Spatial-Temporal Attention through Self-Supervised Geometric Guidance

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

Abstract

This paper explores how the increasingly popular transformer architecture can be guided towards spatially consistent aggregation of features, by exploiting geometric constraints in a self-supervised depth-from-mono learning scheme. We propose a novel spatial-temporal attention mechanism comprising of: 1) a spatial attention module that correlates coarse depth predictions to aggregate local geometric information; 2) temporal attention to further process 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.

1. Introduction

Improving the accuracy of self-supervised monocular depth prediction has been studied extensively over the past years [9] and is essential for many applications in 3D vision such as 3D reconstruction [22], SLAM [30], pose estimation [2], medical applications [4], AR/MR [20], computational photography [3], or autonomous driving [8]. Recent approaches try to leverage transformers to improve depth accuracy [14], but the attentive feature aggregation does not integrate geometric information. 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 can provide temporal aggregations across consecutive frames. Compared to previous methods [1], our geometric constraints do not negatively effect depth accuracy results and yield focused and accurate attention between frames. Fig. 1 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 (Fig. 1 center). With our additional geometric constraints the attention is very focused and spatially coherent, as illustrated for two very challenging examples with thin structures and dynamic objects in Fig. 1 (right).

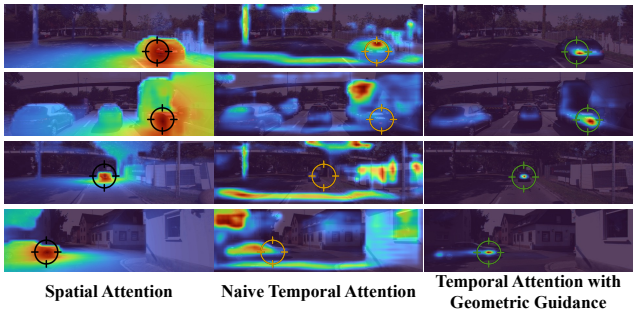

Figure 1: Qualitative visualization of our proposed attention mechanism with self-supervised geometric guidance compared against naive attention.

2. Related Work

Recent works on self-supervised depth estimation from monocular video sequences try to leverage the sequential input images [9] to predict more temporally consistent dense depth. 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. ManyDepth [28] 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 [26] and achieves highly accurate self-supervised depth predictions, but relative poses between frames need to be predicted as well. Mentioned methods however, do not employ the strong capabilities of transformer models, yet.

Self-attention mechanisms are becoming increasingly

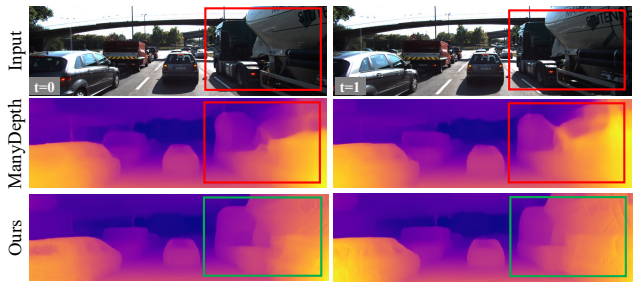

Figure 2: Qualitative depth results: ManyDepth [28] suffers from flickering effects between consecutive images. Our method estimates consistent depth across frames, even capable of handling large dynamic objects.

popular in computer vision [33, 19]. 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. In this regard, Huynh et al. [12] propose a depth-attention volume to favour planar scene structures, well suited for indoor environments, while [25] use attention gates in the decoding stage of depth estimation. In [17] patch-wise attention aggregates information of neighbouring features in the scene to predict dense depth in a supervised setting. Also [29] 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. [14] pioneer the integration of transformers in self-supervised depth prediction for large outdoor scenes, by proposing 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 good accuracy results, the naive self-attention seems non-expressive and incapable of aggregating meaningful features for the task of 3D scene regression.

3. Method

The goal of the proposed method is to integrate a lightweight attention mechanism, for improved depth prediction (compare Fig. 2), which aggregates local spatial and temporally consistent information while training in a self-supervised setting with geometric guidance. We employ the widely used paradigm of regressing 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 [9]. We propose the network architecture as illustrated in Fig. 3. For pose regression, we employ the same strategy as previous methods [9, 28] (not illustrated here).

