

CylinderDepth: Cylindrical Spatial Attention for Multi-View Consistent Self-Supervised Surround Depth Estimation

Samer Abualhanud¹ Christian Grannemann¹ Max Mehlretter¹

¹Institute of Photogrammetry and GeoInformation, Leibniz University Hannover

{abualhanud, grannemann, mehlretter}@ipi.uni-hannover.de

Abstract

Self-supervised surround-view depth estimation enables dense, low-cost 3D perception with a 360° field of view from multiple minimally overlapping images. Yet, most existing methods suffer from depth estimates that are inconsistent between overlapping images. Addressing this limitation, we propose a novel geometry-guided method for calibrated, time-synchronized multi-camera rigs that predicts dense, metric, and cross-view-consistent depth. Given the intrinsic and relative orientation parameters, a first depth map is predicted per image and the so-derived 3D points from all images are projected onto a shared unit cylinder, establishing neighborhood relations across different images. This produces a 2D position map for every image, where each pixel is assigned its projected position on the cylinder. Based on these position maps, we apply an explicit, non-learned spatial attention that aggregates features among pixels across images according to their distances on the cylinder, to predict a final depth map per image. Evaluated on the DDAD and nuScenes datasets, our approach improves the consistency of depth estimates across images and the overall depth compared to state-of-the-art methods. Code is available at <https://github.com/abualhanud/CylinderDepth>

1. Introduction

Depth estimation is an important step in 3D reconstruction and thus a crucial prerequisite for 3D scene understanding, enabling, for example, localization, obstacle avoidance and motion planning in autonomous driving and robotics. Due to the density of observations, the availability of radiometric information, and the comparably low cost, cameras are commonly used for this task. Recent learning-based depth estimation methods, often based on fully-supervised training, produce accurate and dense predictions. However, this requires ground-truth labels, often obtained by additional sensors such as LiDAR, yet, these labels are usually sparse.

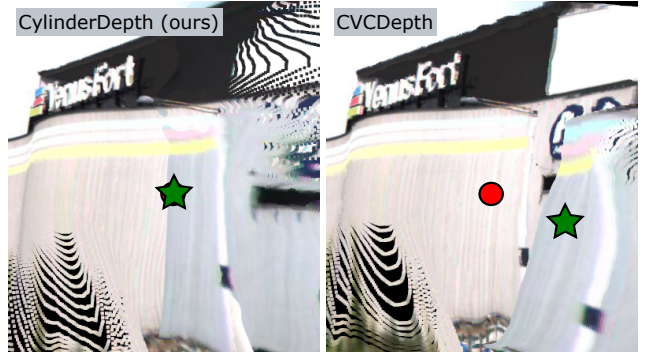


Figure 1. Comparison of multi-view consistency between our method and CVCDepth [4]. The star and circle denote 3D reconstructions of the same 3D object point from two different images. While prior work struggles to achieve consistency in the reconstruction across images, our method overcomes this limitation.

In contrast, self-supervised approaches enforce photometric consistency between images, training on monocular videos, stereo imagery, or both: The pixels of a source image are projected into the coordinate system of a target image, using the estimated depth and known relative orientation parameters, aiming to minimize the color difference between pixels having the same coordinates after projection.

Surround camera setups, which consist of multiple calibrated cameras that are rigidly mounted to each other, provide a full 360° scene coverage and are widely used in autonomous driving [2, 12]. In contrast to a single omnidirectional image, these setups allow to estimate metric-scale depth in overlapping image regions, given that the relative orientation parameters and the length of the baselines between the cameras are known. However, these setups typically provide only minimal spatial overlap. To address this, monocular temporal context is required to increase the effective overlap during training. However, processing each image independently can yield inconsistent depth estimates across cameras; a 3D object point that is visible in multiple

images may get assigned different 3D coordinates per image, resulting in an inconsistent and misaligned reconstruction when combining the results obtained for the individual images. Most prior work enforces multi-view consistency only implicitly during training, e.g., by constraining motion to be consistent across cameras [4, 19, 41], adding loss functions that encourage consistency [4, 14], or using learned attention mechanisms [32, 41]. However, these approaches do not guarantee consistency at inference time, since the cameras’ geometric relationships are not considered.

Addressing this limitation, we propose a novel self-supervised depth estimation method for surround camera setups that enforces multi-view consistency. Given the intrinsic and relative orientation parameters and an initial predicted depth, the 3D points reconstructed from all images are mapped onto a shared unit cylinder. This produces a unified representation across images in which pixels are indexed by cylindrical coordinates and where reconstructions of the same 3D point from multiple images are projected to the same 2D point on the cylinder. Thus, this projection establishes consistent neighborhood relations across images, aligning overlapping image regions. In contrast to approaches that exchange features between images without explicitly modeling their geometric relationship, typically using learned attention, we introduce an explicit, non-learned spatial attention that weights pixel interactions based on the geodesic distances between their cylindrical coordinates. We additionally modulate our spatial attention by feature-space similarity, i.e., we decrease the influence of pixels with dissimilar features even when they are spatially close to each other. Thus, our main contributions are:

- We propose a novel **non-learned geometry-guided spatial attention** mechanism for surround camera setups.
- To enforce multi-view consistency during training and inference, we propose a mapping onto a shared **cylindrical representation**.
- We thoroughly evaluate our proposed method, focusing on **multi-view consistency**. In this context, we further present a novel **depth consistency metric**, closing a relevant gap in the literature.

