and pose consistency among predictions for different views can be enforced. These methods can be grouped into two categories: (1) Separate pose prediction, e.g., FSM [15] predicts the poses for different views separately and adds multi-camera pose consistency constraints. (2) Joint pose Prediction, e.g., SurroundDepth [14] and VFDepth [17] combines the features from all views and decodes the fused features into the vehicle’s pose or canonical pose, and poses for each view can be obtained using camera extrinsics. These approaches are depicted in Figure 3 (a) and Figure 3 (b).

TABLE I

REPRODUCED FSM [15] RESULTS FOR EACH VIEW ON DDAD [10].

Method	Front	F.Left	F.Right	B.Left	B.Right	Back
FSM [15]	0.186	0.245	0.272	0.270	0.274	0.256

Unlike previous methods, we hypothesize that using only front-view images to regress the pose is already effective

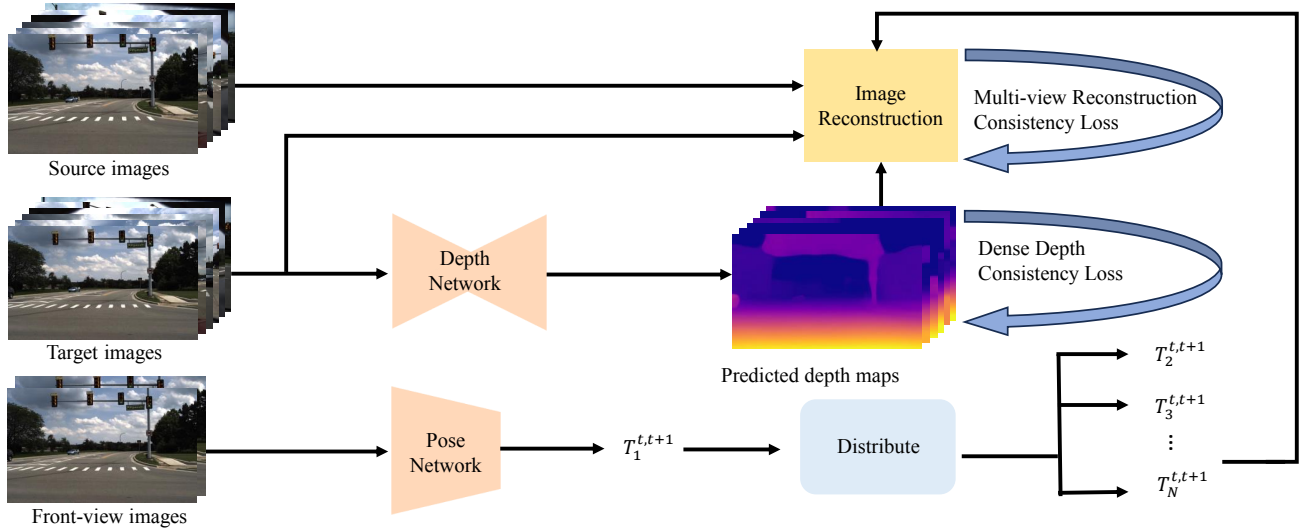


Fig. 2. Overall pipeline of our proposed method. Augmentation is omitted for simplicity.

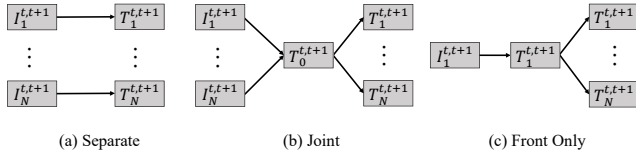


Fig. 3. Comparison of different pose estimation methods.

enough. As shown in Figure 3 (c), we predict the front-view pose using only front-view images and then distribute it to other views. Our motivation is two-fold. First, front view depth estimation is significantly better than other views, as shown in Table I. Due to the tight link between ego-motion and depth predictions [31], we may assume the pose prediction in the front view is better than other views. Second, front-view information is sufficient to predict the front-view pose. Additional information in other views could barely help the front view pose regression. This can be validated by our experiments in Table V. Compared with other pose prediction methods, we only need one pass of encoding and decoding to process a batch of six surround views, while others need at least six passes of encoding, as seen from Figure 3. Consequently, our simple design can reduce memory consumption considerably during training.

