Attention meets Geometry: Geometry Guided Spatial-Temporal Attention for Consistent Self-Supervised Monocular Depth Estimation

Patrick Ruhkamp* [1] Daoyi Gao* [1] Hanzhi Chen* [1] Nassir Navab[1] Benjamin Busam[1]
* Equal contribution. Author ordering determined randomly. [1] Technical University of Munich
{p.ruhkamp,...,b.busam}@tum.de

Abstract

Inferring geometrically consistent dense 3D scenes across a tuple of temporally consecutive images remains challenging for self-supervised monocular depth prediction pipelines. This paper explores how the increasingly popular transformer architecture, together with novel regularized loss formulations, can improve depth consistency while preserving accuracy. We propose a spatial attention module that correlates coarse depth predictions to aggregate local geometric information. A novel temporal attention mechanism further processes the local geometric information in a global context across consecutive images. Additionally, we introduce geometric constraints between frames regularized by photometric cycle consistency. By combining our proposed regularization and the novel spatial-temporal-attention module we fully leverage both the geometric and appearance-based consistency across monocular frames. This yields geometrically meaningful attention and improves temporal depth stability and accuracy compared to previous methods.

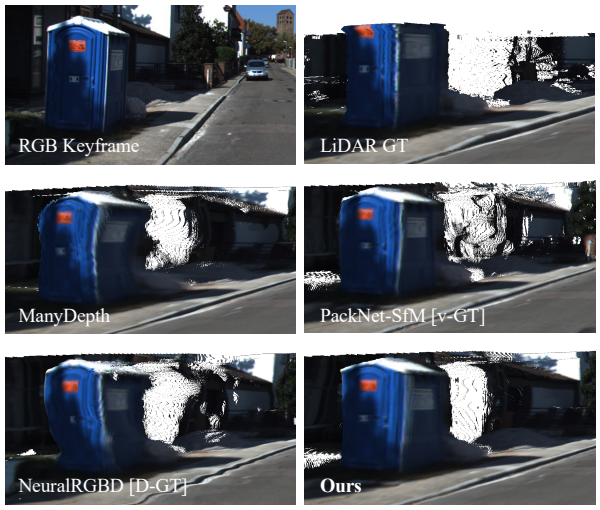

Figure 1: 3D reconstruction from five consecutive depth predictions on Kitti [13]: Our method yields a higher quality reconstruction due to geometrically consistent depth predictions with high accuracy compared against SOTA methods in self-supervised (ManyDepth [43]), semi-supervised (PackNet-SfM [17] with pose velocity [v-GT]), and supervised (NeuralRGBD [28] with depth [D-GT]) methods. Twisted boundaries due to pixel-wise misalignment and "flying pixels" are significantly reduced.

1. Introduction

Improving the accuracy of self-supervised monocular depth prediction has been studied extensively over the past years [15, 5]. However, predicting temporally and geometrically consistent depth over multiple consecutive frames in a self-supervised fashion is mostly unexplored. Consistency is essential for many applications in 3D vision such as reconstruction [35], SLAM [47], pose estimation [2], medical applications [4], AR/MR [30], computational photography [3], or autonomous driving [13].

Consistent Depth Estimation Any downstream tasks may suffer from inconsistent dense depth predictions as for instance, inaccurate 3D object pose estimation in safety-critical applications for autonomous vehicles [19, 18] or RGB-D reconstruction [35]. Geometric consistency has long ago been examined in classical SfM and visual odometry, where usually computationally expensive local and global bundle adjustment aligns sparse triangulated points [33, 34] to account for erroneous initial predictions. Some recent depth prediction pipelines try to enforce consistency [1] using additional ground truth signals such as velocity [17], or take whole sequences into account to train with recurrent units [36].

The evaluation of depth accuracy from monocular self-supervised methods usually employs median scaling against the depth ground truth to account for general scale ambiguity [15]. As this is applied independently per image, the consistency of predictions is completely neglected in such metrics. Hence, the procedure is inadequate to capture the quality of the predictions for real-world scenarios and disregards pixel-wise variations in predictions across multiple

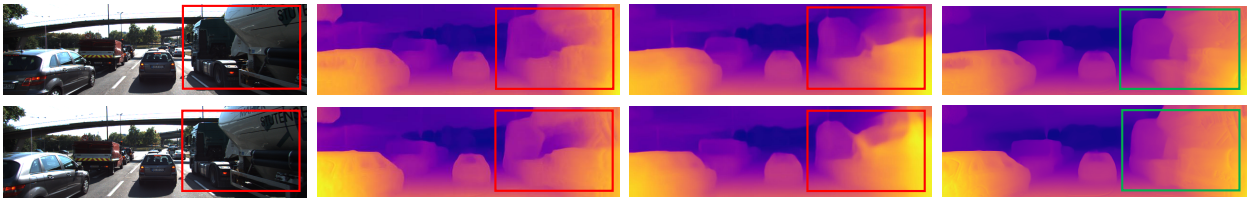

Figure 2: Qualitative depth results: Strong baselines [15, 43] in self-supervised depth prediction suffer from flickering effects between consecutive images. Our method estimates consistent depth across frames, even capable of handling large dynamically moving objects.

frames.

We advocate paying more attention to the temporal consistency of depth predictions by proposing a new metric that allows to quantify this consistency. The qualitative 3D reconstruction [6] from consecutive depth predictions in Fig. 1 gives an impression of the need for consistent depth predictions and justifies our motivation. While the very recently presented ManyDepth [43] currently achieves the best accuracy results (see Table 2), the inconsistent predictions result in noisy reconstructions of the scene. Our model does not only achieve good accuracy metrics but most importantly yields highly consistent depth predictions - even for dynamically moving objects as illustrated in Fig. 2 - and subsequently improves reconstructions of the scene.

Contributions and Key Results Enforcing geometric consistency constraints usually negatively affect depth accuracy towards blurry edges and smooth depth discontinuities in self-supervised methods [1]. Our **t**emporally **c**onsistent **depth** estimation pipeline, short **TC-Depth**[1], enables explicit learning of temporally consistent features for depth prediction in a spatial-temporal attention module, together with geometric regularization, thus achieving high accuracy and unprecedented consistency. An extensive ablation study proves individual contributions on consistency and accuracy, and how our novel geometric constraint with photometric cycle consistency improves the attention mechanism significantly. To this end our contributions are:

1. A novel **spatial attention** formulation which aggregates local geometric information.

2. A **temporal attention** module across tuples of monocular frames which ensures global consistency.

3. A novel **cycle consistency regularization** scheme for our **geometric guidance** of the spatial-temporal attention fusion of feature embeddings.