We opt for a feature encoder with dilated convolutions [31] 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

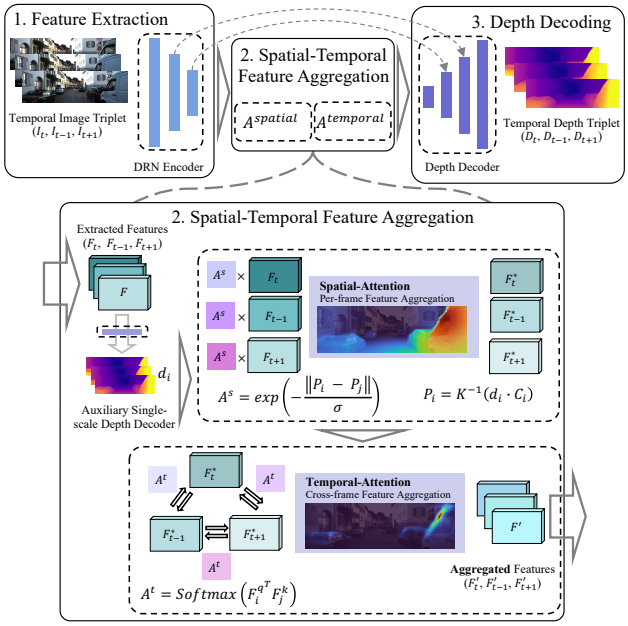

Figure 3: Pipeline Overview: 1. Image features are extracted with a dilated residual network (DRN) 2. An auxiliary low resolution depth map is predicted by a single-stage reference decoder and passed to the spatial attention module for local geometric correlation. The temporal attention aggregates the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

layers to remove checkerboard effects [31]. The feature embedding of the encoder is additionally given to an auxiliary single-scale depth decoder [10, 14] 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 px, 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

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:

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

where $\mathcal{A}(\cdot)$ is a function that produces a similarity score as attention weight between feature embeddings for aggregation. Recent works [27] have shown that transformer models with self- and cross-attention can outperform fully convolution networks [18] 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 [14] 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 [13] and recent dense matching pipelines [27], 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:

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

3.2. 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:

$$\mathcal{L} = \mathcal{L}_{\text{photo}} + \lambda_{\text{s}}\mathcal{L}_{\text{s}} + \lambda_{\text{m}}\mathcal{L}_{\text{m}} + \lambda_{\text{geo}}\mathcal{L}_{\text{geo}} + \mathcal{L}_{\text{ref}}, \tag{5}$$

where the photometric error $\mathcal{L}_{\text{photo}}$, the smoothness loss $\mathcal{L}_{\text{s}}$ and the auto-masking for stationary objects $\mathcal{M}_{\text{auto}}$ follow previous established methods [9, 28] and are therefore not detailed here. We detail all other parts hereafter.

Motion Consistency Loss $\mathcal{L}_{\text{m}}$. Inspired by the knowledge distillation strategy from [24], we train a simplified self-supervised depth prediction network (MonoDepth2 [9] in Table 1) alongside as weak teacher. Following [28], we define a mask where large differences between our prediction D_t and the teacher $\hat{D}_t$ may indicate moving objects as

$$\mathcal{M}_{\text{m}} = \max\left(\frac{D_t - \hat{D}_t}{\hat{D}_t}, \frac{\hat{D}_t - D_t}{D_t} \right) < \tau. \tag{6}$$

This yields our motion consistency term with $\tau = 0.6$

$$\mathcal{L}_{\text{m}} = (1 - \mathcal{M}_{\text{m}}) \cdot \|D_t - \hat{D}_t\|_1. \tag{7}$$

Regularized Geometric 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 [7, 32]. 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. 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:

$$\mathcal{M}_{\text{cycle}} = \left[E_{\text{pe}}(I_t, I_{t\to s\to t}) < \gamma \right], \tag{8}$$

where $[\cdot]$ is the Iverson bracket and E_{pe} is the photometric error [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.