To be specific, taking pose prediction from time t to $t+1$ as an example, once we get the pose prediction $T_1^{t,t+1}$ for the front view, we distribute it to view i by leveraging the camera extrinsics E_i as follows:

$$T_i^{t,t+1} = E_i^{-1} E_1 T_1^{t,t+1} E_1^{-1} E_i \quad (1)$$

D. Dense Depth Consistency Loss

Multi-view consistency is a common challenge in self-supervised depth estimation. SC-Depth [31] introduces temporal geometry consistency in monocular estimation and performs backward warping directly. However, MCDP [19]

notices that due to the large difference in camera viewpoints, the depth maps estimated from spatially neighbouring cameras cannot be directly compared. Thus they propose Depth Consistency Loss (DCL) with forward warping, which leads to sparse supervision due to discretization. VFDepth [17] maintains consistency of depth map at novel viewpoints. Yet, they require a 3D representation of the scene to synthesize the depth. Our idea shares the same insight as above to enforce depth consistency among different views. Instead, we propose a Dense Depth Consistency Loss (DDCL) by transforming the source depth beforehand and performing backward warping later. Consequently, DDCL results both dense and correct supervision. Compared with previous methods, our DDCL is more effective and easy to apply for various architectures.

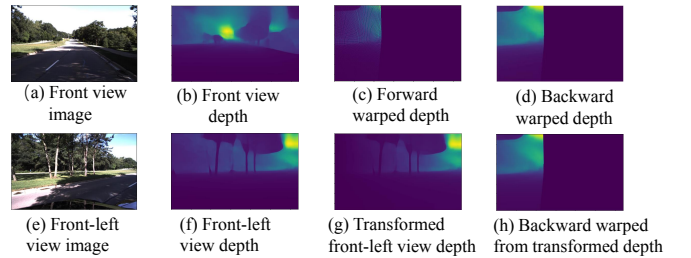


Fig. 4. Different ways to project depth from spatially neighbouring views.

As illustrated in Figure 4, taking the front and front-left images as an example, two depth maps are predicted. To project the depth of points from the front-left-view to the front-view, one may consider backward warping to generate a dense depth map (Figure 4d). However, due to the viewpoint difference, the same objects in the world coordinate would have different depths when projected into different camera views. Thus, backward warping is not the right way. One simple and correct way is to perform forward warping [19]. But, this will result in depth with holes due to discretization

(Figure 4c). To overcome the above disadvantages, which may lead to suboptimal performance, we propose the novel DDCL. We first transform the front-left depth map by assuming it is in the front-view (Figure 4g). That is, for each point, we leverage its homogeneous coordinate, predicted depth, and extrinsics to compute its depth in the front-view. Then we perform backward warping to get the projected depth map (Figure 4h) with bilinear sampling [1]. Compared with forward warping [19], more points in the target depth are supervised since there are no holes now.

For a surround view of N cameras, where for camera i , the predicted depth map is D_i , the depth projected from neighbouring views using aforementioned operations is \tilde{D}_i , the DDCL is calculated as L1 loss:

$$L_{DDCL} = \sum_{i=1}^N \|D_i - \tilde{D}_i\|_1 \quad (2)$$

E. Multi-View Reconstruction Consistency Loss

Following FSM [15], we apply the spatial reconstruction to the recover metric scale and spatial-temporal reconstruction to further incorporate larger spatial-temporal contexts. A natural idea is to enforce the consistency between the reconstruction from spatial and spatial-temporal contexts. This can be achieved by adding one more reconstruction loss between the two reconstructed images, as illustrated in Figure 5. Since this loss involves reconstructions from two views, i.e., the spatially neighbouring view and the spatial-temporally neighbouring view, it is named Multi-View Reconstruction Consistency Loss (MVRCL). From the perspective of the source image, this loss can be interpreted as maintaining temporal consistency under the target view.

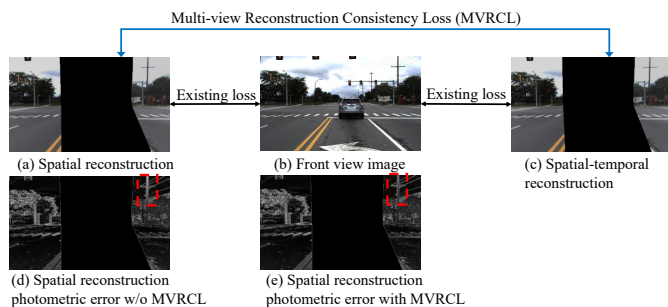


Fig. 5. Our proposed multi-view reconstruction consistency loss can reduce spatial reconstruction errors.