4. A new **temporal consistency metric** (TCM) to quantify depth consistency across frames.

[1] https://daoyig.github.io/attention_meets_geometry/

2. Related Work

Recent pipelines [9, 25, 10] pioneer the task of supervised monocular depth estimation with convolutional neural networks (CNNs). However, acquiring accurate ground-truth depth data remains difficult, especially for outdoor and large-scale scenes [13]. Scholars [45, 12] propose self-supervised learning approaches with photometric consistency losses by leveraging stereo imagery during training. Monodepth [14] explores left-right consistencies in a fully differentiable pipeline which is also extended to the temporal domain in MonoDepth2 [15], where the complementary prediction of relative camera poses is necessary. While initial joint estimation of depth and pose [55] falls short of accuracy compared to traditional methods, the robustness seems interesting [31, 52]. The use of optical flow [49, 48] greatly improves depth results in particular for moving objects in the scene where forward-backward consistency checks are used to automatically detect occlusions [49, 22, 42].

Attention for Depth Estimation Self-attention mechanisms have shown impressive results in the field of natural language processing [41] and are becoming increasingly popular in computer vision [53, 29]. While a trained set of traditional convolutions is applied independently to an image with fixed kernels during test time, self-attention constitutes a set of operations that adapt to the image and feature input. Huynh et al. [20] propose a depth-attention volume to favour planar scene structures, well suited for indoor environments, while [38] use attention gates in the decoding stage of depth estimation. In [26] patch-wise attention aggregates information of neighbouring features in the scene to predict dense depth in a supervised setting. Also [46] proposed the integration of transformers within a large architecture for highly accurate predictions, but only show applicability in a fully supervised setting. Johnston et al. [23] pioneer the integration of transformers in self-supervised depth prediction for large outdoor scenes. They propose a self-attention mechanism on the feature embedding of input frames after a ResNet encoder and integrate a discrete disparity volume as depth decoder. Despite achieving good accuracy results, the naive self-attention seems

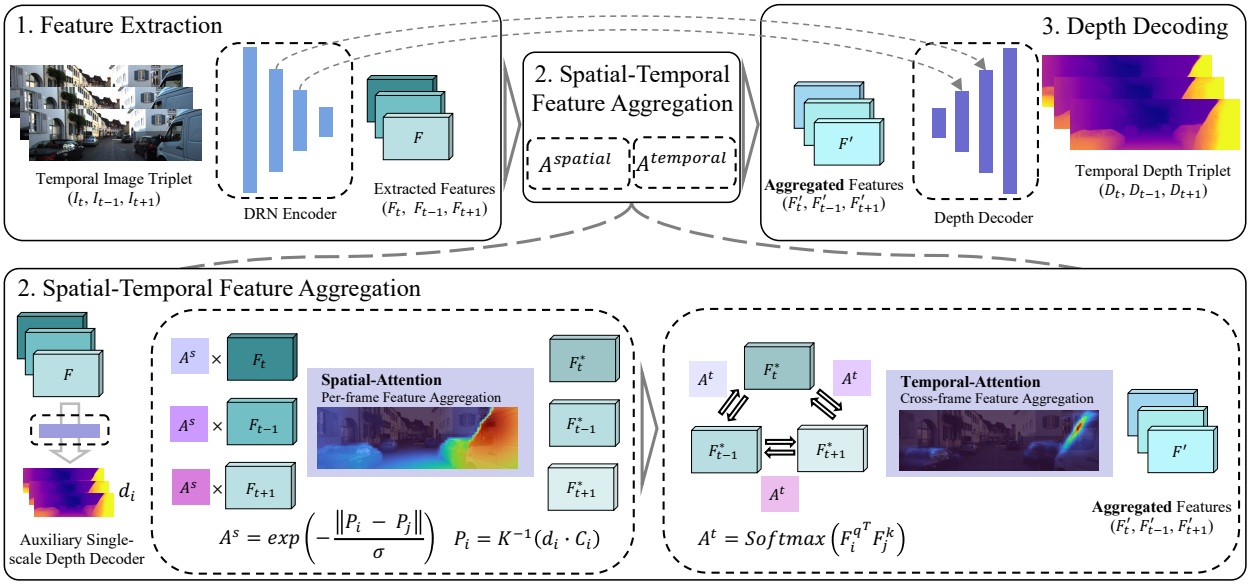

Figure 3: Pipeline Overview: 1. Image features are extracted with a dilated residual network (DRN) 2. An auxiliary low resolution depth map is predicted by a single-stage reference decoder and passed to the spatial attention module for local geometric correlation. The temporal attention aggregates the spatially-aware features globally across frames. 3. Aggregated features are decoded to final depth predictions with skip connections from the encoder.

non-expressive and not capable of aggregating meaningful feature correlation for the task of 3D scene regression.

Consistent Depth Estimation Previous attempts to introduce metrics for depth consistency did not fully capture the geometric and temporal consistency. [52] tried to measure the structural similarity between two consecutive depth maps without aligning them spatially. An optical-flow-based KLT tracker was borrowed in [30] to measure Euclidean distances in 3D between photometrically corresponding points, thus being heavily influenced by the quality of the optical flow estimation.