Geometric Loss $\mathcal{L}_{\text{geo}}$. Inspired by the ratio loss proposed by [16], we design a geometric loss that not only alleviates the problem of penalizing the scale of the depth prediction, but also utilizes the cycle consistency (Eq. 8) to handle occlusions with

$$\mathcal{L}_{\text{geo}} = \mathcal{M}_{\text{m}} \cdot \mathcal{M}_{\text{auto}} \cdot \mathcal{M}_{\text{cycle}} \cdot \left(1 - \frac{\min(D_{s\to t}, D'_t)}{\max(D_{s\to t}, D'_t)} \right), \tag{9}$$

where $D_{s\to t}$ is the depth map warped from the adjacent source frame to the target frame and D'_t is the interpolated target depth map [1, 7].

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

$$\mathcal{L}_{\text{ref}} = \|D_t - d_t\|_1 . \tag{10}$$

4. Experiments

We evaluate our model against recent SOTA quantitatively on well established depth accuracy metrics [9]. We follow previous works on self-supervised depth estimation [9, 28] and conduct extensive experiments on the Eigen split [6] of the Kitti dataset [8] and also report results on Cityscapes [5]. For inference we use an image triplet as

Method		Abs Rel	Sq Rel	$\sigma < 1.25$	$\sigma < 1.25^3$
Monodepth2 [9]		0.115	0.903	0.877	0.981
SC-SfMLearner [1] †		0.119	0.857	0.863	0.981
TrianFlow [34]		0.113	**0.704**	0.871	**0.984**
PackNet-SfM[11]*		0.111	0.829	0.864	0.980
FeatDepth[26] ‡		0.109	0.923	0.886	0.981
ManyDepth [28]		**0.098**	0.770	**0.900**	0.983
Ours (DRN-C-26)		0.106	0.770	0.890	0.983
Ours (DRN-D-54)		0.103	0.746	0.894	0.983
ManyDepth[28]	TTR	0.090	0.713	0.914	**0.997**
Ours (DRN-C-26)	TTR	**0.082**	**0.667**	**0.921**	**0.997**
ManyDepth [28]	S	0.117	0.886	0.872	0.982
Ours (DRN-C-26)	S	0.107	0.784	0.888	**0.983**
Ours (DRN-D-54)	S	**0.104**	**0.760**	**0.982**	**0.983**
Monodepth2 [9]	CS	0.129	1.569	0.849	0.983
ManyDepth [28]	CS	0.114	1.193	**0.875**	0.989
Ours (DRN-C-26)	CS	**0.110**	**0.958**	0.867	**0.991**

Table 1: Accuracy results. Top: Kitti [8] Eigen test split [6]. *: semi-supervision. TTR: With test time refinement [26]. S: Static camera simulation. Bottom: CS: Cityscape dataset [5]. †: new results from GitHub; ‡: retrained results with standard image size for fair comparison. We highlight **best**; 2nd best; 3rd best results.

indicated in Fig. 3, similar to ManyDepth [28] where consecutive images are used as well. Different from [28], our method does not need to predict relative poses between adjacent frames for depth inference.

Depth Accuracy. Table 1 summarizes the depth accuracy results. Our model performs significantly better than comparable self-supervised models (MonoDepth2 [9]), and yields better results than models with larger backbones (FeatDepth [26]), models trained with consistency constraints (SC-SfMLearner [1]) or semi-supervised methods (PackNet-SfM [11]). We also adopt the test time refinement scheme (TTR in Table 1) of [21], for which our method outperforms ManyDepth [28]. Our method also achieves the best accuracy on the challenging Cityscapes dataset [5].

To simulate the scenario of a static camera, where no consecutive images with changing scene structure are provided, we input only a single static image to our method, and compare against ManyDepth [28] which also utilizes consecutive input frames. Despite slightly inferior results for our method with single static frame input compared to temporal images, we do not observe such strong deterioration in accuracy as for ManyDepth [28].

Ablation Study. To quantitatively evaluate the influence of each sub-module, we perform an extensive ablation study and report depth accuracy as before in Table 2. The choice of the backbone (ResNet18 in MD2 [9] against DRN-C-26 in our baseline) has only a marginal effect.