Consequently, the spatial photometric error is reduced by comparing the red region in Figure 5 (d) and (e). This is similar to the phenomenon reported by FSM [15], where photometric error is reduced after applying spatial-temporal reconstruction. Our additional reconstruction consistency loss drives this photometric error from spatial contexts to be smaller. Note that we do not enforce constraints between temporal and spatial or spatial-temporal reconstruction since images from different views can have quite different appearances due to large variations of viewpoint and changes of illumination.

Given a surround view of N cameras, for camera i , the original image is I_i , the reconstructed image from spatial contexts is \tilde{I}_i^s , the reconstructed image from spatial-temporal contexts is \tilde{I}_i^{st} , and the MVRCL is calculated as an image reconstruction loss:

$$L_{MVRCL} = \sum_{i=1}^N (1 - \alpha) \|\tilde{I}_i^s - \tilde{I}_i^{st}\|_1 + \alpha \frac{1 - \text{SSIM}(\tilde{I}_i^s, \tilde{I}_i^{st})}{2} \quad (3)$$

where α is the weight of SSIM loss [30].

F. Augmentation for Self-Supervised Surround Depth Estimation

Augmentation is a crucial component to the success of deep learning methods [26]. Self-supervised monocular depth estimation often utilizes horizontal flipping as a geometric augmentation [3]. However, in the field of SSSDE, horizontal flipping cannot be applied naively. The problem lies in the camera extrinsics, which describes the geometric relationship between different cameras. Flipping the images would destroy such relationships.

To tackle this problem, we propose a novel **H**orizontal-**f**lip augmentation for self-supervised **S**urround depth estimation (**H**flip-**S**) that operates differently on depth network and pose network. The core idea is that outputs from the networks, with flipping applied to inputs, should be appropriately transformed as if the inputs were not flipped.

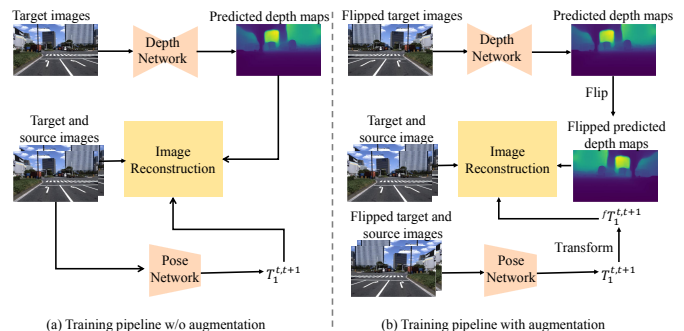


Fig. 6. Comparison of training pipeline using augmentation or not.

We apply the Hflip-S with a probability of 50%. Figure 6 shows the training process with or without flipping. When we apply horizontal flipping, the required operations for depth and pose prediction are different. For depth prediction, we input the flipped version of inputs to the network and flip the depth prediction back. For pose prediction, we transform the pose prediction for flipped inputs back using Eqn. A.5 from the appendix. By making modifications to the pipeline as shown in Figure 6 (b), a novel augmentation for SSSDE is introduced. Furthermore, flipping for depth and pose networks can be used standalone, though we have validated that combining both can lead to better performance.

IV. OVERALL TRAINING LOSS

Our overall training loss follows FSM [15] with our proposed losses added:

$$L = L_t + \lambda_s L_s + \lambda_{st} L_{st} + \lambda_{smooth} L_{smooth} + \lambda_{DCCL} L_{DCCL} + \lambda_{MVRCL} L_{MVRCL} \quad (4)$$

where L_t, L_s, L_{st} are temporal, spatial, spatial-temporal photometric losses, and L_{smooth} is the edge-aware smoothness.

V. EXPERIMENTS

This section will first describe the datasets used and implementation details. We then compare our methods with other state-of-the-art methods quantitatively and qualitatively. Next, abundant ablation studies are performed to validate the effectiveness of proposed techniques. Lastly, we validate the versatility of our techniques on VFDepth [17].

A. Datasets and Implementation Details

Self-supervised surround depth estimation is evaluated on the multi-camera datasets DDAD [10] and nuScenes [16]. Both datasets provide surround-view images, camera intrinsics and extrinsics, and depth ground truth for evaluation. The training resolution is of 384×640 and 352×640 for DDAD and nuScenes respectively, following VFDepth [17]. The model is trained for 20 epochs on both datasets.

