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

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

A. Temporal Consistency Metric (TCM)

The temporal consistency metric (TCM) is proposed to quantify the consistency of depth predictions across consecutive frames from monocular image sequences. The established standard accuracy metrics against ground-truth depth cannot reflect such consistency, due to applying per-frame alignment and individual comparison.

As described in the main paper, we evaluate the temporal consistency of multiple consecutive depth predictions in a sequence by measuring their alignment in 3D. Consecutive predictions are aligned in the same camera view by using the warping flow generated by ground truth depth and pose. We define the central frame of a short sequence of predictions as target depth D_t, and the other frames as the source depths D_s. The length of the short sequences is chosen as $k = \{3, 5, 7\}$. Longer sequences are not suitable for outdoor driving scenarios, as the visual overlap between images is too small (fast forward motion with $10fps$ for Kitti dataset [2]).

We introduce the nomenclature of *track* as the point-wise euclidean distance between the target depth and any source depth in 3D after being aligned in the same camera view:

$$track = \left\| T_{t \rightarrow s} \pi^{-1}(D_t) - \pi^{-1}(D_s^{'}) \right\|_2 \tag{S1}$$

where D_t is the depth from target frame, $T_{t \rightarrow s}$ is the ground-truth pose, $D_s^{'}$ is the interpolated depth from source frame aligned with the warping flow calculated from ground truth pose and depth, and $\pi^{-1}(\cdot)$ is the projective transformation function for 3D lifting.

We can now calculate *track* for ground-truth and predicted depths to acquire $track^{GT}$ and $track^{pred}$. Notably, for monocular methods with scale ambiguity, every frame from a sequence is scaled using **the same ratio** acquired by median-scaling from the target frame and its corresponding ground truth. Finally, we can utilize the computed *track* to define the absolute error (abs), square relative error (sq) and root mean square error (RMSE) to measure the depth consistency of each input:

$$TCM_{\text{abs}} = \frac{1}{H} \sum_{j=1}^{H} \left| track_j^{GT} - track_j^{pred} \right| \tag{S2}$$

$$TCM_{\text{sq}} = \frac{1}{H} \sum_{j=1}^{H} \left(track_j^{GT} - track_j^{pred} \right)^2 \tag{S3}$$

$$TCM_{\text{RMSE}} = \sqrt{\frac{1}{H} \sum_{j=1}^{H} \left(track_j^{GT} - track_j^{pred} \right)^2} \tag{S4}$$

where H is the total amount of valid tracks in current given input after outlier filtering as we suggested in the main paper (20%). To acquire the final TCM evaluation, we simply average over all measured TCM from every testing input.

To sum up, the TCM metrics can be interpreted as: the error between the euclidean distance of consecutive 3D predictions and the euclidean distance of their ground-truths, after alignment in the same camera view.

B. Additional Quantitative Results

B.1. TCM Analysis

A detailed statistical analysis on the absolute TCM metric can be found in Figs. S1, S2, and S3 (left) for different sequence lengths k. Our method has lowest mean and median absolute TCM errors with reduced outliers. This holds true over all sequence lengths $k = \{3, 5, 7\}$.

For a fair comparison between methods, and due to errors in the interpolated ground-truth LiDAR and moving objects, we set a threshold to filter out 20% of the largest

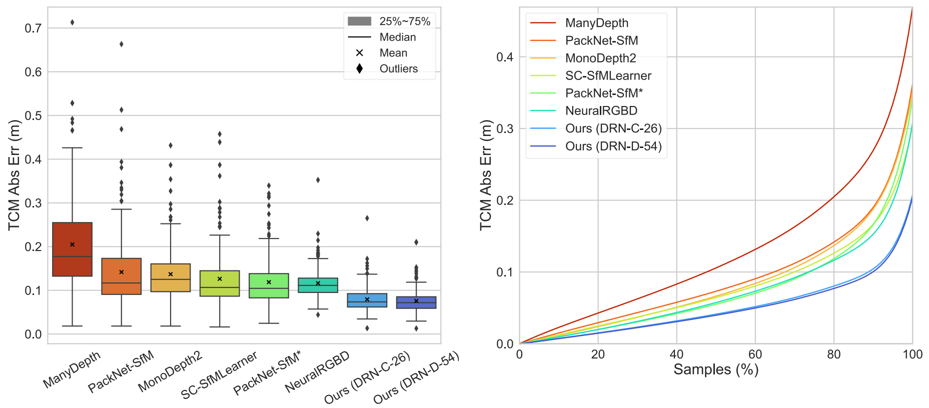

Figure S1: Detailed 3-frame-TCM statistical analysis. Left: distribution of the absolute errors of TCM tested on 3 frames. Right: Absolute errors of TCM measurement with different sampling rate for outliers handling.