2. Related Work

Monocular Depth Estimation In monocular depth estimation, a dense, per-pixel depth map is predicted from a single RGB image, which is an ill-posed task. Learning semantic and geometric cues, supervised methods [1, 5, 7, 25, 29] rely on depth sensors for ground truth labels, which makes the sensor setup and its calibration more complex, while the obtained ground truth is often sparse. Self-supervised approaches commonly optimize for photometric consistency across stereo image pairs [8, 9], image sequences [10, 12, 26, 28, 30, 40, 47, 49] or both [39, 42]. However, these methods commonly focus on images with

narrow fields of view, which are not sufficient to capture an entire scene. Addressing this limitation, another line of work employs omnidirectional images [33, 34]. However, all the aforementioned setups have no baselines, which do not allow scale-aware self-supervised depth estimation.

Multi-View Depth Estimation Given multiple overlapping images, depth can be inferred through multi-view stereo (MVS) reconstruction. Learning-based MVS methods can be grouped into two families: (i) methods based on the classical concept of photogrammetry, i.e., on the identification of image point correspondences and their triangulation to obtain 3D object points [11, 16, 18, 37, 45, 46]. (ii) pointmap regression methods, which directly predict 3D points, often together with the orientation parameters of the images [22, 35, 36]. Typically, such MVS methods assume a 3D object point to be visible in two or more images, requiring sufficient overlap between the images either during training, inference or both.

In contrast, multi-view surround camera setups provide a 360° field of view by combining multiple cameras, following the central projection model, with minimally overlapping image planes. Consequently, for the majority of pixels, depth needs to be estimated monoscopically. Recent work has studied this camera configuration for depth estimation, using both, images from a single [4, 14, 19, 23, 32, 41, 43, 44] and from multiple time steps [6, 31, 50] during inference. The present work also focuses on this camera configuration, using images from a single time step during inference. FSM [14] is among the earliest self-supervised methods for surround-view depth estimation. It leverages the spatio-temporal context for photometric supervision, exploits overlapping image regions to recover metric scale from a single time step, and introduces a loss to enforce consistency in the temporal pose prediction of the individual cameras. Subsequent work [19, 41] assumes a shared rigid motion of the camera rig and estimate the ego motion instead of the individual camera motion. SurroundDepth [41] proposes attention across images to enhance the consistency of the predicted depth maps. To obtain metric scale, a spatial photometric loss on overlapping images is combined with sparse pseudo-depth labels computed via SfM and filtered for outliers using epipolar geometry-based constraints. In contrast, VFDepth [19] model the depth and pose as volumetric feature representations, i.e., operating in 3D instead of 2D space. However, 3D- and attention-based methods are computationally expensive and do not fully exploit the geometric relationships between images to enforce consistency at inference.

Attention-Based Depth Estimation Initially developed for natural language processing, attention mechanisms are now widely used in vision-based tasks, including monocu-

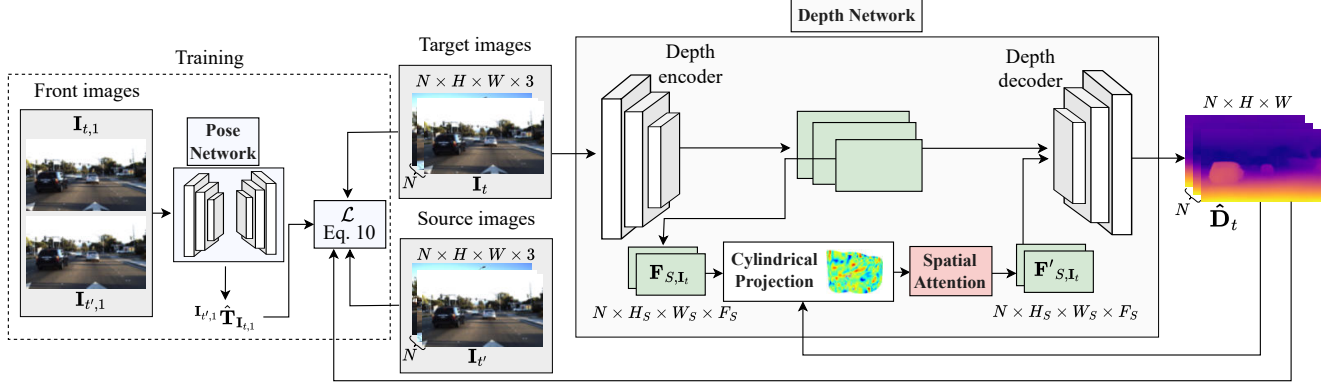


Figure 2. Overview of the proposed network. The depth network takes the target images $\mathbf{I}_t$ as input. The lowest-scale features $\mathbf{F}_{S,\mathbf{I}_t}$ from all target images are projected onto a cylinder, where attention is applied based on cylindrical distances. The pose network takes the source $\mathbf{I}_{t',1}$ and target front $\mathbf{I}_{t,1}$ images as input to predict the temporal pose.

lar [1, 13, 17, 21, 24, 29, 30, 48] and multi-view [27, 32, 37, 41] depth estimation. Early progress was marked by DPT [29], which replaced conventional CNN backbones with Vision Transformers for dense prediction, enabling a global receptive field. Attention can also be used to promote consistency in depth prediction. A work closely related to ours is [30], which employs spatial attention; however it addresses multi-frame monocular depth estimation by aggregating features within each image based on pixel-wise 3D Euclidean distances, relying on estimated depth for the 3D projection, and further adds temporal attention to aggregate features across different time frames to enforce temporal consistency. Different from all previous methods, we introduce a non-learned cross-view spatial attention that fuses features across images by explicitly making use of the geometric relations between the images.