We implement our pipeline using Pytorch [32] framework on four NVIDIA RTX 2080Ti GPUs. The network is trained with the following hyperparameters: Adam [33] optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, a batchsize of 6 per GPU (batchsize is 1 for each view), a learning rate of 10^{-4} , with a StepLr scheduler that reduces the learning rate by $\frac{1}{10}$ at $\frac{3}{4}$ of the training epochs. For hyperparameters in Eqn. 4, we set $\lambda_s = 0.03$, $\lambda_{st} = 0.1$, $\lambda_{smooth} = 0.1$, $\lambda_{DDCL} = 1 \times 10^{-3}$, $\lambda_{MVRCL} = 0.2$. Note that our proposed DDCL and MVRCL do not require extra estimation of the depth of adjacent frames. Thus, little overhead and memory consumption are introduced. Also, focal normalization [34] and intensity alignment [17] are applied for stable training.

For evaluation, we follow previous works [17], [14] to evaluate depth predictions up to 200m for DDAD [10] and 80m for nuScenes [16]. Metrics from Eigen *et al.* [35] are adopted. We do not conduct post-processing [3] unless specified.

B. Quantitative Experiments

We compare our methods against other state-of-the-art methods in this section. Scale-aware and scale-ambiguous results are listed in Table II and Table III, respectively. We also include the training memory using a batchsize of 6 (batchsize is 1 for each view) in Table II. Our baseline method is a reproduced variant of FSM [15] where we use the pose network from MonoDepth2[3] following VFDepth [17]. Note that scale-aware evaluation is more meaningful to real applications but more challenging.

For a fair comparison and easier understanding, the following symbols and acronyms are used in Table II and Table III: (1) The symbol \oplus and \star denote results reproduced by VFDepth [17] and us. (2) The symbol \ddagger denotes entirely scale-ambiguous methods. (3) SurroundDepth-M and SurroundDepth-A are the scale-aware and scale-ambiguous models from SurroundDepth [14]. (4) pp means

post-processing [3]. Newly added results that are different from the reported ones are obtained using the original public trained model and codebase under the common protocol.

TABLE II

SCALE-AWARE EVALUATION ON DDAD [10] AND nuSCENES DATASETS [16]. WE REPORT THE AVERAGE RESULTS FROM ALL VIEWS. BEST DEPTH RESULTS AMONG SIMILAR METHODS ARE **BOLDED**.

Dataset	Method	Abs Rel \downarrow	Sq Rel \downarrow	RMSE \downarrow	$\delta < 1.25^\uparrow$	Memory
DDAD	FSM [15]	0.201	-	-	-	-
	FSM \oplus [15]	0.228	4.409	13.433	0.687	7.2GB
	Our baseline (FSM \star [15])	0.252	4.382	14.684	0.551	7.2GB
	VFDepth [17]	0.218	3.660	13.327	0.674	15.9GB
	SurroundDepth-M [14]	0.208	3.371	12.977	0.693	19.7GB
	Ours	0.210	3.458	12.876	0.704	7.0GB
Ours (res34+pp)	0.203	3.363	12.805	0.706	7.6GB	
nuScenes	FSM [15]	0.297	-	-	-	-
	FSM \oplus [15]	0.319	7.534	7.860	0.716	6.8GB
	Our baseline (FSM \star [15])	0.418	13.271	9.210	0.657	6.8GB
	VFDepth [17]	0.289	5.718	7.551	0.709	15.3GB
	SurroundDepth-M [14]	0.280	4.401	7.467	0.661	17.8GB
	Ours	0.264	5.525	7.178	0.763	6.4GB
Ours (res34+pp)	0.246	4.440	6.789	0.764	7.2GB	

TABLE III

SCALE-AMBIGUOUS EVALUATION ON DDAD [10] AND nuSCENES DATASETS [16] WITH PER-FRAME MEDIAN SCALING. WE REPORT THE AVERAGE RESULTS FROM ALL VIEWS. BEST DEPTH RESULTS AMONG SIMILAR METHODS ARE **BOLDED**.