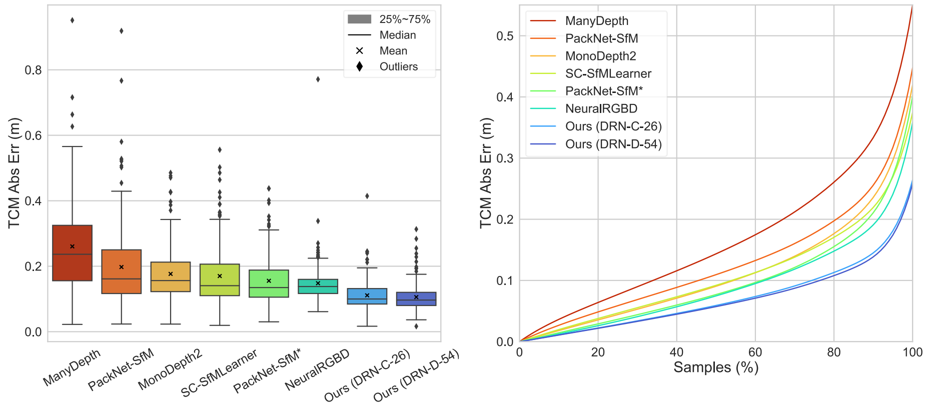

Figure S2: Detailed 5-frame-TCM statistical analysis. Left: distribution of the absolute errors of TCM tested on 5 frames. Right: Absolute errors of TCM measurement with different sampling rate for outliers handling.

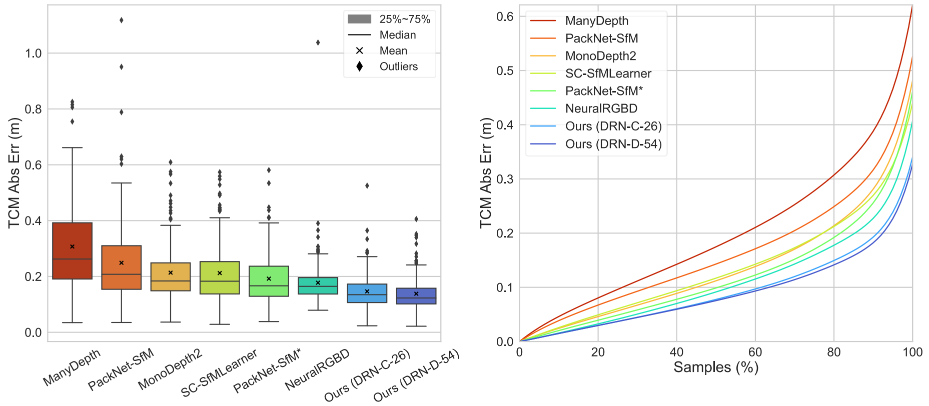

Figure S3: Detailed 7-frame-TCM statistical analysis. Left: distribution of the absolute errors of TCM tested on 7 frames. Right: Absolute errors of TCM measurement with different sampling rate for outliers handling.

Method	Test-time input	Abs Rel	Sq Rel	RMSE	RMSE log	$\sigma < 1.25$	$\sigma < 1.25^2$	$\sigma < 1.25^3$
ManyDepth [7]	Temporal Frames (Standard)	**0.098**	0.770	**4.459**	**0.176**	**0.900**	**0.965**	**0.983**
Ours (DRN-C-26)	Temporal Frames (Standard)	0.106	0.770	4.558	0.182	0.890	0.964	**0.983**
Ours (DRN-D-54)	Temporal Frames (Standard)	0.103	**0.746**	4.483	0.180	0.894	**0.965**	**0.983**
ManyDepth [7]	Single Frame (Static)	0.117	0.886	4.754	0.191	0.872	0.959	0.982
Ours (DRN-C-26)	Single Frame (Static)	0.107	0.784	4.596	0.184	0.888	0.963	**0.983**
Ours (DRN-D-54)	Single Frame (Static)	**0.104**	**0.760**	**4.515**	**0.181**	**0.982**	**0.964**	**0.983**