3. Methodology

Given a surround camera setup capturing N time-synchronized images with spatial overlap and known intrinsic parameters and metric relative poses, i.e., known relative orientations and baselines in metric units between the cameras, we aim to estimate a depth map for every image. The depth network employed in our work follows an encoder-decoder architecture (see Fig. 2). In a first forward pass, input images $\mathbf{I}_t \in \mathbb{R}^{N \times H \times W \times 3}$ at time t , with H and W denoting the height and width of the images, respectively, are processed separately by a shared encoder to produce multi-scale feature maps $\mathbf{F}_{S,\mathbf{I}_t} \in \mathbb{R}^{N \times H_s \times W_s \times F_s}$, where $s \in \{1, \dots, S\}$ is the scale, H_s and W_s are the height and width in s , respectively, and F_s is the feature dimension. Passing these feature maps through the decoder, this first forward pass yields a preliminary depth prediction. In a second forward pass, we reuse the encoded feature maps and project their pixel positions onto a shared unit cylinder,

based on the preliminary depth predictions and the known camera parameters. This enables feature aggregation via attention based on the pixels’ geodesic distance on the cylinder to enforce consistent depth predictions across images (see Sec. 3.1). We apply the proposed spatial attention mechanism only at the lowest scale S for efficiency, while using skip connections to preserve high-frequency information. The resulting feature maps are then decoded to predict per-pixel depth $\hat{\mathbf{D}}_t \in \mathbb{R}^{N \times H \times W}$ for each of the N images.

To train our model (see Sec. 3.2), the depth network takes the target frame $\mathbf{I}_t$ and predicts a depth map for each of the N images. The network is supervised based on the spatial photometric consistency between the target images in $\mathbf{I}_t$. However, since the spatial overlap between images in such a setup is typically minimal, we additionally supervise our model temporally. For that, a pose network takes the front view images from the target frame $\mathbf{I}_t$ and from a source frame $\mathbf{I}_{t'}$, where t' is either a past frame $t - 1$ or a future frame $t + 1$, and predicts the transformation of the camera poses between t and t' . This transformation is used to warp the source frame into the target frame, to enforce temporal photometric consistency.

3.1. Multi-View Consistency

In a multi-view setup, processing each image in isolation can yield inconsistent depth predictions across the images, i.e., the same point in 3D object space observed in multiple images may be predicted to be at different 3D locations for each image, since the individual feature representations are unshared and image-specific. To address this issue, we propose an explicit geometry-guided enforcement of multi-view consistency. We first project the pixel coordinates of all individual feature maps onto a shared unit cylinder. This results in cylindrical position maps $\mathbf{O}_{S,\mathbf{I}_t} \in \mathbb{R}^{N \times H_s \times W_s \times 2}$ for each image $\mathbf{I}_{t,i}$, where $i \in N$. Attention between pixels

for a pixel pair u, v from $\mathbf{F}_{S, \mathbf{I}_t}$, and their positions on the cylinder $\mathbf{o}_u$ and $\mathbf{o}_v$ from $\mathbf{O}_{S, \mathbf{I}_t}$, is given as:

$$d_{ij}^2 = (d_{geo}(\mathbf{o}_i, \mathbf{o}_j))^\top \Sigma^{-1} d_{geo}(\mathbf{o}_i, \mathbf{o}_j), \quad (6)$$

$$a_{ij}^{sp} = \begin{cases} \exp(-\frac{1}{2} d_{ij}^2), & d_{ij}^2 \leq \tau^2, \\ 0, & \text{otherwise,} \end{cases} \quad (7)$$

where Σ is a pre-defined non-learned covariance matrix defining the shape and size of the 2D Gaussian kernel, τ is the truncation threshold and d_{geo} is the geodesic distance.

Relying solely on spatial proximity to define attention weights is often suboptimal. This can cause pixels to attend to other pixels that are located nearby on the cylinder and are yet unrelated to each other based on their learned contextual features, e.g., if these pixels show different objects. Therefore, we modulate the spatial attention weight of two pixels u, v with a contextual similarity term defined as the cosine similarity between the feature vectors $\mathbf{f}_u, \mathbf{f}_v$ from $\mathbf{F}_{S, \mathbf{I}_t}$, and denoted as a_{uv}^f . The attention features for a pixel u , for all possible pixels of v , is given as:

$$\mathbf{f}'_u = \sum_v a_{uv} \cdot \mathbf{f}_v, \quad a_{uv} = a_{uv}^{sp} \cdot a_{uv}^f. \quad (8)$$

For all pixels in $\mathbf{F}_{S, \mathbf{I}_t}$, the resulting attention feature map is given as $\mathbf{F}'_{S, \mathbf{I}_t} \in \mathbb{R}^{N \times H_S \times W_S \times F_S}$.

3.2. Self-Supervision

Our method is trained in a self-supervised manner, enforcing photometric consistency between images. The photometric loss [9] compares a target image $\mathbf{I}_{t,i} \in \mathbb{R}^{H \times W \times 3}$ with a warped source image $\hat{\mathbf{I}}_{t,i}$ and is defined as:

$$\mathcal{L}_{photo} = \frac{1}{M} \sum_M \alpha \frac{1 - \text{SSIM}(\hat{\mathbf{I}}_{t,i}, \mathbf{I}_{t,i})}{2} + (1 - \alpha) \left\| \hat{\mathbf{I}}_{t,i} - \mathbf{I}_{t,i} \right\|. \quad (9)$$

where $\alpha = 0.85$, SSIM [38] is the structural similarity, and $M = H \cdot W$ is the number of pixels in the image. The warping can either be done temporally, between images from two consecutive frames, spatially, between different cameras on the rig, or spatio-temporally as a combination of both. These three configurations result in three variants of the photometric loss, described in more detail in the following. Our overall loss is defined as the weighted sum of these photometric loss terms and a set of auxiliary losses:

$$\begin{aligned} \mathcal{L} = & \mathcal{L}_{photo,temp} + \lambda_{sp} \mathcal{L}_{photo,sp} + \lambda_{spt} \mathcal{L}_{photo,spt} \\ & + \lambda_{sm} \mathcal{L}_{sm} + \lambda_{DCCL} \mathcal{L}_{DCCL} + \lambda_{MVRCL} \mathcal{L}_{MVRCL}, \end{aligned} \quad (10)$$

where $\mathcal{L}_{sm}$ is an edge-aware smoothing loss of the depth [9], $\mathcal{L}_{DCCL}$ [4] is a dense depth consistency loss that

enforces consistency of the depth predictions between spatially adjacent images, and $\mathcal{L}_{MVRCL}$ [4] enforces photometric consistency of the spatial and spatio-temporal reconstructions. λ are weighting factors.

Spatial Loss Given the metric relative poses, we make use of the spatial overlap between images from the same frame to obtain a supervision signal based on stereo matching. This enables the network to predict depth that is consistent in scale and given in metric units in the overlapping regions and, due to the propagation of information, also beyond. In this setting, we employ inverse warping: each pixel $\mathbf{p}_{\mathbf{I}_{t,i}}$ in a target image $\mathbf{I}_{t,i}$ is projected into the coordinate system of a spatially adjacent source image $\mathbf{I}_{t,j}$ using the predicted depth $\hat{\mathbf{D}}_{\mathbf{I}_{t,i}}$ and the metric relative pose ${}^{\mathbf{I}_{t,j}}\mathbf{T}_{\mathbf{I}_{t,i}}$ between these images:

$$\hat{\mathbf{p}}_{\mathbf{I}_{t,j}} = \mathbf{K}_{\mathbf{I}_{t,j}} {}^{\mathbf{I}_{t,j}}\mathbf{T}_{\mathbf{I}_{t,i}} \hat{\mathbf{D}}_{\mathbf{I}_{t,i}} \mathbf{K}_{\mathbf{I}_{t,i}}^{-1} \mathbf{p}_{\mathbf{I}_{t,i}}. \quad (11)$$

The spatial loss $\mathcal{L}_{photo,sp}$ is given as the photometric loss (Eq. 9) between a target image and a spatially warped source images.

Temporal Loss Due to the limited spatial overlap between images from the same frame, spatial supervision alone is insufficient for learning accurate depth estimation. To address this limitation, we use temporal context by enforcing photometric consistency between $\mathbf{I}_{t,i}$ and its temporally adjacent source image $\mathbf{I}_{t',i}$, based on a predicted temporal pose ${}^{\mathbf{I}_{t',i}}\hat{\mathbf{T}}_{\mathbf{I}_{t,i}}$. The temporal loss $\mathcal{L}_{photo,temp}$ is given as the photometric loss (Eq. 9) between a target image and a temporally warped source image. To estimate the temporal pose, we assume that all cameras share the same motion, i.e., that they are mounted rigidly to each other. Following [4], we use only the front camera pose to predict the front image temporal pose ${}^{\mathbf{I}_{t',1}}\hat{\mathbf{T}}_{\mathbf{I}_{t,1}}$ using the pose network, ensuring lightweight computations. The temporal pose ${}^{\mathbf{I}_{t',i}}\hat{\mathbf{T}}_{\mathbf{I}_{t,i}} = {}^{\mathbf{I}_{t',1}}\hat{\mathbf{T}}_{\mathbf{I}_{t,1}}^{-1} {}^{\mathbf{I}_{t',1}}\hat{\mathbf{T}}_{\mathbf{I}_{t,1}} {}^{\mathbf{I}_{t,1}}\mathbf{T}_{\mathbf{I}_{t,i}}$ is derived based on the given camera pose w.r.t the front camera ${}^{\mathbf{I}_{t,1}}\mathbf{T}_{\mathbf{I}_{t,i}}$.

Spatio-Temporal Loss Following [14], we employ a spatio-temporal loss, enforcing photometric consistency between images taken by different cameras and at different points in time. This allows us to further increase the number of object points that are seen in more than one image and, thus, to better learn metric scale. The warping follows the same principle as in the previous losses, where a target image $\mathbf{I}_{t,i}$ is warped into the coordinate system of a source image $\mathbf{I}_{t',j}$ based on the relative spatio-temporal pose ${}^{\mathbf{I}_{t',j}}\hat{\mathbf{T}}_{\mathbf{I}_{t,i}} = {}^{\mathbf{I}_{t',j}}\hat{\mathbf{T}}_{\mathbf{I}_{t,j}} {}^{\mathbf{I}_{t,j}}\mathbf{T}_{\mathbf{I}_{t,i}}$. The spatio-temporal loss $\mathcal{L}_{photo,spt}$ is given as the photometric loss (Eq. 9) between a target image and a spatio-temporally warped source images.

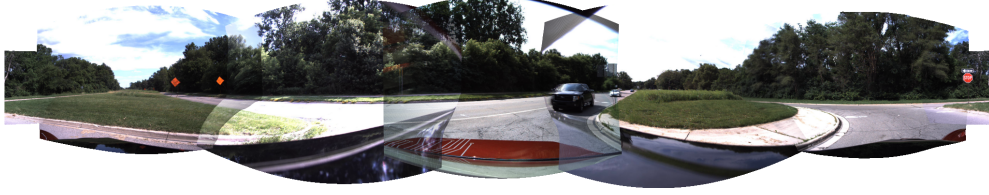


Figure 4. Panoramic visualization of the cylindrical projection of RGB inputs. Note that in our method, only pixel positions are projected, not RGB values. This figure is provided solely for illustration, to show how objects captured from different views are mapped to nearby locations in cylindrical coordinates.