Until now, self-supervised monocular depth pipelines mainly focus on maintaining a constant overall scale of depth predictions which also affects the auxiliary pose network being more suitable for odometry applications. For this purpose, Bian et al. [1] propose a scale consistent depth and ego-motion approach by adding a depth consistency loss. This leads to reduced scale drift of inferred poses and depth but decreases depth accuracy. Zhao et al. [54] propose a method without direct regression of the 6-DOF camera transformation. They first estimate the optical flow between frames to sample correspondences for relative pose estimation via epipolar geometry. Consistency between triangulated points and the predicted depth ensures scale consistency. MonoRec [44] also focuses on visual odometry applications and achieves impressive results by building a photometric error cost volume to handle static and dynamic elements in the scene in a multi-view stereo setup. However, they employ additional supervision on dense stereo depth predictions and require a complex training scheme. Other works focus on static small-scale indoor scenes. Luo et al. [30] use learning-based priors and test-time training for such scenarios. Their optimization method involves all pixels in a monocular video to achieve highly consistent small-scale reconstructions. NeuralRGBD [28] specifically focuses on consistency, by integrating multiple depth estimates from video sequences in a probability volume, thus aggregating consistent 3D scene information for indoor scene reconstruction in a supervised setting. In order to exploit input image sequences further, Patil et al. [36] use recurrent units to learn more accurate depth predictions utilizing multiple frames in a self-supervised approach. The limitation of this approach, however, is the need for long sequences during both training and test time. ManyDepth [43] proposes to utilize nearby frames of the monocular video sequence during inference time by proposing a cost volume which aggregates the encoded features of multiple frames. This approach is more efficient than previous test time refinement procedures [39] and achieves highly accurate self-supervised depth predictions, yet relative poses between frames need to be predicted as well. In our analysis, however, the improved accuracy and the full utilization of multiple consecutive test frames do not necessarily manifest in temporally consistent depth predictions.

3. Method

The goal of **TC-Depth** is to learn consistent and accurate depth from monocular image sequences in a self-supervised manner. We employ the widely used paradigm of regress-

ing depth and relative camera poses jointly, by minimizing the image reconstruction loss after warping adjacent frames into a common central view via backwards warping with predicted dense depth and pose [15]. We propose the network architecture as illustrated in Fig. 3. For pose regression, we employ the same strategy as previous methods [15, 43] (not illustrated here).

We opt for a feature encoder with dilated convolutions [50] to align resolutions with the attention module in the bottleneck. The DRN-C-26 encoder is similar to a ResNet18 but with dilated strides and additional de-griding layers to remove checkerboard effects [50]. The feature embedding of the encoder is additionally given to an auxiliary single-scale depth decoder [23, 16] which produces a coarse initial depth prediction for the spatial-temporal attention module. The attention mechanism is applied on the coarsest resolution at 24×80 which is $1/8th$ of the input resolution. Inspired by optical flow approaches, the temporal attention takes the encoded input features, together with the spatial attention, to aggregate temporally consistent scene content, before passing through the final depth decoder.

3.1. Attention Module

Convolutional neural networks are constrained by a receptive field which prohibits the correlation of features from spatially distant inputs. Transformers have been proposed in NLP [41] to correlate word items that share some semantic correlation but are far apart in the structure of a sentence. Similar approaches have been applied in computer vision [8] where items are now pixels or some patch of pixels. The inputs for the attention layer are usually named query (Q), key (K), and value (V). Q retrieves information from V based on the attention weight. The attention is defined as:

$$\text{Attention}(Q, K, V) = \mathcal{A}(Q, K)V, \tag{1}$$

where $\mathcal{A}(\cdot)$ is a function that produces a similarity score as attention weight between feature embeddings for aggregation.

Recent works [40] have shown that transformer models with self- and cross-attention can outperform fully convolution networks [27] for the task of finding dense correspondences between image pairs. Inspired by these findings we propose our spatial-temporal attention module.

Spatial-Attention Layer Self-attention as proposed in [23] correlates information within the same image to attend to visually similar parts of the scene. The dot-product in the attention module can introduce some feature aggregation from geometrically distant parts in the 3D scene, which may not be desirable for the task of dense depth regression.

We propose explicit modelling of self-attention with 3D spatial awareness by exploiting a coarse predicted initial depth estimate. Given known camera intrinsics $\mathbf{K}$, a pair of coordinates $\mathbf{C_i} = (u_i, v_i)$ and $\mathbf{C_j} = (u_j, v_j)$, together with their depth d_i and d_j. We first back-project the two pixel coordinates to 3D space:

$$\mathbf{P}_i = \mathbf{K}^{-1}(d_i \cdot \mathbf{C}_i), \quad \mathbf{P}_j = \mathbf{K}^{-1}(d_j \cdot \mathbf{C}_j). \tag{2}$$

Then we formulate the spatial-attention explicitly as:

$$\mathcal{A}_{i,j}^{spatial} = \exp\left(- \frac{\|\mathbf{P}_i - \mathbf{P}_j\|_2}{\sigma} \right), \tag{3}$$

where $\mathbf{P}_i$, $\mathbf{P}_j$ can be treated as key and query, respectively. This can be interpreted as 3D positional encoding via 3D spatial correlation.

Temporal-Attention Layer Inspired by the correlation layer in optical flow [21] and recent dense matching pipelines [40], we formulate a novel temporal attention across frames by exploiting the temporal image sequence input of the self-supervised training scheme.

As a result, given a triplet of feature maps from consecutive image inputs, we can iteratively choose one of them as query and the rest as key features, and then acquire the key-query similarities using Softmax. Here we define $\mathbf{F}_i^q$ as query feature and $\mathbf{F}_j^k$ as key feature, and temporal-attention is formulated as:

$$\mathcal{A}_{i,j}^{temporal} = \text{Softmax}_j(\mathbf{F}_i^{q\top} \mathbf{F}_j^k). \tag{4}$$

Spatial-Temporal Attention The unique formulation of our proposed spatial-temporal attention model can explicitly correlate geometrically meaningful and spatially coherent features - by first passing through the spatial attention - and at the same time provide temporal correlations across subsequent frames. Fig. 4 visualizes the spatial and temporal attention individually for a queried pixel. The spatial attention aggregates geometrically consistent parts of the scene (notice large attention gradients towards the background at object edges). The appearance-based temporal attention correlates global information, which may be difficult and imprecise in a naive approach. With our additional geometric constraints (as later discussed in 3.2 and defined in Eq. 13) the attention is very focused and spatially coherent, as illustrated for two very challenging examples with thin structures and dynamic objects.

3.2. Regularized Geometric Consistency

Scale-invariant Consistent Depth Loss Constraining the absolute depth or disparity values between frames after projecting into the same camera view would either shrink or

enlarge the overall scene depth-scale. Scale-invariant formulations have been proposed [54, 1, 30], but they do not provide strong gradients for depth values that exhibit small alignment errors. Therefore, we adopt a formulation of [24], with additional regularization as detailed in Eq. 13, to constraint depth predictions to be consistent between frames.