Dataset	Method	Abs Rel \downarrow	Sq Rel \downarrow	RMSE \downarrow	$\delta < 1.25^\uparrow$
DDAD	EGA-Depth-LR \ddagger [18]	0.195	3.211	12.117	0.743
	EGA-Depth-HR \ddagger [18]	0.191	3.126	11.922	0.747
	MCDP \ddagger [19]	0.193	3.111	12.264	0.811
	SurroundDepth-A \ddagger [14]	0.200	3.392	12.270	0.740
	FSM [15]	0.202	-	-	-
	FSM \oplus [15]	0.219	4.161	13.163	0.703
	Our baseline (FSM \star [15])	0.239	4.648	13.461	0.671
	VFDepth [17]	0.221	3.549	13.031	0.681
	SurroundDepth-M[14]	0.205	3.348	12.641	0.716
	Ours	0.208	3.380	12.640	0.716
Ours (res34+pp)	0.204	3.327	12.489	0.720	
nuScenes	EGA-Depth-LR \ddagger [18]	0.239	2.357	6.801	0.723
	EGA-Depth-HR \ddagger [18]	0.223	1.987	6.599	0.732
	MCDP [19] \ddagger	0.237	3.030	6.822	0.719
	SurroundDepth-A \ddagger [14]	0.245	3.067	6.835	0.719
	FSM [15]	0.299	-	-	-
	FSM \oplus [15]	0.301	6.180	7.892	0.729
	Our baseline (FSM \star [15])	0.374	10.243	8.754	0.696
	VFDepth [17]	0.271	4.496	7.391	0.726
	SurroundDepth-M[14]	0.271	3.749	7.279	0.681
	Ours	0.258	4.540	7.030	0.756
Ours (res34+pp)	0.247	3.791	6.704	0.756	

1) *Scale-Aware Evaluation*: For scale-aware evaluation on the DDAD dataset, as shown in Table II, our method outperforms previous arts while using much less training memory. Compared with VFDepth [17], we achieve a boost of 0.008 on *Abs Rel* metric using a resnet18 [36] encoder. To compare with SurroundDepth-M [14], we use a resnet34 [36] encoder and conduct post-processing. We obtain better performance without multi-scale loss [3] and transformer-based fusion. Unfortunately, due to different implementation techniques and hyperparameter settings, both us and the authors from VFDepth [17] cannot reproduce the original FSM [15] results. SurroundDepth [14] even reports that they cannot recover the metric scale when the authors try to reproduce FSM [15]. Nevertheless, our proposed techniques can boost our reproduced FSM [15] by 0.042 on *Abs Rel*.

For scale-aware evaluation on the more challenging nuScenes dataset, where images are taken under different weather conditions and times of day, and the overlapping

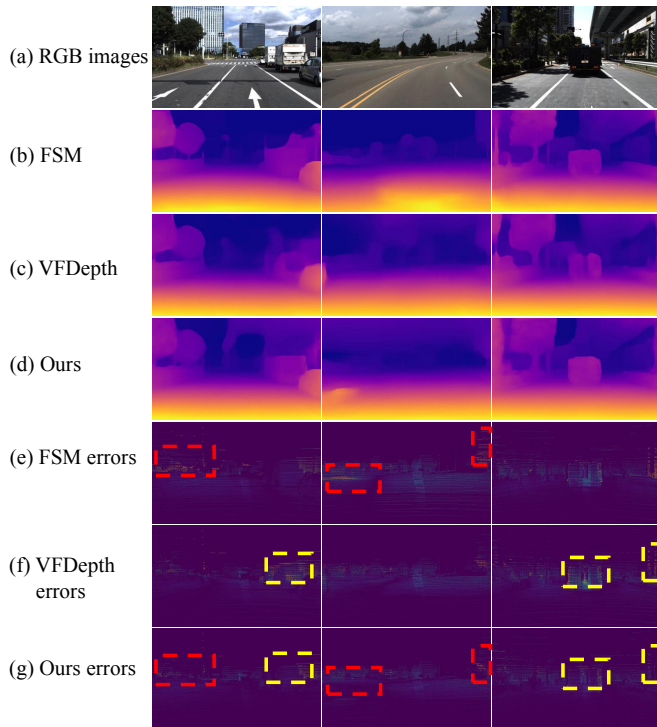


Fig. 7. Qualitative comparisons among our reproduced FSM [15] (our baseline), VFDepth [17] and our full model on the DDAD dataset [10].

regions among cameras are smaller, we achieve even greater performance gains. We obtain an improvement of 0.025 and 0.034 on the *Abs Rel* metric compared with VFDepth [17] and SurroundDepth-M [14], respectively.

2) *Scale-Ambiguous Evaluation*: For scale-ambiguous evaluation, where the predicted depth is per-frame median-scaled [1], the results are shown in Table III.