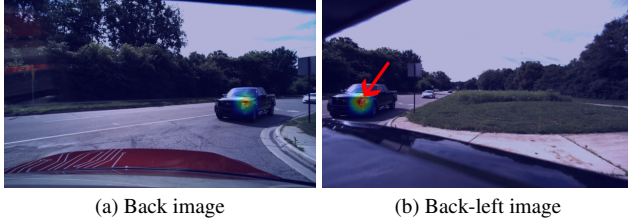


Figure 5. Attention maps for a query token (indicated by the arrow in the back-left image), as overlays on the respective RGB images, showing that this token attends to itself, nearby regions, and to the corresponding region in the spatially adjacent image. High attention is shown in red, low attention in yellow to blue.

Dataset	Method	Abs Rel	Sq Rel [m]	RMSE [m]	$\delta < 1.25$
DDAD	FSM	0.201	-	-	-
	FSM*	0.228	4.409	13.43	68.7
	VFDepth	0.218	3.660	13.32	67.4
	SurroundDepth	0.208	3.371	12.97	69.3
	CVCDepth	0.210	3.458	12.87	70.4
	CylinderDepth (ours)	0.208	3.480	12.85	70.2
nuScenes	FSM	0.297	-	-	-
	FSM*	0.319	7.534	7.860	71.6
	VFDepth	0.289	5.718	7.551	70.9
	SurroundDepth	0.280	4.401	7.467	66.1
	CVCDepth	0.264	5.525	7.178	76.3
	CylinderDepth (ours)	0.238	5.662	6.732	80.5

Table 1. Comparison of our method with state-of-the-art methods. FSM* denotes results reproduced with the implementation of [19]. δ is given in [%]. Abs Rel is unit-free.

4. Experiments

4.1. Experimental Setup

Dataset We train and evaluate our method on DDAD [12] and nuScenes [2]. Both datasets provide images from a six-camera surround rig mounted on a vehicle, capturing 360° of the vehicle’s surrounding, along with LiDAR-derived reference depth. We resize the images to 384×640 pixels for DDAD and 352×640 pixels for nuScenes before providing them as input to our model. Depth is evaluated up to 200 m for DDAD and 80 m for nuScenes, correspond-

Dataset	Method	Abs Rel	Depth Cons. [m]
DDAD	VFDepth (3D)	0.222	4.82
	SurroundDepth (2D)	0.217	7.86
	CVCDepth (2D)	0.212	6.35
	CylinderDepth (ours) (2D)	0.210	5.61
nuScenes	VFDepth (3D)	0.277	3.57
	SurroundDepth (2D)	0.295	6.33
	CVCDepth (2D)	0.388	3.02
	CylinderDepth (ours) (2D)	0.215	2.85

Table 2. Comparison of our method with state-of-the-art 2D and 3D methods in overlapping regions. The best results per category are shown bold. Abs Rel is unit-free.

ing to the range of the ground-truth depth labels. Following [19, 41], we apply self-occlusion masks for DDAD to remove the ego-vehicle from the images during training.

Implementation Details We use a ResNet-18 [15] encoder pre-trained on ImageNet [3] for the depth and pose networks. The decoder in both networks is adopted from [10] and is randomly initialized. Training is conducted on 8 NVIDIA RTX 3060 GPUs with a batch size of 1 (consisting of six surround images) per GPU. We optimize the network using Adam [20] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate is 10^{-4} with a StepLR scheduler decreasing the learning rate by a factor of 0.1 after completing $\frac{3}{4}$ of the total 20 training epochs. For the Gaussian distribution in Eq. 6, we use a covariance matrix $\Sigma = \text{diag}(0.02, 0.02)$, and $\tau = 1.2$. These values are selected based on the feature-map resolution. For the hyperparameter in Eq. 10, we choose $\lambda_{sp} = 0.03$, $\lambda_{spt} = 0.1$, $\lambda_{sm} = 0.1$, $\lambda_{DCCL} = 1 \times 10^{-3}$ and $\lambda_{MVRCL} = 0.2$ based on preliminary experiments.

Evaluation Metrics We adopt standard depth evaluation metrics [5]: Absolute relative difference (Abs Rel), Squared Relative difference (Sq Rel), RMSE, and the percentage of pixels with an error below a threshold δ . In addition, we propose a novel quality metric to assess the multi-view depth consistency (Depth Cons.): First, we identify cor-

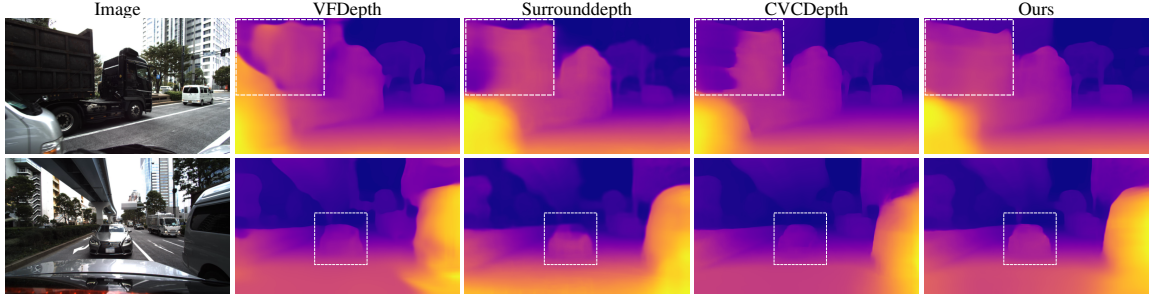


Figure 6. Comparison of depth maps predicted by our method and by state-of-the-art methods on DDAD. Our results show better preserved details and well-defined object boundaries (in white bounding boxes). Depth is shown from close to yellow to distant in blue.

responding pixels in spatially adjacent images using the ground truth depth and the known metric relative pose. We then express the predicted depth of the corresponding pixels in the vehicle coordinate system, i.e., as the Euclidean distance to the origin of a common coordinate system, and compute the RMSE between the two depth predictions.