Cycle-Mask from Photometric Consistency Aggregating the pixel-wise mean geometric loss over different views violates the scene structure as occluded regions would contribute to the loss computation, resulting in blurry edges and low depth accuracy [1]. The pixel-wise minimum depth error was already proposed to avoid this issue [51, 11]. However, quantitative and qualitative evaluations show that this strategy, while mostly solving the issue of occluded regions, often also excludes major regions of the scene. These can be regions with large inconsistency due to imprecise transformation of adjacent depth maps. The minimum operator can mask out large regions of the scene (see Fig. 5), which harms the training signal. Instead, we propose a novel masking scheme by exploiting the assumption of photo-consistency. For this purpose, the central target image I_t is projectively transformed to the view of the adjacent source frame $I_{t\to s}$ and then transformed back again $I_{t\to s\to t}$. Our cycle-masking can be formulated as:

$$\mathcal{M}_{\text{cycle}} = \left[E_{\text{pe}}(I_t, I_{t\to s\to t}) < \gamma\right], \tag{5}$$

where $[\cdot]$ is the Iverson bracket and E_{pe} is the photometric error as defined in Eq. 9. We set an adaptive threshold γ as the 70% percentile of the photometric error among all pixels of I_s for binarization of $\mathcal{M}_{\text{cycle}}$. With our cycle-masking, we can successfully rule out occluded regions while preserving most of the non-occluded regions for more exhaustive geometric consistency checking as illustrated in Fig. 5.

3.3. Loss Formulation

Our model is trained with a set of loss terms based on content-based image reconstruction and geometric properties of our depth map. It reads:

$$\mathcal{L} = \mathcal{L}_{\text{photo}} + \lambda_{\text{s}}\mathcal{L}_{\text{s}} + \lambda_{\text{geo}}\mathcal{L}_{\text{geo}} + \lambda_{\text{m}}\mathcal{L}_{\text{m}} + \mathcal{L}_{\text{ref}}, \tag{6}$$

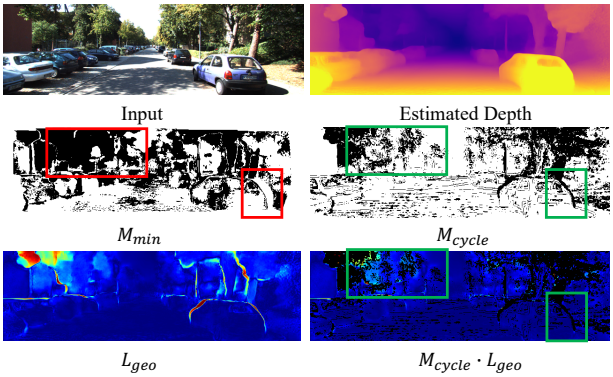

Figure 5: Visualization of occlusion handling for $\mathcal{L}_{\text{geo}}$ with $\mathcal{M}_{\text{cycle}}$ as opposed to pixel-wise minimum: The $\mathcal{M}_{\text{min}}$ cannot robustly account for all occlusions and falsely masks out large image regions. $\mathcal{M}_{\text{cycle}}$ can instead model such cases, resulting in better gradients for training with $\mathcal{M}_{\text{cycle}} \cdot \mathcal{L}_{\text{geo}}$.

where $\mathcal{L}_{\text{photo}}$ and $\mathcal{L}_{\text{s}}$ follow previous established methods [15, 43] and will therefore be only briefly mentioned. We detail all other parts hereafter.

Motion Consistency Loss $\mathcal{L}_{\text{m}}$ Inspired by the knowledge distillation strategy from [37], we train a simplified self-supervised depth prediction network (MonoDepth2 [15] in Table 2) alongside as weak teacher. Following [43], we define a mask where large differences between our prediction D_t and the teacher $\hat{D}_t$ may indicate moving objects, which is also utilized for the photometric loss later, as

$$\mathcal{M}_{\text{m}} = \max\left(\frac{D_t - \hat{D}_t}{\hat{D}_t}, \frac{\hat{D}_t - D_t}{D_t}\right) < 0.6. \tag{7}$$

This yields our motion consistency loss term to help the student to learn from the weak teacher as

$$\mathcal{L}_{\text{m}} = (1 - \mathcal{M}_{\text{m}}) \cdot \|D_t - \hat{D}_t\|_1. \tag{8}$$

Photometric Loss $\mathcal{L}_{\text{photo}}$ The photometric reconstruction error [15, 43] between image I_x and I_y given by

$$E_{\text{pe}}(I_x, I_y) = \alpha\frac{1-\text{SSIM}(I_x, I_y)}{2} + (1-\alpha)\|I_x - I_y\|_1 \tag{9}$$

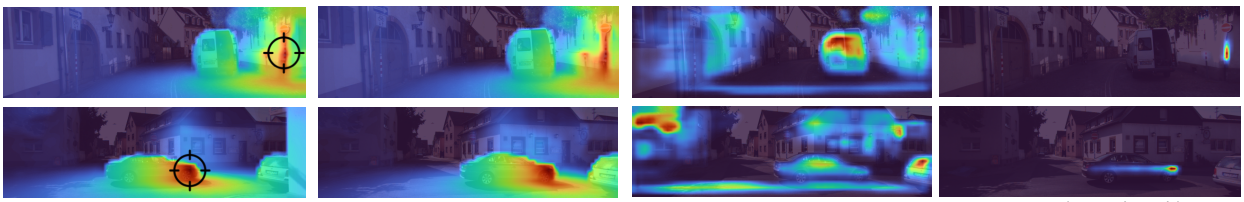

Figure 4: Spatial and temporal attention for a queried pixel (indicated by cross) between frames: The boundary of the spatial attention conforms to the scene structure. The appearance-based naive temporal attention seems unspecific. Our spatially-aware temporal attention focuses on visually similar features with geometric reasoning.

is computed between the target frame I_t and each source frame I_s with $s \in S$ and the pixel-wise minimum error is retrieved. An auto-mask accounts for objects moving with the same velocity and direction as the camera-ego motion

$$\mathcal{M}_{\text{auto}} = \left[\min_{s \in S} E_{\text{pe}}(I_t, I_{s \to t}) < \min_{s \in S} E_{\text{pe}}(I_t, I_s) \right]. \quad (10)$$