Table S1: Simulation of static camera scenario: Accuracy results on Kitti Eigen test split [1] for standard temporal frames input compared to static single frame input.

outliers. For this reason, we provide TCM results over varying outlier sampling ratios in Figs. S1, S2, and S3 (right). Our method consistently outperforms other methods over all sampling ratios.

B.2. Static Camera Performance

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. For this example, we only report accuracy metrics, as TCM metrics would not be meaningful for a static scene with a non-moving camera.

As ManyDepth [7] is also utilizing consecutive input frames, we use this method as our baseline. Table S1 summarizes the accuracy measures for the identical split as presented in the main paper. 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 [7].

C. Additional Qualitative Results

Figs. S4 and S5 show more qualitative 3D reconstruction results like the ones provided in the main paper. 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 [7] - despite achieving the best results in accuracy - suffers from deformed objects, ghosting effects, and ”flying pixels”. Similar artifacts are observed for the semi-supervised PackNet-SfM* [4]. Our method yields the most consistent reconstructions from consecutive depth maps. We also refer to the supplementary video for more qualitative results on the TCM metrics, 3D reconstructions, and temporal depth map predictions.

D. Network Architecture

An overview of the spatial-temporal attention module is illustrated in Fig. S6, and details on the individual attention mechanism are given in Fig. S7 for the spatial attention and the temporal attention in Fig. S8, respectively.

E. Implementation Details

We implement our model in PyTorch [6] and train for 25 epochs using Adam [5] with a batch size of 6 for our full DRN-C-26 [8] 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 [3]. 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.