4.2. Experimental Results

We compare our method against four state-of-the-art methods: FSM [14], SurroundDepth [41], VFDepth [19], and CVCDepth [4]. Since the code of FSM is not publicly available, we report the related results from the original paper and as reproduced in [19]. For CVCDepth, we compare against their ResNet18 version. As shown in Fig. 6 and 7 and Tab. 1 and 2, our approach achieves substantial improvements in multi-view depth consistency, qualitatively and quantitatively, over other 2D-based depth estimation methods [4, 41]. The improvements are especially visible under strong lighting variation between the images (see Fig. 7). This is to be expected, as under these conditions the feature similarity across images is limited, but the geometric guidance provided by our method still applies unconditionally. Our method also achieves slightly higher depth accuracy in both, overlapping regions and full-image evaluations on both datasets. However, it is to be noted that in nuScenes, cameras are synchronized with the LiDAR sweep, leading to clear time differences between images captured by different cameras (up to 40ms). In dynamic scenes and under rig motion, larger deviations from the time-synchronization assumption degrade result quality. This issue affects all methods modeling shared camera motion and relying on spatial supervision, including [4, 14, 19, 41] and ours. VFDepth achieves better results on DDAD compared to ours in multi-view consistency by processing features directly in 3D space. However, this method performs worse than ours under strong illumination differences (see Fig. 7), where severe feature mismatches limit their consistency. Moreover, our method has a considerably smaller memory footprint than VFDepth, as we operate on a two-dimensional cylindrical surface instead of 3D

Method	Train [GB]	Inference [GB]
FSM*	5.6	0.5
VFDepth	11.0	3.3
SurroundDepth	12.6	1.4
CVCDepth	5.4	0.6
CylinderDepth (ours)	8.0	0.7

Table 3. Efficiency comparison of our method against state-of-the-art in terms of peak allocated memory during training and inference. FSM* denotes the implementation from [19].

space (see Tab. 3). Similar is true for SurroundDepth, which relies on multi-head learned attention with attention matrices eight times larger than ours. Yet, SurroundDepth underperforms compared to our non-learned geometry-based attention, since learned attention does not guarantee feature aggregation from the correct tokens across images. CVCDepth faces a similar limitation, as multi-view consistency is only enforced implicitly through its loss functions. In contrast, our method takes advantage of the known camera parameters to project all views into a shared cylindrical representation (see Fig. 4), explicitly ensuring multi-view consistency as illustrated by the attention weight maps in Fig. 5. For more results, refer to the supp. material.

4.3. Ablation Studies

To better assess our contribution and validate its effectiveness, we conduct ablation studies examining the impact of the proposed geometry-guided spatial attention during both training and inference, compare applying the attention only at a low scale versus at all scales, and analyze the role of feature similarity within our spatial attention.

Spatial Attention To evaluate the influence of the proposed spatial attention mechanism, we keep the architecture unchanged and replace our attention weights (cf. Eq. 8) with an identity matrix, i.e., each token attends only to itself. We evaluate two settings: (i) identity-train, where the network is trained with identity attention, and (ii) identity-inference, where a model trained with our spatial attention is tested us-

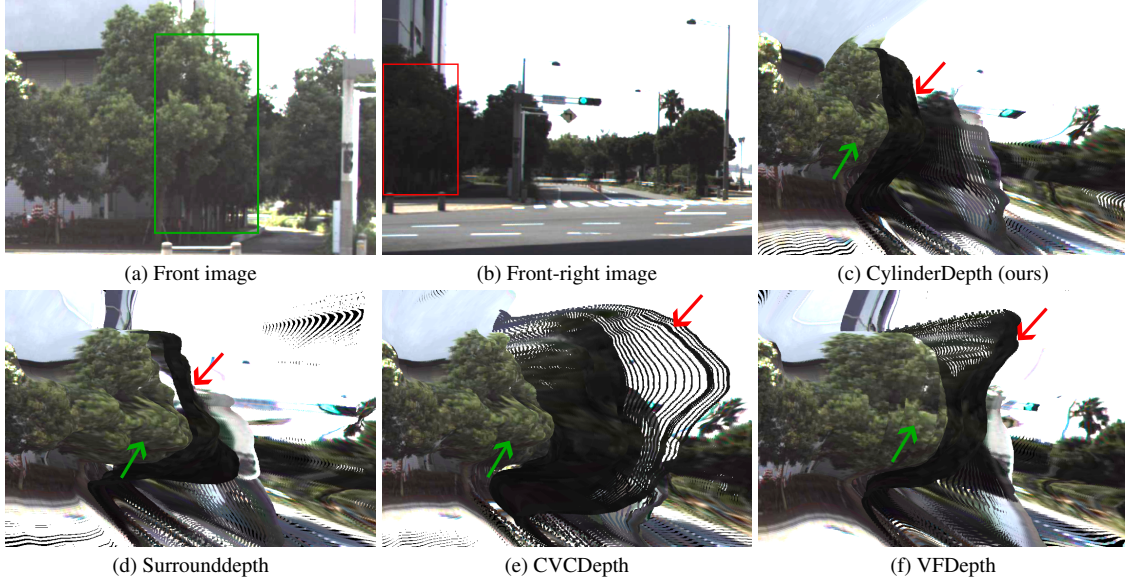


Figure 7. Exemplary 3D reconstructions, comparing our method to the state-of-the-art on DDAD. While our method maps overlapping regions in two images to nearby 3D locations, clear distortions and displacements can be seen for the other methods. The red and green bounding boxes highlight a tree region visible in both images, the green and red arrows indicate the respective regions in the 3D reconstruction.