$\mathcal{L}_{\text{photo}}$ is finally defined over $S \in \{t-1, t+1\}$ as

$$\mathcal{L}_{\text{photo}} = \mathcal{M}_{\text{m}} \cdot \mathcal{M}_{\text{auto}} \cdot \min_{s \in S} E_{\text{pe}}(I_t, I_{s \to t}). \quad (11)$$

Edge-aware Smoothness Loss $\mathcal{L}_{\text{s}}$ The edge-aware smoothness is applied as in previous works [14, 15] to encourage locally smooth depth estimations with the mean-normalized inverse depth $\overline{d_t}$ as

$$\mathcal{L}_{\text{s}} = \left| \partial_x \overline{d_t} \right| e^{-|\partial_x I_t|} + \left| \partial_y \overline{d_t} \right| e^{-|\partial_y I_t|}. \quad (12)$$

Geometric Loss $\mathcal{L}_{\text{geo}}$ As motivated in Sec. 3.2, we design a geometric loss to encourage consistent depth predictions between frames that not only alleviates the problem of penalizing the scale of the depth prediction, but also utilizes the cycle consistency (Eq. 5) to handle occlusions with

$$\mathcal{L}_{\text{geo}} = \mathcal{M}_{\text{m}} \cdot \mathcal{M}_{\text{auto}} \cdot \mathcal{M}_{\text{cycle}} \cdot \left(1 - \frac{\min(D_{s \to t}, D'_t)}{\max(D_{s \to t}, D'_t)} \right), \quad (13)$$

where the $D_{s \to t}$ is the depth map warped from the adjacent source frame to the target frame and D'_t is the interpolated target depth map [1, 11].

Reference Loss $\mathcal{L}_{\text{ref}}$ To train the single-stage auxiliary depth decoder D_{ref} for spatial attention acquisition, we minimize its difference against the (detached) final depth prediction of our full pipeline D_t:

$$\mathcal{L}_{\text{ref}} = \left\| D_t - D_{ref} \right\|_1. \quad (14)$$

4. Temporal Consistency Metric (TCM)

We propose to measure the consistency directly on the predicted depth output after aligning a number of k frames in 3D via projective transformation, where k is chosen to be in $\{3, 5, 7\}$ (longer sequences usually do not have enough visual overlap for outdoor driving scenes). To transform all predictions from I_s in a common reference frame of I_t, we use the ground-truth depth and pose. Monocular methods (with scale-ambiguity) are first aligned with the same median scaling ratio. Our temporal consistency metric (TCM) measures the track difference between estimated pixel-wise depth and GT across multiple frames. A visual impression of TCM is given in Fig. 6. To account for errors in the interpolated ground-truth LiDAR and moving objects, we filter out 20% of the largest outliers for a fair comparison. In Sec. 5.1 more details on TCM results are given.

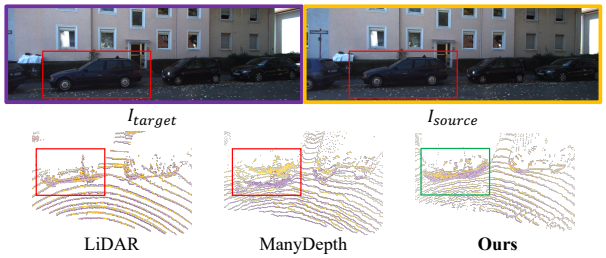

Figure 6: Visualization of TCM: Consecutive depth predictions are aligned in 3D and evaluated pixel-wise across multiple frames. The car shows significantly better alignment between frames for our method.

5. Experiments

We evaluate our model against recent SOTA quantitatively on temporal consistency with our proposed TCM and on well established depth accuracy metrics [15]. We follow previous works on self-supervised depth estimation [15, 43] and conduct extensive experiments on the Eigen split [9] of the Kitti dataset [13] and also report results on Cityscapes [7]. TCM is computed on a test split from the Kitti odometry data (details in Supplementary Material). Cityscapes is neglected for TCM as too many moving objects prevent reliable evaluation. For inference we use an image triplet as indicated in Fig. 3 similar to ManyDepth [43] where consecutive images are used. Different from ManyDepth [43], our method does not need to predict relative poses between adjacent frames for depth inference. In our extensive ablation studies we observe that our model is not majorly influenced by the encoder with dilated convolutions, compared to the standard ResNet as in [15], for consistent depth predictions, likewise for depth accuracy results. Additional qualitative results give an impression of the superiority of our model, especially on temporally consistent 3D reconstructions from consecutive depth predictions (see Figs. 1, 2, 10). Here we focus on the most important findings. For implementation details, and additional quantitative and qualitative results, we refer the interested reader to the supplementary material and video on our project website.

5.1. Depth Consistency

Table 1 summarizes the results on relative TCM for depth consistency over an increasing number of test frames. **TC-Depth** outperforms strong self-supervised baselines such as MonoDepth2 [15], SC-SfMLearner [1] which focuses on temporal consistency, and ManyDepth [43] which specifically utilizes adjacent frames during inference. Our method is also superior to the semi-supervised method of PackNet-SfM [17] and even NeuralRGBD [28] with full supervision and the use of GT poses during testing.

Method	Abs Err			Sq Err			RMSE		
# Test Frames	3	5	7	3	5	7	3	5	7
ManyDepth [43]	0.204	0.260	0.307	0.087	0.147	0.206	0.256	0.319	0.373
MonoDepth2 [15]	0.137	0.177	0.215	0.039	0.068	0.104	0.176	0.223	0.268
SC-SfMLearner [1]	0.126	0.170	0.211	0.032	0.062	0.099	0.159	0.210	0.259
PackNet-SfM [17]	0.141	0.196	0.247	0.044	0.090	0.147	0.177	0.240	0.299
PackNet-SfM [17]*	0.118	0.154	0.190	0.030	0.052	0.083	0.154	0.197	0.240
NeuralRGBD [28]**	0.116	0.148	0.179	0.024	0.044	0.066	0.147	0.186	0.222
Ours DRN-C-26	**0.079**	**0.111**	**0.147**	**0.011**	**0.025**	**0.047**	**0.099**	**0.139**	**0.184**
Ours DRN-D-54	**0.076**	**0.106**	**0.138**	**0.010**	**0.022**	**0.041**	**0.095**	**0.131**	**0.172**

Table 1: Temporal consistency metric (TCM) for increasing number of test frames [3, 5, 7]. *: semi-supervision with velocity. **: supervision with GT depth and inference with GT pose. Our self-supervised model improves TCM about 60% across all metrics compared to strong baselines that leverage temporal frames [43]. It even outperforms semi-supervised [17] and fully supervised [28] pipelines that aim to estimate temporally coherent depth.

