Stabilizing Streaming Video Geometry via Dynamic Feature Normalization

Supplementary Material

S1. Details of Loss Functions

In addition to our proposed temporal alignment losses ($\mathcal{L}_{align}$ and $\mathcal{L}_{temp}$), we retain the original monocular supervision from MoGe [46] to preserve single-frame geometric fidelity. The total monocular loss $\mathcal{L}_{MoGe}$ is composed of three terms:

$$\mathcal{L}_{MoGe} = \mathcal{L}_{local} + \mathcal{L}_{normal} + \mathcal{L}_{mask}, \tag{10}$$

where $\mathcal{L}_{local}$, $\mathcal{L}_{normal}$, and $\mathcal{L}_{mask}$ supervise local geometry, surface normals, and validity masks, respectively.

Multi-scale local geometry loss ($\mathcal{L}_{local}$). This term explicitly supervises local geometric structures. Given a ground-truth anchor point $\mathrm{p_j}$, we define a local spherical neighborhood $\mathcal{S}_j$ as:

$$\mathcal{S}_j = \{i \mid ||\mathrm{p_i} - \mathrm{p_j}|| \leq \mathrm{r_j}, \mathrm{i} \in \mathcal{M}\}. \tag{11}$$

Following MoGe, the radius r_j is depth-adaptive, defined as $r_j = \alpha \cdot z_j \cdot \frac{\sqrt{W^2+H^2}}{2 \cdot f}$, where z_j is the depth of $\mathrm{p_j}$, f is the focal length, and $\alpha \in (0, 1)$ is a scalar controlling the neighborhood size relative to the image diagonal. Within each neighborhood, we solve for the optimal affine parameters (s_j^*, t_j^*) to align predictions with the ground truth. We sample anchor sets $\mathcal{H}_\alpha$ across multiple scales $\alpha \in \{\frac{1}{4}, \frac{1}{16}, \frac{1}{32}\}$ and compute the accumulated error:

$$\mathcal{L}_{local} = \sum_{\alpha} \sum_{j \in \mathcal{H}_\alpha} \sum_{i \in \mathcal{S}_j} \frac{1}{z_i} ||s_j^* \hat{\mathrm{p}}_\mathrm{i} + \mathrm{t}_\mathrm{j}^* - \mathrm{p_i}||_1. \tag{12}$$

Normal loss ($\mathcal{L}_{normal}$). To enforce high-quality surface details, we minimize the angular error between predicted and ground-truth normals:

$$\mathcal{L}_{normal} = \sum_{i \in \mathcal{M}} \angle(\hat{n}_i, n_i), \tag{13}$$

where the predicted normal $\hat{n}_i$ is derived from the cross-product of adjacent vectors on the predicted point map grid, and $\angle(\cdot, \cdot)$ measures the angular difference.

Mask loss ($\mathcal{L}_{mask}$). This loss is employed to identify valid geometric regions (e.g., suppressing sky or infinity in outdoor scenes). It is formulated as the mean squared error between the predicted mask $\hat{M}$ and the valid region label:

$$\mathcal{L}_{mask} = ||\hat{M} - (1 - M_{\text{inf}})||_2^2, \tag{14}$$

where M_{inf} denotes the infinity mask. During inference, $\hat{M}$ is binarized with a threshold of 0.5.

S2. Details of Evaluation

Evaluation Metrics. We adopt the Absolute Relative Error (AbsRel) and the inlier ratio δ_1 as our primary metrics. Averaged over all valid pixels $\mathcal{M}$, these are defined as:

$$\text{AbsRel} = \frac{1}{|\mathcal{M}|} \sum \frac{|d - \hat{d}|}{d}, \tag{15}$$

$$\delta_1 = \frac{1}{|\mathcal{M}|} \sum \mathbb{I}\left[\max\left(\frac{d}{\hat{d}}, \frac{\hat{d}}{d}\right) < 1.25\right], \tag{16}$$

where d is the ground truth depth, $\hat{d}$ is the predicted depth (after alignment, if applicable), and $\mathbb{I}[\cdot]$ denotes the indicator function, which evaluates to 1 if the condition is met and 0 otherwise.