Method	Overall				Overlap	
	Abs Rel	Sq Rel	RMSE $\delta < 1.25$	Abs Rel	Depth Cons.	
Ours (*)	0.212	3.741	13.21	70.0	0.214	5.59
Ours (**)	0.207	3.503	12.76	70.5	0.207	5.68
Ours (***)	0.208	3.500	12.90	70.2	0.211	6.72
Ours (****)	0.211	3.546	12.90	69.8	0.215	7.04
Ours (*****)	0.208	3.480	12.85	70.2	0.210	5.61

Table 4. Ablation study on our method. (*) applying attention at all scales; (**) geometric attention only (no feature similarity-based weighting); (***) identity attention during training; (****) identity attention during inference with the full model; (*****) our full model. RMSE, Sq Rel and Depth Cons. are given in [m]. δ is given in [%]. Abs Rel is unit-free. Results are reported for the entire images and for overlapping regions only.

ing identity attention. This study isolates the contribution of our spatial attention mechanism and demonstrates the benefit of cross-image feature sharing for multi-view consistency, particularly at inference (see Tab. 4).

Low-Scale Spatial Attention We apply spatial attention only at the coarsest feature scale (cf. Sec. 3), as cross-image attention behaves like a smoothing operator on the feature maps. By restricting attention to the lowest resolution, we enforce global multi-view consistency while preserving fine-scale structures in the higher-resolution features. In contrast, SurroundDepth applies attention at all scales and, for ablation, we do the same. The predictions of this variant

of our method exhibit reduced edge sharpness and appear over-smoothed, with slightly worse overall depth accuracy. Yet, the multi-view consistency does not improve significantly (see Tab. 4 and supp. material).

Geometric and Contextual Attention We combine cylindrical spatial distance with feature similarity, reducing attention between tokens that are geometrically close but feature-wise dissimilar, and reinforcing it when tokens are feature-wise similar (cf. Eq. 8). As shown in Tab. 4, integrating feature-similarity weighting into the geometric spatial attention improves the multi-view consistency compared to a variant with geometric attention only, but slightly reduces overall depth accuracy. We attribute this limitation to the rather narrow feature space of the encoder, requiring further investigation in the future.

5. Conclusion

In this paper, we presented a method for self-supervised surround depth estimation, with a particular focus on enforcing multi-view consistency. Our approach projects pixels from all input images into a shared cylindrical representation, where attention is applied based on their distances on the cylinder. As shown by the results, this enables effective cross-image feature sharing, leading to improvements in multi-view consistency and overall depth accuracy. A limitation of the current design is that attention, due to its high computational cost, is applied only at the lowest feature resolution. While this enforces global consistency,

the coarse scale aggregates large regions and restricts fine-grained detail, leading to suboptimal pixel-level consistency; we aim to address this issue in future work by adapting the distance computations. Moreover, we aim to model the vehicle’s trajectory as a continuous function, instead of discrete time steps, allowing us to effectively account for asynchronously taken images, as in nuScenes [2].

References

- [1] Ashutosh Agarwal and Chetan Arora. Attention attention everywhere: Monocular depth prediction with skip attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5861–5870, 2023. 2, 3
- [2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multi-modal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1, 6, 9
- [3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6
- [4] Laiyan Ding, Hualie Jiang, Jie Li, Yongquan Chen, and Rui Huang. Towards cross-view-consistent self-supervised surround depth estimation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10043–10050. IEEE, 2024. 1, 2, 5, 7
- [5] David Eigen, Christian Puhersch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 2, 6
- [6] Xin Fei, Wenzhao Zheng, Yueqi Duan, Wei Zhan, Masayoshi Tomizuka, Kurt Keutzer, and Jiwen Lu. Driv3r: Learning dense 4d reconstruction for autonomous driving. *arXiv preprint arXiv:2412.06777*, 2024. 2
- [7] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011, 2018. 2
- [8] Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European conference on computer vision*, pages 740–756. Springer, 2016. 2
- [9] Clément Godard, Oisín Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017. 2, 5
- [10] Clément Godard, Oisín Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3828–3838, 2019. 2, 6
- [11] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504, 2020. 2
- [12] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raven-tos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2485–2494, 2020. 1, 2, 6
- [13] Vitor Guizilini, Rares Ambrus, Dian Chen, Sergey Zakharov, and Adrien Gaidon. Multi-frame self-supervised depth with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 160–170, 2022. 3
- [14] Vitor Guizilini, Igor Vasiljevic, Rares Ambrus, Greg Shakhnarovich, and Adrien Gaidon. Full surround monodepth from multiple cameras. *IEEE Robotics and Automation Letters*, 7(2):5397–5404, 2022. 2, 5, 7
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6
- [16] Sunghoon Im, Hae-Gon Jeon, Stephen Lin, and In So Kweon. Dpsnet: End-to-end deep plane sweep stereo. *arXiv preprint arXiv:1905.00538*, 2019. 2
- [17] Adrian Johnston and Gustavo Carneiro. Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4756–4765, 2020. 3
- [18] Tejas Khot, Shubham Agrawal, Shubham Tulsiani, Christoph Mertz, Simon Lucey, and Martial Hebert. Learning unsupervised multi-view stereopsis via robust photometric consistency. *arXiv preprint arXiv:1905.02706*, 2019. 2
- [19] Jung-Hee Kim, Junhwa Hur, Tien Phuoc Nguyen, and Seong-Gyun Jeong. Self-supervised surround-view depth estimation with volumetric feature fusion. *Advances in Neural Information Processing Systems*, 35:4032–4045, 2022. 2, 6, 7
- [20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [21] Sihaeng Lee, Janghyeon Lee, Byungju Kim, Eojindl Yi, and Junmo Kim. Patch-wise attention network for monocular depth estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1873–1881, 2021. 3
- [22] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024. 2
- [23] Ruihang Li, Shanding Ye, Zhe Yin, Tao Li, ZeHua Zhang, KaiKai Xiao, and Zhijie Pan. M2depth: A novel self-supervised multi-camera depth estimation with multi-level supervision. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024. 2