5.2. Depth Accuracy

Table 2 shows the depth accuracy results. Our model performs significantly better than comparable self-supervised models such as MonoDepth2 [15] and can also yield better results than models with larger backbones (FeatDepth [39]), models trained with consistency constraints (SC-SfMLearner [1]) or semi-supervised metods (PackNet-SfM [17]). We also adopt the test time refinement scheme (TTR in Table 2) of [32], for which our method actually outperforms ManyDepth [43]. Our method also achieves the best accuracy on the challenging Cityscapes dataset [7].

5.3. Ablation Study

To quantitatively evaluate the influence of each submodule of our pipeline, we perform an extensive ablation study and report TCM results and depth accuracy as before in Table 3. The choice of the backbone (ResNet18 in MD2 [15] against DRN-C-26 in our baseline) has only a marginal effect on accuracy and TCM.

The ablation study reveals that the spatial-temporal attention (ST-A) has a major influence on accuracy, as well as a distinct influence on TCM results. The spatial attention (S-A) improves accuracy but has almost no influence

Method		Abs Rel	Sq Rel	RMSE	$\sigma < 1.25$	$\sigma < 1.25^3$
Monodepth2 [15]		0.115	0.903	4.863	0.877	0.981
SC-SfMLearner [1] †		0.119	0.857	4.950	0.863	0.981
TrianFlow [54]		0.113	**0.704**	4.581	0.871	**0.984**
PackNet-SfM[17]*		0.111	0.829	4.788	0.864	0.980
FeatDepth[39] ‡		0.109	0.923	4.819	0.886	0.981
ManyDepth [43]		**0.098**	0.770	**4.459**	**0.900**	0.983
Ours (DRN-C-26)		0.106	0.770	4.558	0.890	0.983
Ours (DRN-D-54)		0.103	0.746	4.483	0.894	0.983
ManyDepth[43]	TTR	0.090	0.713	4.137	0.914	**0.997**
Ours (DRN-C-26)	TTR	**0.082**	**0.667**	**4.104**	**0.921**	**0.997**
Monodepth2 [15]	CS	0.129	1.569	6.876	0.849	0.983
ManyDepth [43]	CS	0.114	1.193	6.223	**0.875**	0.989
Ours (DRN-C-26)	CS	**0.110**	**0.958**	**5.820**	0.867	**0.991**

Table 2: Accuracy results on Kitti Eigen test split [9] for self-supervised monocular methods (* indicate semi-supervision). Middle: with test time refinement (TTR) [39]. Bottom: Cityscape dataset [7]. †: new results from GitHub; ‡: retrained results with standard image size for fair comparison. We highlight **best**; 2nd best; 3rd best results.

(a) (b) (c) (d)

Figure 7: The baseline model without our contributions (a) suffers from strong ghosting effects due to erroneous pixel-wise misalignment. Constraining with $\mathcal{M}_{\text{cycle}} \cdot \mathcal{L}_{\text{geo}}$ (b) or applying spatial-temporal attention (c) mitigate such issue to a large extent. Our full model (d) yields the result with the highest quality.

on TCM (Tab. 3). While the temporal attention (T-A) alone can neither improve TCM and even harms accuracy, as feature aggregation may be highly noisy and imprecise without positional information [43]. We therefore introduce S-A with correlated 3D information serving as 3D positional encoding to ensure the temporal feature aggregation in T-A is spatially-aware, preventing accuracy degradation.

$\mathcal{L}_{\text{m}}$ enforces consistency between the weak teacher and our prediction where they deviate significantly (i.e. $1 - \mathcal{M}_{\text{m}}$), due to e.g. moving objects. When only $\mathcal{L}_{\text{m}}$ is enforced in the training (w/o geometric guidance), $\mathcal{M}_{\text{m}}$ can already help identify regions with large deviations due to e.g. moving objects, leading to improved accuracy [43]. From the ablations we can see $\mathcal{L}_{\text{geo}}$ is crucial for highly consistent depth predictions. The effectiveness of $\mathcal{L}_{\text{geo}}$ to respect occlusions is guaranteed with $\mathcal{M}_{\text{cycle}}$, whereas potential dynamic objects are masked by $\mathcal{M}_{\text{m}} \cdot \mathcal{M}_{\text{auto}}$. Hence, dynamic objects violating the static assumption of $\mathcal{L}_{\text{geo}}$ are explicitly handled with more exhaustive and accurate masking, thus the consistency performance improves. $\mathcal{L}_{\text{geo}}$ on its own with $\mathcal{M}_{\text{min}}$ actually reduces accuracy slightly for the accuracy measure $\sigma < 1.25$ (which is in accordance with the observations from SC-SfMLearner [1]). The additional cycle mask $\mathcal{M}_{\text{cycle}}$ can mitigate this issue by better accounting for occluded regions based on photometric cues. $\mathcal{L}_{\text{geo}}$ together with $\mathcal{M}_{\text{cycle}}$ also significantly improves TCM. $\mathcal{L}_{\text{m}}$ further reduces the outlier rate as indicated by Sq.Rel. error, as moving objects are handled explicitly.

When spatial-temporal attention is combined with $\mathcal{L}_{\text{geo}}$ and $\mathcal{M}_{\text{cycle}}$, the additional loss function together with appropriate regularization can push the attention module to learn geometrically more consistent aggregation of temporal information by geometric guidance, thus significantly improving on TCM and depth accuracy. The full model achieves the best results, and a larger encoder can improve results further.

Our findings are also conformed by qualitative results in Fig. 7. While the baseline without our contributions shows a deteriorated 3D reconstruction, the proposed geometric constraints and the novel spatial-temporal attention module both improve results individually.