Evaluation Protocols. We employ three distinct protocols to evaluate different capabilities:
(1) **Metric Depth Protocol:** For models designed to predict absolute metric depth (e.g., DepthPro, MoGe-v2), we evaluate the raw predictions directly without any post alignment.
(2) **Video Depth Protocol (Global Alignment):** To evaluate temporal consistency, we align the entire predicted sequence using a **single** global transformation. Given predictions $\{\hat{\mathrm{d}}_\mathrm{j}\}_{j=1}^L$ and ground truth $\{\mathrm{d_j}\}_{j=1}^L$, we solve for the optimal global scale s^* and shift t^* that minimize the error across all frames simultaneously:

$$(s^*, t^*) = \underset{s,t}{\text{argmin}} \sum_{j=1}^{L} \sum_{i \in \mathcal{M}} \frac{1}{d_j^i} ||s\hat{\mathrm{d}}_j^i + t - \mathrm{d}_j^i||_1. \tag{17}$$

This global transformation is then applied uniformly to the sequence: $\{\mathrm{d}_j^{align}\} = \{s^* \cdot \hat{\mathrm{d}}_j + t^*\}$. This protocol strictly penalizes scale drift over time.
(3) **Image Depth Protocol (Per-Frame Alignment):** To evaluate per-frame geometric quality in isolation, we align each frame independently. For each frame j, we compute specific parameters (s_j^*, t_j^*):

$$(s_j^*, t_j^*) = \underset{s,t}{\text{argmin}} \sum_{i \in \mathcal{M}} \frac{1}{d_j^i} ||s\hat{\mathrm{d}}_j^i + t - \mathrm{d}_j^i||_1. \tag{18}$$

The metrics are then computed on the individually aligned frames: $\{\mathrm{d}_j^{align}\} = \{s_j^* \cdot \hat{\mathrm{d}}_j + t_j^*\}$.

S3. Dataset Configuration

S3.1. Training Dataset

To finetune our newly designed DyFN module, we use seven different synthetic datasets which contain continuous frames and depth annotations. Details are shown in the

Tab 4. Our training dataset contains total around 1M images and our main experiment are trained with the sequence length 12. To increase the robustness of our model training, we randomly select stride from 1 to 5 to sample the continuous frames.

Table 4. Datasets used for training.

Name	Domain	# Frames	Weight
IRS [44]	Indoor	101K	20.1%
PointOdyssey [59]	Indoor	79K	27.8%
Dynamic Replica [23]	Indoor	143K	17.4%
Spring [27]	In-the-wild	5K	2.4%
MidAir [14]	In-the-wild	423K	9.3%
KenBurns3D [29]	In-the-wild	76K	5.6%
TartanAir [49]	In-the-wild	306K	17.4%

S3.2. Evaluation Dataset

Video & Image Depth Estimation. We evaluate both video and image depth estimation performance using four diverse benchmarks. The details are as follows:

- **Sintel [6]**. We utilize all 23 sequences for evaluation. We evaluate directly at the original 1024×436 resolution without resizing.
- **ScanNet [12]**. We use the standard test split, comprising 100 scenes. We extract **90 continuous frames** per scene at a rate of 15 frames per second (FPS). To handle the black borders resulting from calibration, we **follow** DepthCrafter [22] and crop 8 pixels from the top and bottom edges, and 11 pixels from the left and right edges.
- **KITTI [16]**. We sample **110 frames** across all sequences in the official KITTI Depth split, maintaining the original frame rate.
- **Bonn [31]**. We selected 5 scenes from this dataset, each contributing 110 frames for evaluation.

Long Sequence Depth Estimation. To assess long-term stability and error accumulation, we adopt the same **ScanNetV2** test split. For this specific protocol, we extract **500 continuous frames** per scene, sampled at the depth camera's original frame rate. Furthermore, the same cropping strategy used for the short sequence evaluation is applied.

S4. Reconstruction Comparison

S4.1. Reconstruction Algorithm

To rigorously evaluate the scale-shift consistency of our proposed model, we employ a geometric alignment protocol based on point correspondences. Given a sequence of L continuous frames $\{\mathcal{I}_j\}_{j=1}^{L}$ and their predicted point clouds in the camera coordinate system $\{\mathcal{P}_j^{\text{cam}}\}_{j=1}^{L}$, the objective is to estimate the rigid transformation (pose) $\{R_j|t_j\}$ for the frame at timestamp j.

We select a set of reference frames $\mathcal{K} = \{j-1, j-5, j-21\}$ whose ground-truth poses are assumed known, providing their corresponding world coordinates $\{\mathcal{P}_k^{\text{world}}\}_{k\in\mathcal{K}}$. We first leverage PDCNet to establish reliable point correspondences, composing the matched 3D point pairs $\{\mathbf{p}_j^{\text{cam}}, \mathbf{p}_k^{\text{world}}\}_{k\in\mathcal{K}}$.