Compared with scale-aware methods from Table II, we still achieve better results. For example, on DDAD dataset, we get the best performance on three out of four depth metrics, i.e., *Abs Rel*, *RMSE*, $\delta < 1.25$. Furthermore, we outperform the original FSM [15], VFDepth, and SurroundDepth-M [14] by 0.052, 0.074, and 0.024 on *Abs Rel* on the more challenging nuScenes dataset.

Also, note that SurroundDepth-A and SurroundDepth-M achieve a *Abs Rel* of 0.245 and 0.271 on nuScenes [16]. This difference indicates that scale-aware SSSDE is harder to learn. Still, we obtain comparable results against EGA-Depth-LR [18] and MCDP [19], which are entirely scale-ambiguous.

C. Qualitative Results

We provide qualitative results to visualize the predictions of our methods and compare them with previous methods in Figure 7. The FSM [15] and VFDepth [17] results here are reproduced by us. Input images, predicted depth maps, and *Abs Rel* error maps are included in Figure 7. The red rectangles indicate where our method outperforms our reproduced baseline FSM [15]. For example, in the second column, with our proposed techniques applied, errors around

the trees on the right side of the image are reduced, indicating that our techniques improve performance in overlapped regions. Furthermore, the yellow rectangles indicate where our method can outperform our reproduced VFDepth [15]. For instance, in the third column, we achieve lower *Abs Rel* on the car in the middle of the image and the building on the right of the image.

D. Ablation Study

In this section, we first provide an overall ablation study to verify the effectiveness of the proposed techniques. Then, we compare our proposed method with variants to validate the superiority of our method. Lastly, we apply our techniques to VFDepth [17] to demonstrate versatility. All the experiments are conducted on DDAD [10] dataset and scale-aware.

TABLE IV
ABLATION STUDY ON OUR PROPOSED TECHNIQUES.

Front pose	DDCL	MVRCL	Hflip-S	Abs Rel↓	Sq Rel↓	$\delta < 1.25$ ↑
				0.252	4.382	0.551
✓				0.229	4.361	0.676
✓				0.215	3.634	0.693
✓		✓		0.224	4.397	0.697
✓			✓	0.222	4.182	0.702
✓	✓	✓		0.214	3.587	0.694
✓	✓		✓	0.211	3.539	0.692
✓	✓	✓	✓	0.208	3.380	0.716

1) *Overall Ablation Study*: As shown in Table IV, all of our proposed techniques can improve the performance individually and jointly. Overall, we can obtain a large reduction on the *Abs Rel* metric of 0.044. By comparing row 3, row 4, and row 5, we notice that DDCL can reduce both *Abs Rel* and *Sq Rel* greatly, while MVRCL and Hflip-S seem to improve more on the $\delta < 1.25$ metric. This may indicate that DDCL has more impact on the farther points while MVRCL and Hflip-S improve more on the nearer points.

2) *Effectiveness of Front View Pose Only Design*: We verify that using front-view images and camera extrinsics is sufficient to acquire the poses for all views. We test four variants: (a) Pose Consistency: We use the pose consistency loss from FSM [15], which is also our baseline; (b) Joint pose: We follow SurroundDepth [14] to extract features from all views, combine them and then decode them into the vehicle’s poses; (c) Joint front pose: the same as (b) except the features are decoded into poses for the front-view. This is similar to VFDepth [17], except we conduct 2D fusions; (d) Front pose: only use the front-view images to get the poses of the front-view and distribute it to other views.

TABLE V
COMPARISONS BETWEEN DIFFERENT POSE ESTIMATION METHODS.

Method	#Enc↓	#Dec↓	Abs Rel↓	Memory↓
Pose Consistency [15]	6	6	0.252	7.227GB
Joint Pose [14]	6	1	0.228	7.186GB
Joint Front Pose [17]	6	1	0.230	7.186GB
Front pose (Ours)	1	1	0.229	6.113GB

From Table V, we can see that the front-view pose-only design requires one pass of encode and decode, greatly

reducing the memory consumption during training. Furthermore, it achieves very similar results with joint pose or joint front pose. This verifies our intuition that using front-view images is sufficiently effective.

3) *Effectiveness of Dense Depth Consistency Loss*: As mentioned in Section III-D, there exist two correct ways to implement the correct depth consistency loss: (1) DCL [19]: This can only provide sparse constraints on points that are projected. (2) DDCL: This applies transformation first and then uses backward warping to avoid holes.

TABLE VI
EFFECTIVENESS OF OUR DENSE DEPTH CONSISTENCY LOSS.