Model	Ablations				Accuracy							TCM (3 Frames)		
	$\mathcal{L}_{geo}$	$\mathcal{M}_{cycle}$	Attention	$\mathcal{L}_{m}$	Abs Rel	Sq Rel	RMSE	RMSE log	$\sigma < 1.25$	$\sigma < 1.25^2$	$\sigma < 1.25^3$	Abs Err	Sq Err	RMSE
MD2 [15]					0.115	0.903	4.863	0.193	0.877	0.959	0.981	0.137	0.039	0.176
DRN-C-26					0.115	1.027	5.004	0.197	0.879	0.958	0.979	0.136	0.039	0.173
	✓				0.113	0.904	4.773	0.193	0.877	0.959	0.980	0.124	0.032	0.157
	✓	✓			0.111	0.878	4.761	0.190	0.882	0.961	0.981	0.113	0.026	0.141
			S-A		0.113	0.958	4.861	0.192	0.882	0.960	0.980	0.134	0.038	0.172
			T-A		0.116	1.028	5.024	0.197	0.879	0.957	0.979	0.133	0.037	0.171
			ST-A		0.112	0.974	4.921	0.194	0.882	0.960	0.980	0.130	0.035	0.165
				✓	0.112	0.840	4.683	0.189	0.880	0.961	0.982	0.132	0.036	0.169
	✓	✓	ST-A		0.108	0.819	4.655	0.186	0.886	0.962	0.982	0.105	0.022	0.133
	✓	✓	ST-A	✓	**0.106**	**0.770**	**4.558**	**0.182**	**0.890**	**0.964**	**0.983**	**0.079**	**0.011**	**0.099**
DRN-D-54	✓	✓	ST-A	✓	0.103	0.746	4.483	0.180	0.894	0.965	0.983	0.076	0.010	0.095

Table 3: Ablation study on depth accuracy and depth consistency (TCM). The consecutive activation of individual pipeline components as indicated all positively influence the overall performance of our method.

5.4. Limitations

Fig. 9 (top) illustrates that the ball-query of the spatial attention can correlate spatially nearby structures. The temporal attention does not always provide one distinct maximum attention for the queried pixel, as multiple non-identical objects of similar appearance can correlate, yielding ambiguous attention (multiple pedestrians, multiple cars). Only objects in a close depth layer are correlated, while other similar distant objects are ignored (e.g. cars in the background). This behavior actually confirms our hypothesis that the spatial attention and geometry constraints guide the temporal attention towards geometry-aware aggregation of consistent features.

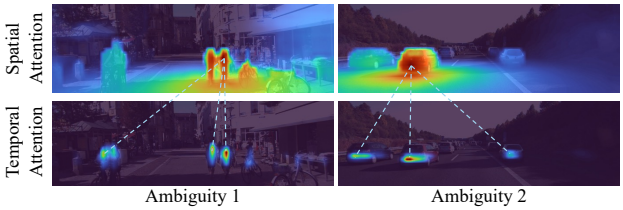

Figure 9: Illustration of spatial and temporal attention for a difficult scene with multiple similar objects.

6. Conclusion

To the best of our knowledge, we have for the first time presented a model that fully leverages the spatial-temporal domain to predict self-supervised consistent depth estimations by introducing a unique and novel attention model based on geometric and appearance-based information. Our method **TC-Depth** has proven that geometric constraints, together with cycle consistency regularization, can further improve such consistency by guiding the spatial-temporal attention aggregation. Future research on temporally consistent depth estimation can now be objectively compared with the new temporal consistency metric (TCM).

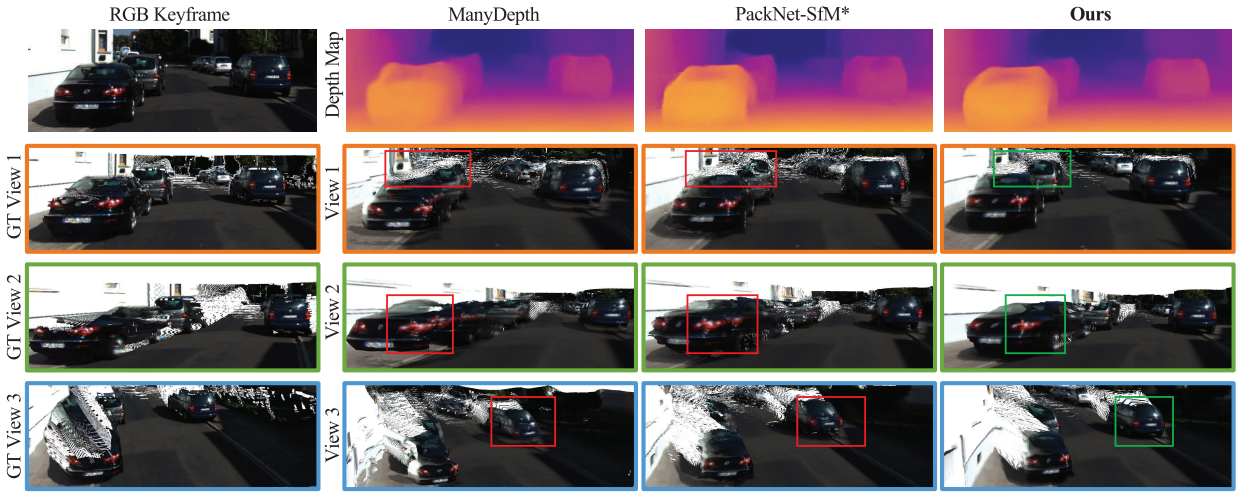

Figure 10: Qualitative reconstruction results from five consecutive depth predictions. Both, ManyDepth [43] and PackNet-SfM* [17] with velocity semi-supervision, suffer from "flying pixels" (View 1), ghosting effects (View 2), and deformed objects (View 3), due to temporal inconsistencies. This is not directly apparent in a single frame depth prediction, but unfold when changing the viewpoint. Our method mitigates these artifacts to a large extend.

References

[1] Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. In *Advances in neural information processing systems*, pages 35–45, 2019.

[2] Benjamin Busam, Tolga Birdal, and Nassir Navab. Camera pose filtering with local regression geodesics on the riemannian manifold of dual quaternions. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2436–2445, 2017.