This set of correspondences is then used to solve for the optimal rigid transformation $\{R_j|t_j\}$ that aligns the predicted camera-centric point cloud $\mathcal{P}_j^{\text{cam}}$ into the global world coordinate system. To handle the inevitable noise and outliers present in the correspondences, we employ the Random Sample Consensus (RANSAC) algorithm. Within the RANSAC iterative loop, the rigid transformation $\{R_j, t_j\}$ is determined by solving the Absolute Orientation Problem (Procrustes problem). Specifically, we seek to minimize the squared error between the aligned source points and the target points:

$$\min_{R_j,t_j} \sum_{m=1}^{N} \left\| R_j \mathbf{p}_{j,m}^{\text{cam}} + t_j - \mathbf{p}_{k,m}^{\text{world}} \right\|_2^2, \tag{19}$$

where N is the number of inlier correspondence pairs identified by RANSAC. The closed-form solution for the optimal rotation R_j and translation t_j is obtained efficiently using the Singular Value Decomposition (SVD) method applied to the centered cross-covariance matrix.

S4.2. Qualitative Comparison

We utilize the robust reconstruction algorithm detailed in Section S4.1 for qualitative video reconstruction. Our results are presented in Figure 7 (short sequences) and Figure 8 (long/dynamic sequences).

Figure 7 provides comparative results, demonstrating our approach's superior geometric consistency and clearer structural reconstruction in both indoor and outdoor scenes compared to baselines such as VideoDepthAnything (VDA) and FlashDepth. This highlights the immediate benefits of our dynamic feature stabilization.

Furthermore, Figure 8 showcases our method's robust performance in complex scenarios. The results confirm sustained scale-shift consistency over **long-term sequences**, and crucially, illustrate the capability of our model to produce coherent geometric reconstructions even in the presence of dynamic scene elements.

S5. Training Length Influence

To investigate the influence of training sequence length, we conduct an ablation study across four evaluation benchmarks, as detailed in Table 5. The results clearly demonstrate that varying the input frame length from 8 to 24