The ablation study reveals that the spatial-temporal attention has a major influence on accuracy. $\mathcal{L}_{\text{geo}}$ 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. $\mathcal{L}_{\text{m}}$ reduces the outlier

Method	$\mathcal{L}_{\text{geo}}$	$\mathcal{M}_{\text{cycle}}$	Attention	$\mathcal{L}_{\text{m}}$	Abs Rel	Sq Rel	$\sigma < 1.25$	$\sigma < 1.25^3$
MD2 [9]					0.115	0.903	0.877	0.981
DRN-C-26					0.115	1.027	0.879	0.979
DRN-C-26	✓				0.113	0.904	0.877	0.980
DRN-C-26	✓	✓			0.111	0.878	0.882	0.981
DRN-C-26			✓		0.112	0.974	0.882	0.980
DRN-C-26				✓	0.112	0.840	0.880	0.982
DRN-C-26	✓	✓	✓		0.108	0.819	0.886	0.982
DRN-C-26	✓	✓	✓	✓	**0.106**	**0.770**	**0.890**	**0.983**
DRN-D-54	✓	✓	✓	✓	0.103	0.746	0.894	0.983

Table 2: Ablation study on depth accuracy for Kitti [8] Eigen test split [6].

rate as indicated by the 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 guide the attention module to learn geometrically more consistent aggregation of temporal information, thus significantly improving depth accuracy. The full model achieves the best results, and a larger encoder can improve results further.

Attention Ambiguities. Fig. 9 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 a queried pixel, as multiple non-identical objects of similar appearance may show high correlation, hence yielding ambiguous attention (multiple pedestrians or cars). Note, only objects in a close depth layer are correlated, while other similar distant objects are ignored (e.g. cars in the background). This behavior confirms our hypothesis that the spatial attention and geometric constraints guide the temporal attention towards geometry-aware aggregation of consistent features.

5. Conclusion

Our model fully leverages the spatial-temporal domain to predict self-supervised consistent depth estimations by introducing a unique and novel attention model, based on spatial and appearance-based information, with geometric guidance. Our method has proven that geometric constraints, together with cycle consistency regularization, can guide the spatial-temporal attention aggregation towards focused and distinct feature aggregation for the task of self-supervised depth (from monocular images) prediction.

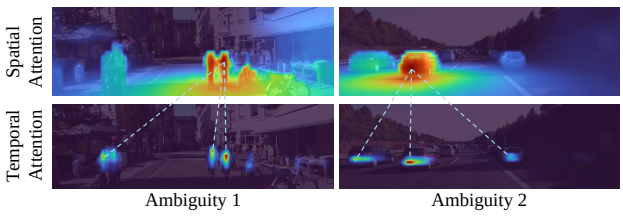

Figure 9: Attention ambiguities: Illustration of spatial and temporal attention for difficult scenes with multiple related objects at similar distance.

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. 1, 3, 4

[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. 1

[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. 1

[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. 1

[5] 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. 3, 4

[6] 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. 3, 4

[7] 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. 3

[8] 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. 1, 3, 4

[9] 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. 1, 2, 3, 4, 7

[10] 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. 2

[11] 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. 4, 7, 8, 9

[12] 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. 2

[13] 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. 3

[14] 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. 1, 2, 3

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[16] 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. 3

[17] 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. 2

[18] Xinghui Li, Kai Han, Shuda Li, and Victor Prisacariu. Dual-resolution correspondence networks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 2

[19] 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. 2

[20] 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. 1

[21] Robert McCraith, Lukas Neumann, Andrew Zisserman, and Andrea Vedaldi. Monocular depth estimation with self-supervised instance adaptation. *arXiv preprint arXiv:2004.05821*, 2020. 4

[22] 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. 1

[23] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NIPS-W*, 2017. 7

[24] 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. 3

[25] 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. 2

[26] Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-metric loss for self-supervised learning of depth and egomotion. In *ECCV*, 2020. 1, 4

[27] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. *CVPR*, 2021. 2, 3

[28] 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. 1, 2, 3, 4, 7, 8, 9

[29] 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. 2

[30] 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. 1

[31] 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. 2, 7

[32] 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. 3

[33] 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. 2

[34] 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. 4

A. Appendix

A.1. Qualitative Results

Figs. A1, A2 and A3 show more qualitative 3D reconstruction results. The importance of temporally consistent depth predictions is apparent in such reconstructions. A single depth map cannot capture inconsistencies, but observing a reconstruction of fused depth maps from different view points can intuitively demonstrate such effects. In the examples illustrated here, the strong baseline ManyDepth [28] - despite achieving state-of-the-art results in accuracy - suffers from deformed objects, ghosting effects, and "flying pixels". Similar artifacts are observed for the semi-supervised PackNet-SfM* [11]. Our method yields the most consistent reconstructions from consecutive depth maps.