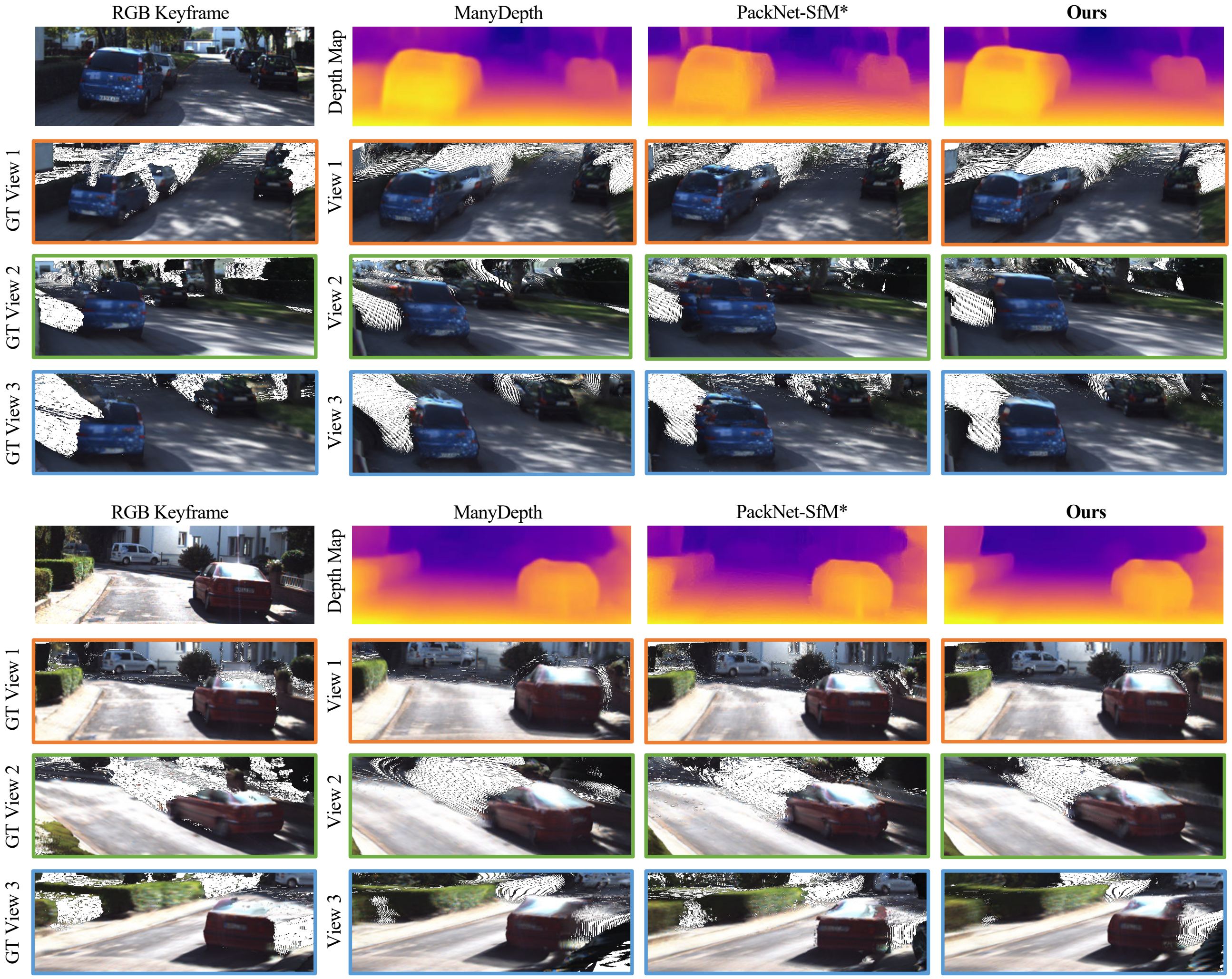

Figure S4: Qualitative reconstruction results from five consecutive depth predictions. Both, ManyDepth [7] and PackNet-SfM* [4] with velocity semi-supervision, suffer from ”flying pixels”, ghosting effects, and deformed objects, due to temporal inconsistencies. This is not directly apparent in a single frame depth prediction, but unfold when changing the viewpoint. Our method mitigates these artifacts to a large extent.