[3] Benjamin Busam, Matthieu Hog, Steven McDonagh, and Gregory Slabaugh. SteReFo: efficient image refocusing with stereo vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019.

[4] Benjamin Busam, Patrick Ruhkamp, Salvatore Virga, Beatrice Lentes, Julia Rackerseder, Nassir Navab, and Christoph Hennersperger. Markerless inside-out tracking for 3d ultrasound compounding. In *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation*, pages 56–64. Springer, 2018.

[5] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8001–8008, 2019.

[6] Sungjoon Choi, Qian-Yi Zhou, and Vladlen Koltun. Robust reconstruction of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5556–5565, 2015.

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[9] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.

[10] 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.

[11] Feng Gao, Jincheng Yu, Hao Shen, Yu Wang, and Huazhong Yang. Attentional separation-and-aggregation network for self-supervised depth-pose learning in dynamic scenes. *CoRL*, 2020.

[12] 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.

[13] A Geiger, P Lenz, C Stiller, and R Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, Aug 2013.

[14] Clement Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[15] Clément Godard, Oisin 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.

[16] Juan Luis GonzalezBello and Munchurl Kim. Forget about the lidar: Self-supervised depth estimators with med probability volumes. *Advances in Neural Information Processing Systems*, 33:12626–12637, 2020.

[17] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, 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.

[18] Hou-Ning Hu, Qi-Zhi Cai, Dequan Wang, Ji Lin, Min Sun, Philipp Krähenbühl, Trevor Darrell, and Fisher Yu. Joint monocular 3d vehicle detection and tracking. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[19] Hou-Ning Hu, Yung-Hsu Yang, Tobias Fischer, Fisher Yu, Trevor Darrell, and Min Sun. Monocular quasi-dense 3d object tracking. *ArXiv:2103.07351*, 2021.

[20] Lam Huynh, Phong Nguyen-Ha, Jiri Matas, Esa Rahtu, and Janne Heikkilä. Guiding monocular depth estimation using depth-attention volume. In *European Conference on Computer Vision*, pages 581–597. Springer, 2020.

[21] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[22] Joel Janai, Fatma Güney, Anurag Ranjan, Michael J. Black, and Andreas Geiger. Unsupervised learning of multi-frame optical flow with occlusions. In *European Conference on Computer Vision (ECCV)*, volume Lecture Notes in Computer Science, vol 11220, pages 713–731. Springer, Cham, Sept. 2018.

[23] 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.

[24] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1611–1621, 2021.

[25] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016.

[26] 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*, volume 35, pages 1873–1881, 2021.

[27] Xinghui Li, Kai Han, Shuda Li, and Victor Prisacariu. Dual-resolution correspondence networks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[28] Chao Liu, Jinwei Gu, Kihwan Kim, Srinivasa G Narasimhan, and Jan Kautz. Neural RGB-D sensing: Depth and uncertainty from a video camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10986–10995, 2019.

[29] 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.

[30] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)*, 39(4):71–1, 2020.

[31] V. Madhu Babu, K. Das, A. Majumdar, and S. Kumar. UnDEMoN: Unsupervised deep network for depth and ego-motion estimation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1082–1088, Oct 2018.

[32] Robert McCraith, Lukas Neumann, Andrew Zisserman, and Andrea Vedaldi. Monocular depth estimation with self-supervised instance adaptation. *arXiv preprint arXiv:2004.05821*, 2020.

[33] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. ORB-SLAM: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.

[34] Raul Mur-Artal and Juan D Tardós. ORB-SLAM2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017.

[35] Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, pages 127–136, 2011.

[36] Vaishakh Patil, Wouter Van Gansbeke, Dengxin Dai, and Luc Van Gool. Don't forget the past: Recurrent depth estimation from monocular video. *IEEE Robotics and Automation Letters*, 5(4):6813–6820, 2020.

[37] Andrea Pilzer, Stephane Lathuiliere, Nicu Sebe, and Elisa Ricci. Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9768–9777, 2019.

[38] Assem Sadek and Boris Chidlovskii. Self-supervised attention learning for depth and ego-motion estimation. *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10054–10060, 2020.

[39] Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-metric loss for self-supervised learning of depth and egomotion. In *ECCV*, 2020.

[40] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. *CVPR*, 2021.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 2017.

[42] Yang Wang, Yi Yang, Zhenheng Yang, Liang Zhao, Peng Wang, and Wei Xu. Occlusion aware unsupervised learning of optical flow. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018.

[43] Jamie Watson, Oisin 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.

[44] Felix Wimbauer, Nan Yang, Lukas von Stumberg, Niclas Zeller, and Daniel Cremers. MonoRec: Semi-supervised dense reconstruction in dynamic environments from a single moving camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[45] Junyuan Xie, Ross Girshick, and Ali Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *European Conference on Computer Vision*, pages 842–857. Springer, 2016.

[46] Guanglei Yang, Hao Tang, Mingli Ding, Nicu Sebe, and Elisa Ricci. Transformers solve the limited receptive field for monocular depth prediction. *arXiv preprint arXiv:2103.12091*, 2021.

[47] Nan Yang, Rui Wang, Jörg Stückler, and Daniel Cremers. Deep Virtual Stereo Odometry: Leveraging deep depth prediction for monocular direct sparse odometry. *Lecture Notes in Computer Science*, page 835–852, 2018.

[48] Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, and Ram Nevatia. Every Pixel Counts: Unsupervised geometry learning with holistic 3d motion understanding. *Computer Vision – ECCV 2018 Workshops*, page 691–709, 2019.

[49] Zhichao Yin and Jianping Shi. GeoNet: Unsupervised learning of dense depth, optical flow and camera pose. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018.

[50] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 472–480, 2017.

[51] H. Zhan, C. S. Weerasekera, J. W. Bian, and I. Reid. Visual odometry revisited: What should be learnt? In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4203–4210, 2020.

[52] Haokui Zhang, Chunhua Shen, Ying Li, Yuanzhouhan Cao, Yu Liu, and Youliang Yan. Exploiting temporal consistency for real-time video depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1725–1734, 2019.

[53] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10076–10085, 2020.

[54] Wang Zhao, Shaohui Liu, Yezhi Shu, and Yong-Jin Liu. Towards better generalization: Joint depth-pose learning without posenet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9151–9161, 2020.

[55] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017.