- [24] Zhenyu Li, Zehui Chen, Xianming Liu, and Junjun Jiang. Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. *Machine Intelligence Research*, 20(6):837–854, 2023. 3
- [25] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):2024–2039, 2015. 2
- [26] Jinfeng Liu, Lingtong Kong, Bo Li, Zerong Wang, Hong Gu, and Jinwei Chen. Mono-vifi: A unified learning framework for self-supervised single and multi-frame monocular depth estimation. In *European Conference on Computer Vision*, pages 90–107. Springer, 2024. 2
- [27] Keyang Luo, Tao Guan, Lili Ju, Yuesong Wang, Zhuo Chen, and Yawei Luo. Attention-aware multi-view stereo. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1590–1599, 2020. 3
- [28] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5667–5675, 2018. 2
- [29] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 2, 3
- [30] Patrick Ruhkamp, Daoyi Gao, Hanzhi Chen, Nassir Navab, and Benjamin Busam. Attention meets geometry: Geometry guided spatial-temporal attention for consistent self-supervised monocular depth estimation. In *2021 International Conference on 3D Vision (3DV)*, pages 837–847. IEEE, 2021. 2, 3
- [31] Aron Schmied, Tobias Fischer, Martin Danelljan, Marc Pollefeys, and Fisher Yu. R3d3: Dense 3d reconstruction of dynamic scenes from multiple cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3216–3226, 2023. 2
- [32] Yunxiao Shi, Hong Cai, Amin Ansari, and Fatih 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*, pages 119–129, 2023. 2, 3
- [33] Igor Vasiljevic, Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Wolfram Burgard, Greg Shakhnarovich, and Adrien Gaidon. Neural ray surfaces for self-supervised learning of depth and ego-motion. In *2020 International Conference on 3D Vision (3DV)*, pages 1–11. IEEE, 2020. 2
- [34] Fu-En Wang, Hou-Ning Hu, Hsien-Tzu Cheng, Juan-Ting Lin, Shang-Ta Yang, Meng-Li Shih, Hung-Kuo Chu, and Min Sun. Self-supervised learning of depth and camera motion from 360 videos. In *Asian Conference on Computer Vision*, pages 53–68. Springer, 2018. 2
- [35] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 2
- [36] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 2
- [37] Xiaofeng Wang, Zheng Zhu, Guan Huang, Fangbo Qin, Yun Ye, Yijia He, Xu Chi, and Xingang Wang. Mvster: Epipolar transformer for efficient multi-view stereo. In *European conference on computer vision*, pages 573–591. Springer, 2022. 2, 3
- [38] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5
- [39] Jamie Watson, Michael Firman, Gabriel J Brostow, and Daniyar Turmukhambetov. Self-supervised monocular depth hints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2162–2171, 2019. 2
- [40] Jamie Watson, Oisín Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1164–1174, 2021. 2
- [41] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Yongming Rao, Guan Huang, Jiwen Lu, and Jie Zhou. Surround-depth: Entangling surrounding views for self-supervised multi-camera depth estimation. In *Conference on robot learning*, pages 539–549. PMLR, 2023. 2, 3, 6, 7
- [42] Felix Wimbauer, Nan Yang, Christian Rupprecht, and Daniel Cremers. Behind the scenes: Density fields for single view reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9076–9086, 2023. 2
- [43] Jialei Xu, Xianming Liu, Yuanchao Bai, Junjun Jiang, and Xiangyang Ji. Self-supervised multi-camera collaborative depth prediction with latent diffusion models. *IEEE Transactions on Intelligent Transportation Systems*, 2025. 2
- [44] Yuchen Yang, Xinyi Wang, Dong Li, Lu Tian, Ashish Sirasao, and Xun Yang. Towards scale-aware full surround monodepth with transformers. *arXiv preprint arXiv:2407.10406*, 2024. 2
- [45] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018. 2
- [46] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5525–5534, 2019. 2
- [47] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1983–1992, 2018. 2
- [48] Ilwi Yun, Hyuk-Jae Lee, and Chae Eun Rhee. Improving 360 monocular depth estimation via non-local dense prediction transformer and joint supervised and self-supervised learn-

- ing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3224–3233, 2022. [3](#)
- [49] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017. [2](#)
- [50] Yingshuang Zou, Yikang Ding, Xi Qiu, Haoqian Wang, and Haotian Zhang. M² depth: Self-supervised two-frame multi-camera metric depth estimation. In *European Conference on Computer Vision*, pages 269–285. Springer, 2024. [2](#)