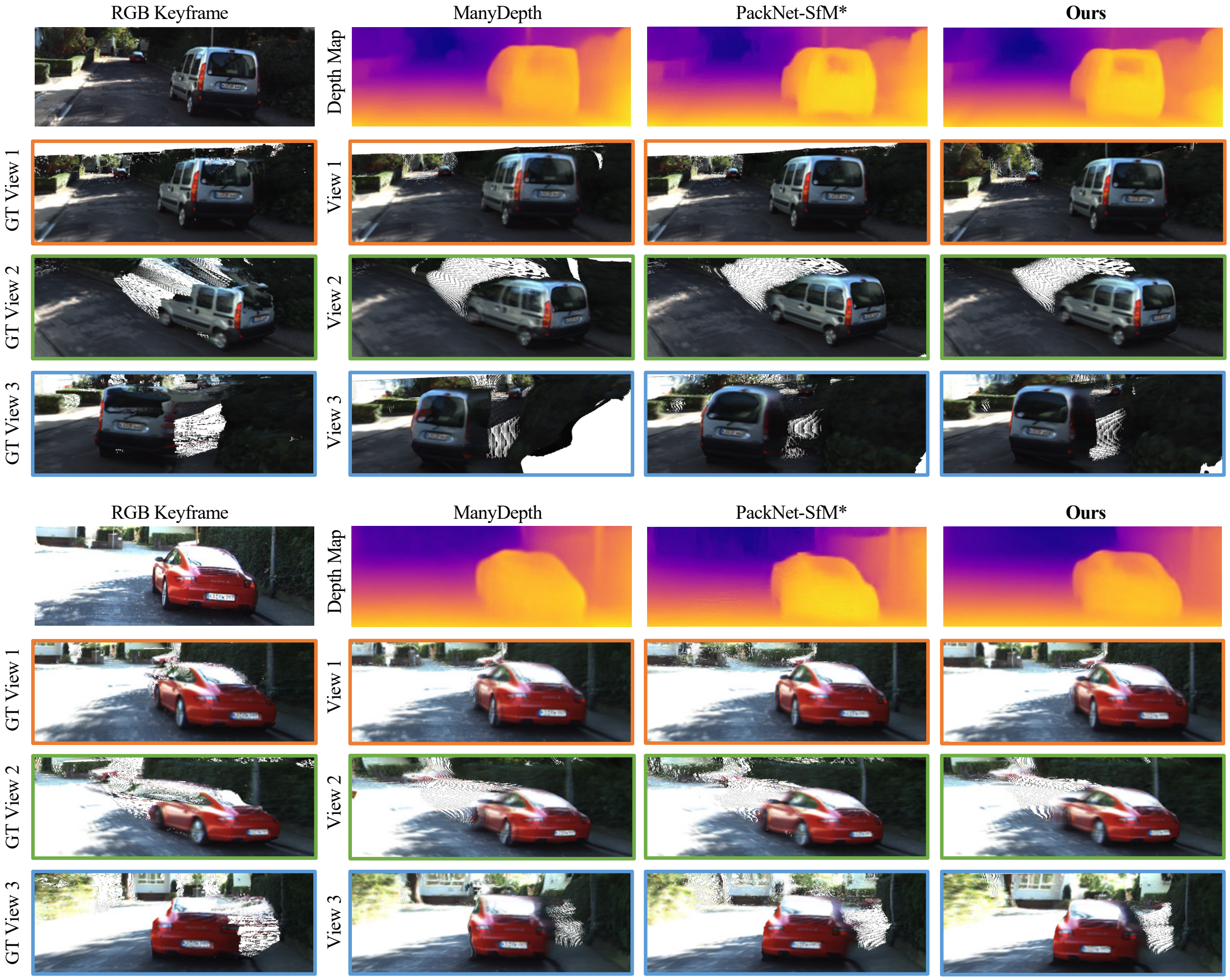

Figure S5: Qualitative reconstruction results from five consecutive depth predictions. Both, ManyDepth [7] and PackNet-SfM* [4] with velocity semi-supervision, suffer from ”flying pixels”, ghosting effects, and deformed objects, due to temporal inconsistencies. This is not directly apparent in a single frame depth prediction, but unfold when changing the viewpoint. Our method mitigates these artifacts to a large extent.

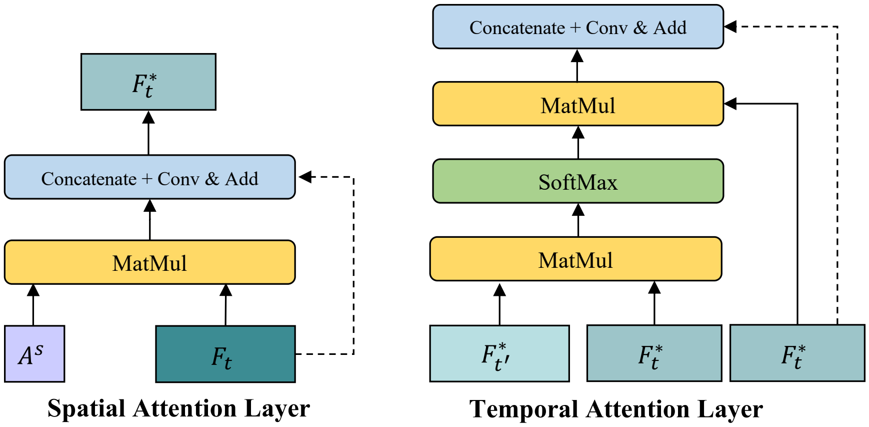

Figure S6: Overview of Spatial and Temporal Attention Modules.