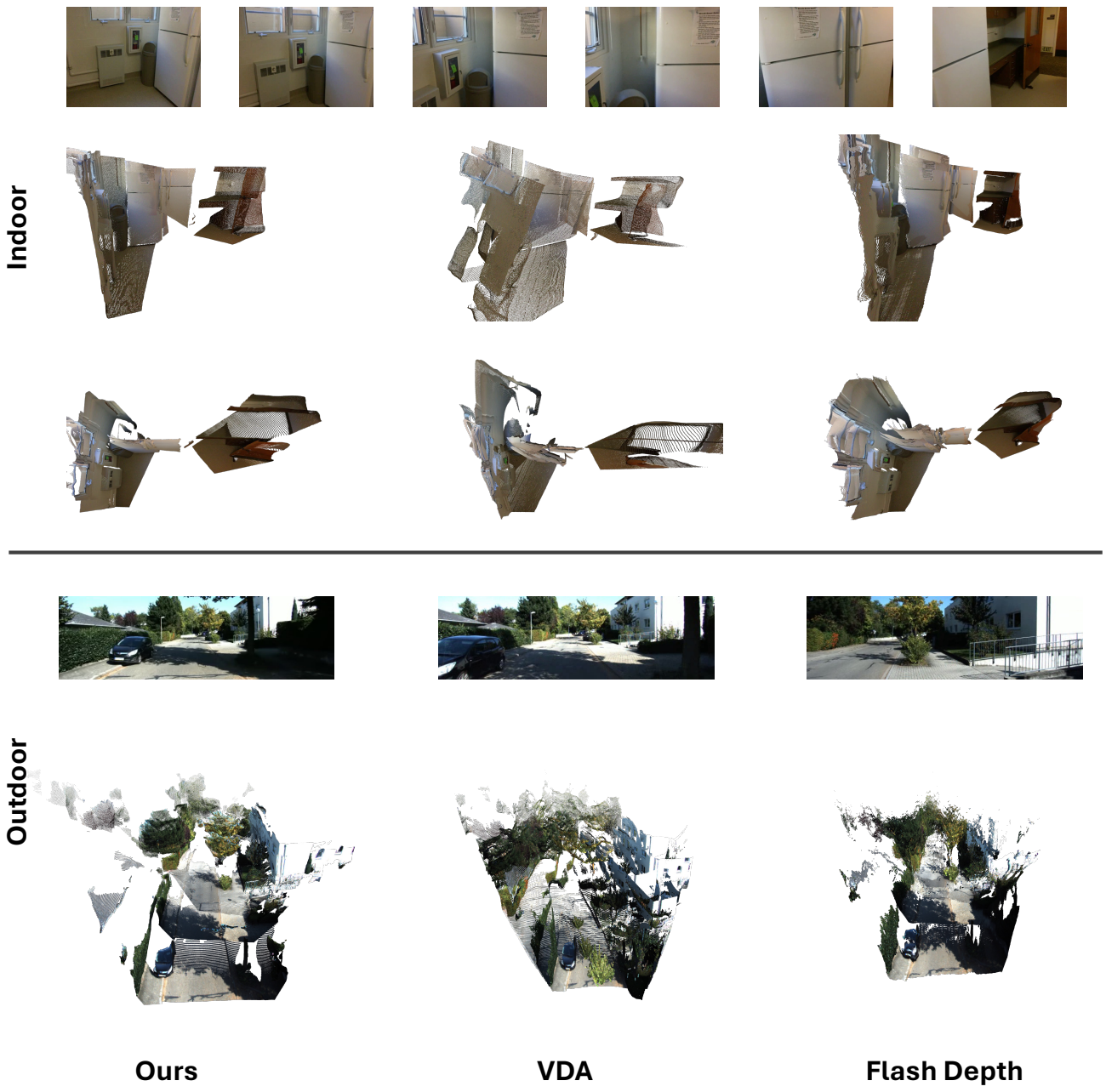

Figure 7. **Qualitative 3D Reconstruction Comparison in Diverse Scenes.** We present a qualitative comparison of 3D reconstruction results across challenging **indoor** and **outdoor** environments. Compared to key video depth baselines (VDA and FlashDepth), our method consistently demonstrates **superior geometric fidelity** and **enhanced temporal consistency**, resulting in noticeably more stable and accurate 3D structures.

frames has a **negligible impact** on the final accuracy metrics across all datasets. This observation strongly confirms the inherent stability of our DyFN module and demonstrates that the necessary temporal alignment and scale-shift information are learned highly **efficiently**, even from relatively short sequence clips. This stability allows us to select 12 frames as the standard training length, optimizing the balance between computational efficiency and stable performance.