A.2. Implementation Details

We implement our model in PyTorch [23] and train for 25 epochs using Adam [15] with a batch size of 6 for our full DRN-C-26 [31] model, trained on one NVIDIA RTX-3090 GPU. We choose an initial learning rate of 1×10^{-4} for 15 epochs, which we decrease to 2.5×10^{-5} for 5 epochs, and 6.25×10^{-6} for the last 5 epochs. We perform the same augmentations as [9]. We set $\lambda_{\text{geo}} = 0.1$ and $\lambda_{\text{s}} = 10^{-3}$. $\lambda_{\text{m}} = 1.0$ for the first 20 epochs, after which $\lambda_{\text{m}} = 0.0$ to allow our network better finetuning.

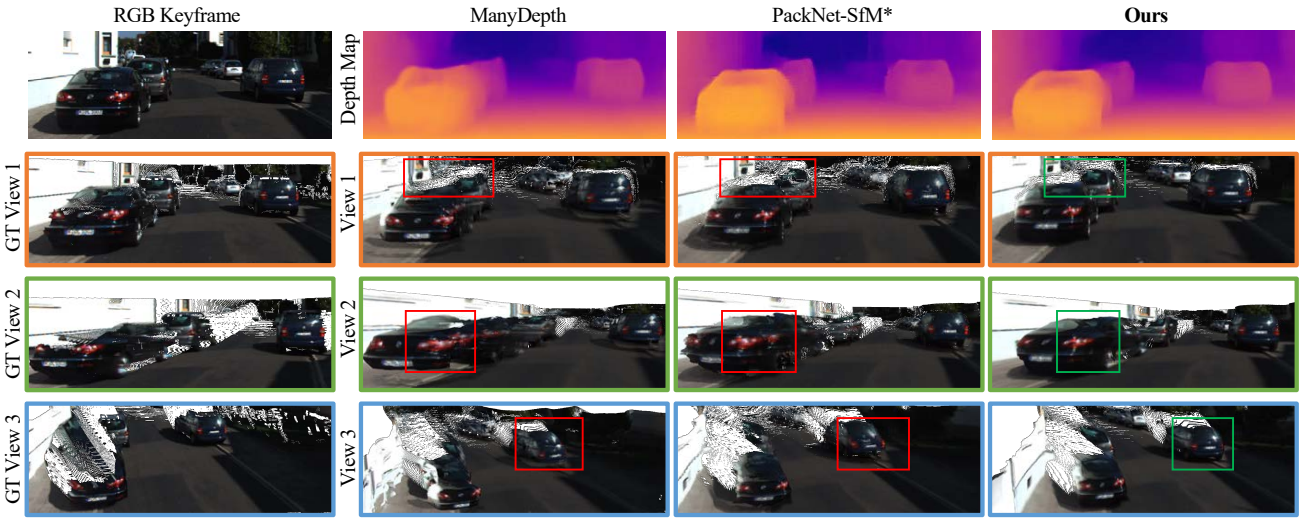

Figure A1: Qualitative reconstruction results from five consecutive depth predictions. Both, ManyDepth [28] and PackNet-SfM* [11] with velocity semi-supervision, suffer from "flying pixels" (View 1), ghosting effects (View 2), and deformed objects (View 3), due to temporal inconsistencies. This is not directly apparent in a single frame depth prediction, but unfold when changing the viewpoint. Our method mitigates these artifacts to a large extent.