Method	Abs Rel↓	Sq Rel↓
w/o DC	0.229	4.361
DCL [19]	0.222	3.830
DDCL	0.215	3.634

The comparisons in table VI shows that both implementations can have a performance boost. Nevertheless, our proposed DDCL can be more effective, benefiting from its dense supervision.

TABLE VII
COMPARISONS BETWEEN DIFFERENT AUGMENTATION METHODS.

Method	Abs Rel↓	Sq Rel↓
w/o augmentation	0.229	4.361
depth aug	0.225	4.354
pose aug	0.227	4.362
depth aug and pose aug	0.222	4.182

E. Effectiveness of Our Proposed Augmentation

As shown in Table VII, augmentation for depth network and pose network can be effective alone compared with not using flipping augmentation. And, we find that using the two techniques together yields the best performance.

F. Versatility of Proposed Techniques

In this section, we choose VFDepth [17] to validate the versatility of proposed techniques. By comparing rows one

TABLE VIII
APPLYING OUR PROPOSED TECHNIQUES ON VFDEPTH [17].

Front pose	DDCL	MVRCL	Abs Rel↓
			0.231
✓			0.230
	✓		0.226
		✓	0.222
✓	✓	✓	0.215

and two, we again validate that the front pose design is effective and efficient. Furthermore, DDCL and MVRCL are still effective, though the network architecture between VFDepth [17] and ours are quite different. By using these three techniques together, we achieve an improvement of 0.016 *Abs Rel* against the reproduced VFDepth [17]. However, it is not easy to apply our augmentation techniques since VFDepth would project features into 3D space, thus

prohibiting flipping of the input images. So, we omit the experiment of using Hflip-S.

VI. CONCLUSION AND FUTURE WORKS

In this paper, we have presented a simple model for SSSDE. We introduced four contributions which enable a simple model to achieve superior performance. Nevertheless, some limitations still exist, which may be handled in future works. First, though we have proposed several ways to maintain cross-view consistency, we do not conduct cross-view feature fusions. It is possible to apply techniques from SurroundDepth [14] and MCDP [19] to enhance the model further. Second, it is possible to apply techniques from MVSNet [37], including 3D cost volume building to obtain better performance.

APPENDIX

Here, we provide a proof of conversion between the relative motions of target and source views, and their horizontally flipped counterparts, i.e., conversion between T_t^s and ${}^f T_t^s$.

Suppose that the camera corresponding to the target view is at the world origin, then coordinate transformation from the world coordinate (X_w, Y_w, Z_w) to the source image coordinate (u, v) can be expressed as,

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \simeq \underbrace{\begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}}_K \underbrace{\begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix}}_{T_t^s} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \quad (\text{A.1})$$

When the image of size $h \times w$ is horizontally flipped, the following transformation between the new point on image (u', v') and the original point (u, v) holds:

$$\begin{bmatrix} u' \\ v' \\ 1 \end{bmatrix} = \begin{bmatrix} w-u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} -1 & 0 & w \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \quad (\text{A.2})$$

By combining Eqn. A.1 and Eqn. A.2, we can get:

$$\begin{bmatrix} u' \\ v' \\ 1 \end{bmatrix} \simeq \underbrace{\begin{bmatrix} f_x & 0 & w-c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}}_{K'} \underbrace{\begin{bmatrix} r_{11} & -r_{12} & -r_{13} & -t_1 \\ -r_{21} & r_{22} & r_{23} & t_2 \\ -r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix}}_{{}^f T_t^s} \begin{bmatrix} -X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \quad (\text{A.3})$$

We assume that the principle point is at the center of the image, so we can ignore the difference between K and K' .

Furthermore, after flipping, the world coordinate (X'_w, Y'_w, Z'_w) has the following relationship with the original world coordinate:

$$\begin{bmatrix} X'_w \\ Y'_w \\ Z'_w \\ 1 \end{bmatrix} = \begin{bmatrix} -X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \quad (\text{A.4})$$

From Eqn. A.1, Eqn. A.3 and Eqn. A.4, we can see that the relative motion of two views (T_t^s) and their horizontally flipped counterparts ($^fT_t^s$) has the following relationship :

$$\underbrace{\begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix}}_{T_t^s} \leftrightarrow \underbrace{\begin{bmatrix} r_{11} & -r_{12} & -r_{13} & -t_1 \\ -r_{21} & r_{22} & r_{23} & t_2 \\ -r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix}}_{^fT_t^s} \quad (\text{A.5})$$