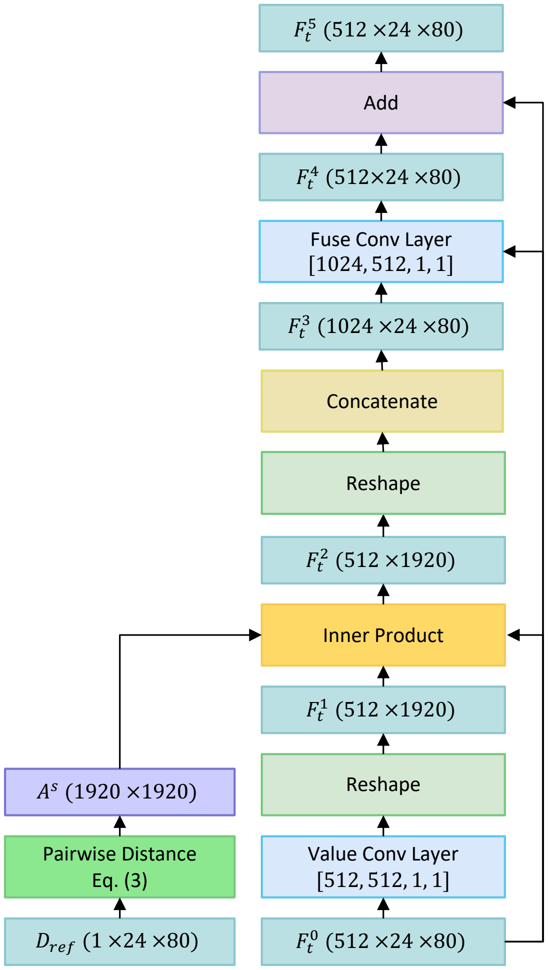

Figure S7: Detailed Spatial Attention Architecture.

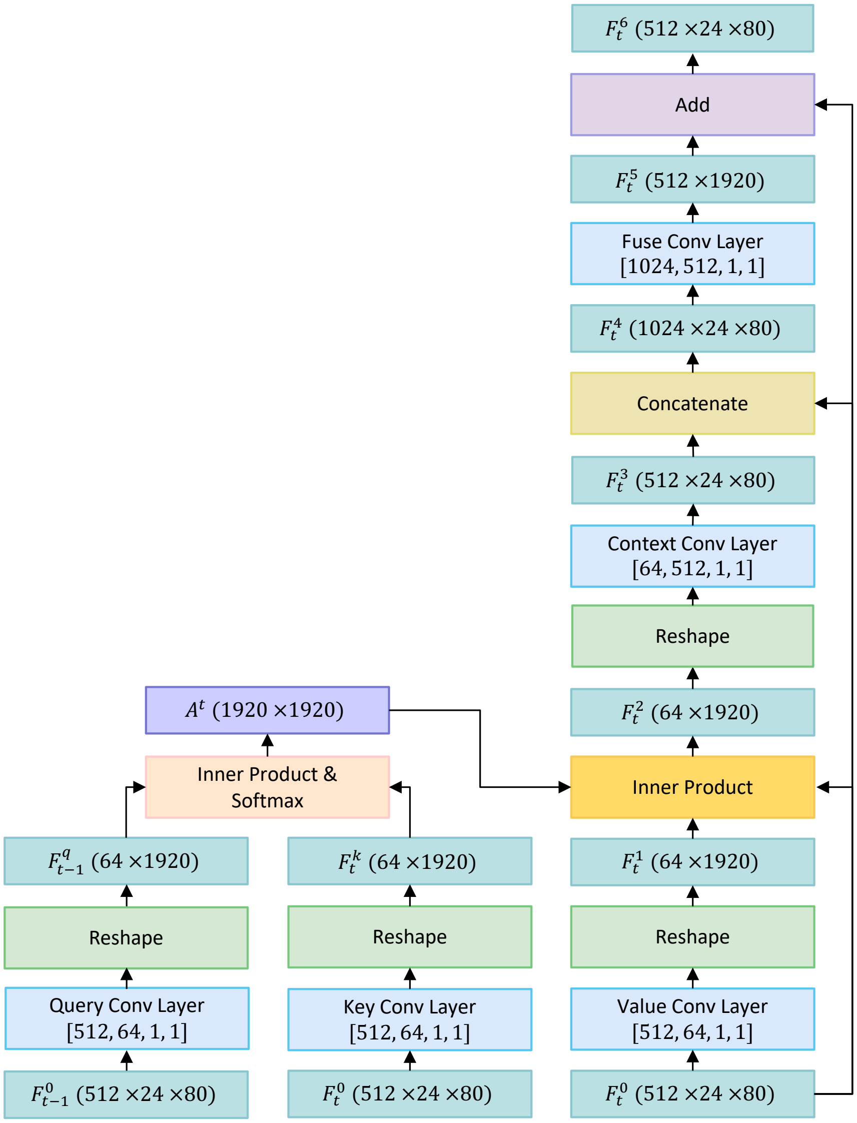

Figure S8: Detailed Temporal Attention Architecture.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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

[6] 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] 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.

[8] 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.