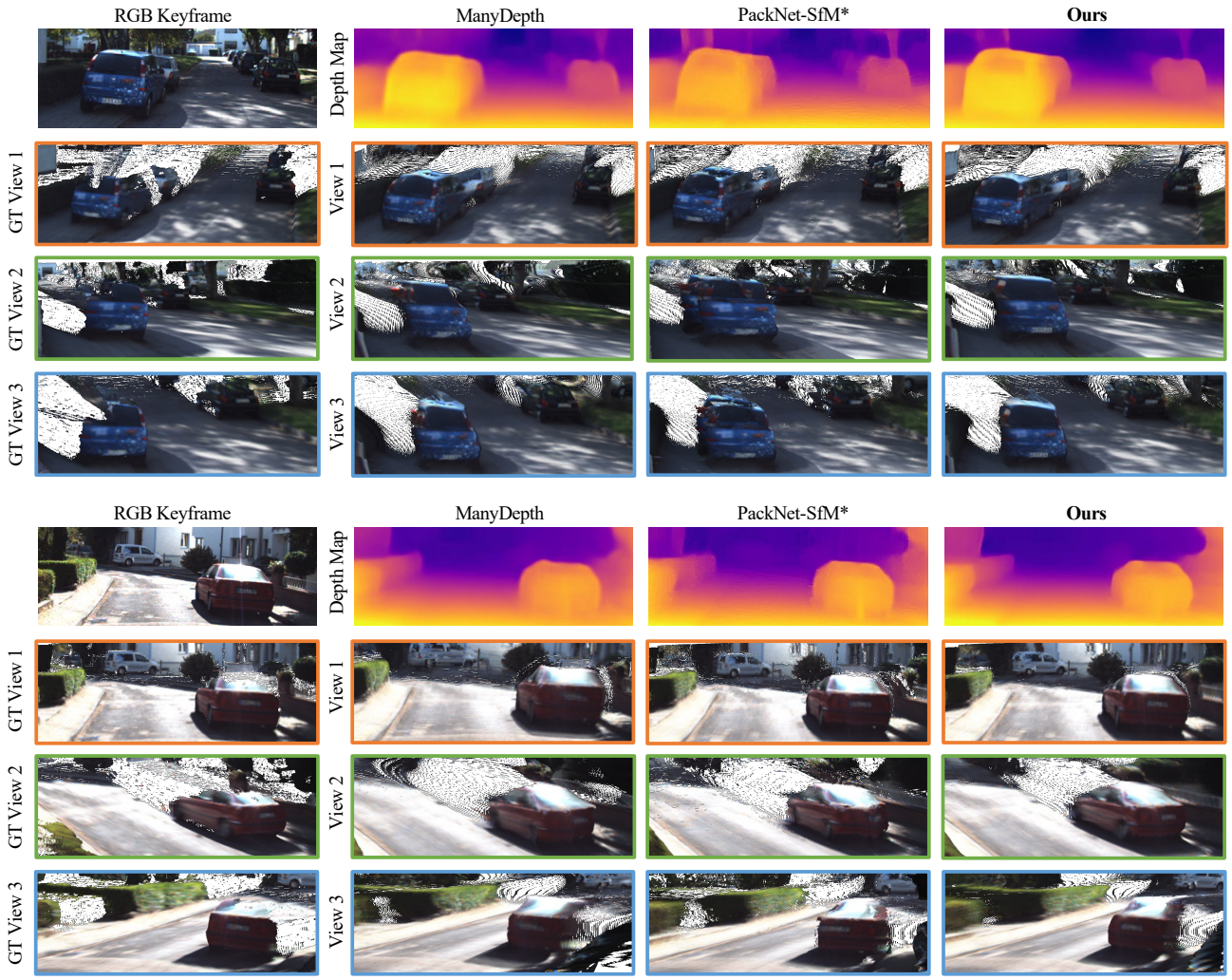

Figure A2: Qualitative reconstruction results from five consecutive depth predictions. Both, ManyDepth [28] and PackNet-SfM* [11] with velocity semi-supervision, suffer from "flying pixels", ghosting effects, and deformed objects, due to temporal inconsistencies. This is not directly apparent in a single frame depth prediction, but unfold when changing the viewpoint. Our method mitigates these artifacts to a large extent.