REFERENCES

- [1] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1851–1858.
- [2] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5410–5418.
- [3] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3828–3838.
- [4] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 179–12 188.
- [5] S. F. Bhat, R. Birkel, D. Wofk, P. Wonka, and M. Müller, "Zoedepth: Zero-shot transfer by combining relative and metric depth," *arXiv preprint arXiv:2302.12288*, 2023.
- [6] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, and Z. Li, "Bevdepth: Acquisition of reliable depth for multi-view 3d object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 1477–1485.
- [7] Z. Wang, C. Min, Z. Ge, Y. Li, Z. Li, H. Yang, and D. Huang, "Sts: Surround-view temporal stereo for multi-view 3d detection," *arXiv preprint arXiv:2208.10145*, 2022.
- [8] J. Phillion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer, 2020, pp. 194–210.
- [9] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.
- [10] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, "3d packing for self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2485–2494.
- [11] H. Jiang, L. Ding, Z. Sun, and R. Huang, "Dipe: Deeper into photometric errors for unsupervised learning of depth and ego-motion from monocular videos," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 061–10 067.
- [12] C. Zhao, Y. Zhang, M. Poggi, F. Tosi, X. Guo, Z. Zhu, G. Huang, Y. Tang, and S. Mattoccia, "Monovit: Self-supervised monocular depth estimation with a vision transformer," in *2022 International Conference on 3D Vision (3DV)*. IEEE, 2022, pp. 668–678.
- [13] N. Zhang, F. Nex, G. Vosselman, and N. Kerle, "Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 537–18 546.
- [14] Y. Wei, L. Zhao, W. Zheng, Z. Zhu, Y. Rao, G. Huang, J. Lu, and J. Zhou, "Surrounddepth: Entangling surrounding views for self-supervised multi-camera depth estimation," in *Conference on Robot Learning*. PMLR, 2023, pp. 539–549.
- [15] V. Guizilini, I. Vasiljevic, R. Ambrus, G. Shakhnarovich, and A. Gaidon, "Full surround monodepth from multiple cameras," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 5397–5404, 2022.
- [16] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nusenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.
- [17] J.-H. Kim, J. Hur, T. P. Nguyen, and S.-G. Jeong, "Self-supervised surround-view depth estimation with volumetric feature fusion," *Advances in Neural Information Processing Systems*, vol. 35, pp. 4032–4045, 2022.
- [18] Y. Shi, H. Cai, A. Ansari, and F. Porikli, "Ega-depth: Efficient guided attention for self-supervised multi-camera depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 119–129.
- [19] J. Xu, X. Liu, Y. Bai, J. Jiang, K. Wang, X. Chen, and X. Ji, "Multi-camera collaborative depth prediction via consistent structure estimation," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 2730–2738.
- [20] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova, "Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 8001–8008.
- [21] R. Peng, R. Wang, Y. Lai, L. Tang, and Y. Cai, "Excavating the potential capacity of self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 560–15 569.
- [22] C. Shu, K. Yu, Z. Duan, and K. Yang, "Feature-metric loss for self-supervised learning of depth and egomotion," in *European Conference on Computer Vision*. Springer, 2020, pp. 572–588.
- [23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [24] A. Schmied, T. Fischer, M. Danelljan, M. Pollefeys, and F. Yu, "R3d3: Dense 3d reconstruction of dynamic scenes from multiple cameras," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3216–3226.
- [25] Z. Teed and J. Deng, "Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras," *Advances in neural information processing systems*, vol. 34, pp. 16 558–16 569, 2021.
- [26] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.
- [27] X. Lyu, L. Liu, M. Wang, X. Kong, L. Liu, Y. Liu, X. Chen, and Y. Yuan, "Hr-depth: High resolution self-supervised monocular depth estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 2294–2301.
- [28] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2002–2011.
- [29] M. Jaderberg, K. Simonyan, A. Zisserman, et al., "Spatial transformer networks," *Advances in neural information processing systems*, vol. 28, 2015.
- [30] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [31] 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," *Advances in neural information processing systems*, vol. 32, 2019.
- [32] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.
- [33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [34] J. M. Facil, B. Ummerhofer, H. Zhou, L. Montesano, T. Brox, and J. Civera, "Cam-convs: Camera-aware multi-scale convolutions for single-view depth," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 826–11 835.
- [35] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in neural information processing systems*, vol. 27, 2014.
- [36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [37] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "Mvsnet: Depth inference for unstructured multi-view stereo," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 767–783.