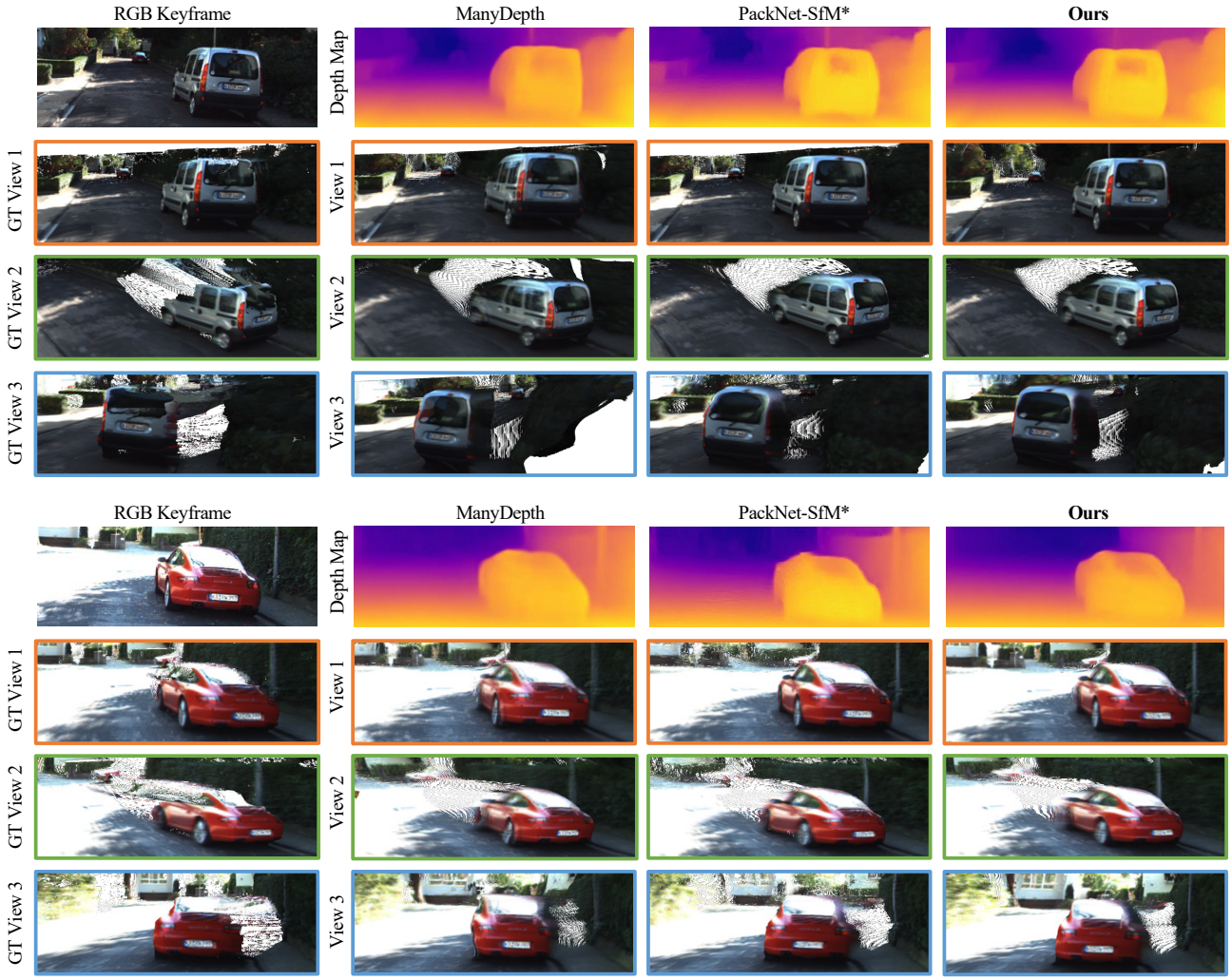

Figure A3: Qualitative reconstruction results from five consecutive depth predictions. Both, ManyDepth [28] and PackNet-SfM* [11] with velocity semi-supervision, suffer from "flying pixels", ghosting effects, and deformed objects, due to temporal inconsistencies. This is not directly apparent in a single frame depth prediction, but unfold when changing the viewpoint. Our method mitigates these artifacts to a large extent.