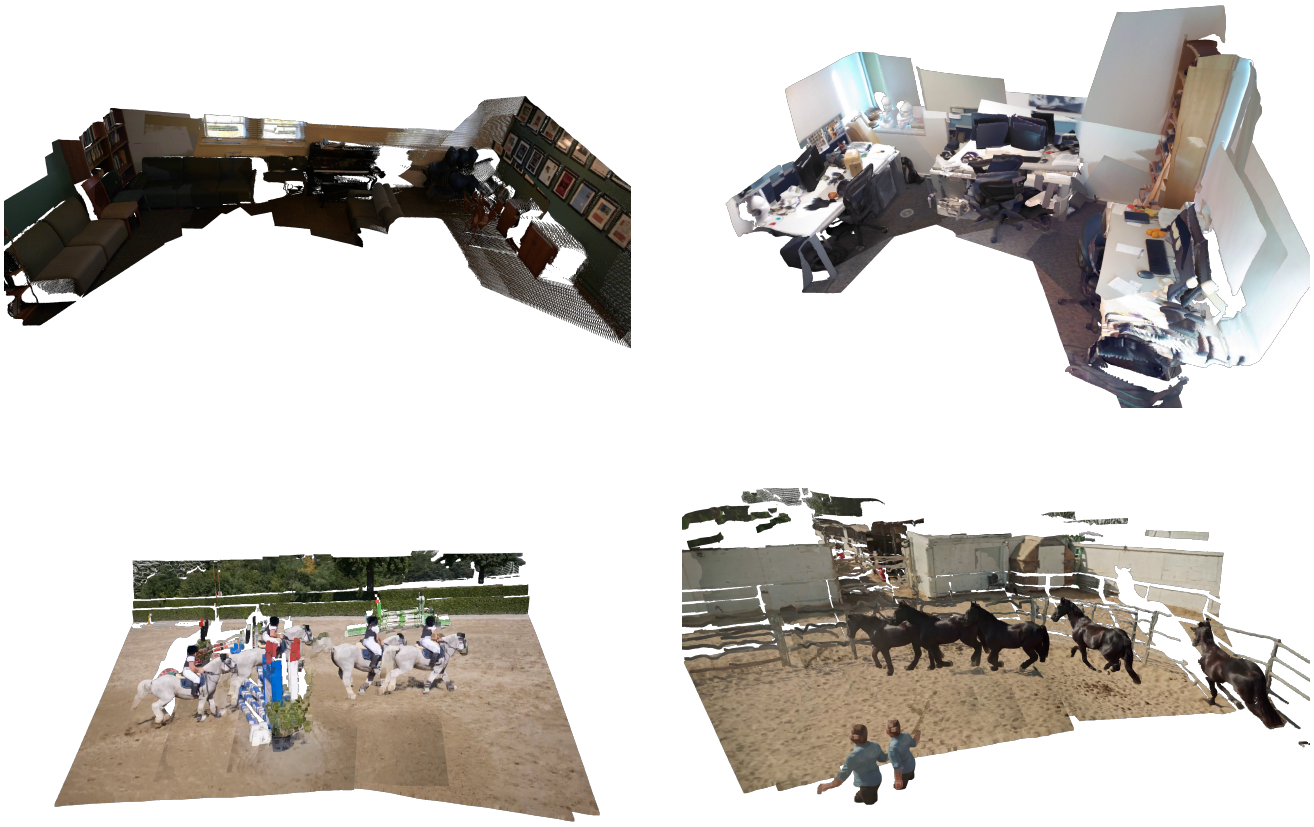

Figure 8. Qualitative results on long sequence reconstruction results (more than 500 frames) and dynamic reconstruction results.

	Sintel		Scannet		KITTI		Bonn	
Frame Length	Abs Rel↓	$\delta < 1.25$ ↑	Abs Rel↓	$\delta < 1.25$ ↑	Abs Rel↓	$\delta < 1.25$ ↑	Abs Rel↓	$\delta < 1.25$ ↑
8	0.182	73.3	0.072	96.2	0.064	97.3	0.043	98.4
12	0.180	73.0	0.073	96.6	0.062	97.3	0.044	98.4
16	0.181	73.4	0.072	96.2	0.064	97.0	0.044	98.4
24	0.182	73.3	0.071	96.3	0.067	97.6	0.045	98.4

Table 5. **Ablation Study on Training Sequence Length.** The negligible variation in performance across different sequence lengths (8 to 24 frames) confirms the stability and efficient learning of temporal consistency in our method.

S6. Implementation Details

As illustrated in Figure 9, our proposed method utilizes a **single ConvGRU** recurrent structure to model temporal dependencies and generate the hidden state. This targeted design leads to superior parameter efficiency: we only need to optimize the weights of the ConvGRU module, which constitutes a dramatically reduced parameter budget of 2% **(approximately** 5**M)** of the total network parameters. This is orders of magnitude lower than previous video-trained methods like DepthCrafter (1422.8M) and VideoDepthAnything (381.8M), allowing for highly efficient fine-tuning and deployment.

S7. Generalization Capability

To rigorously demonstrate the **generalization capability** of our proposed DyFN module, we integrate it into the **DepthAnythingV2 (DAv2)** framework. Unlike our primary Monocular Geometry Estimation (MGE) backbone, DAv2 is a standard monocular depth model designed to output **disparity**, rendering the MoGe-specific geometry losses ($\mathcal{L}_{\text{MoGe}}$) unsuitable. To adapt, we utilize two key supervision signals: the standard **scale-shift invariant loss** ($\mathcal{L}_{\text{ssi}}$) proposed in MiDaS [3] and our inter-frame temporal loss ($\mathcal{L}_{\text{temp}}$) defined in Equation 8. We adopt the same parameter-efficient fine-tuning strategy (freezing the backbone). As shown in Table 6, integrating the DyFN module dramatically boosts DAv2's performance across diverse domains. Specifically, the δ_1 accuracy on Sintel improves substantially from 55.4 to $\mathbf{63.0}$, and on KITTI, it rises sharply from 80.4 to $\mathbf{92.9}$. This significant uplift validates that DyFN's mechanism for stabilizing feature statistics is broadly effective across different architectural types and output representations.

S8. Limitation and Future Work

While our Dynamic Feature Normalization (DyFN) module successfully mitigates temporal inconsistencies and main-

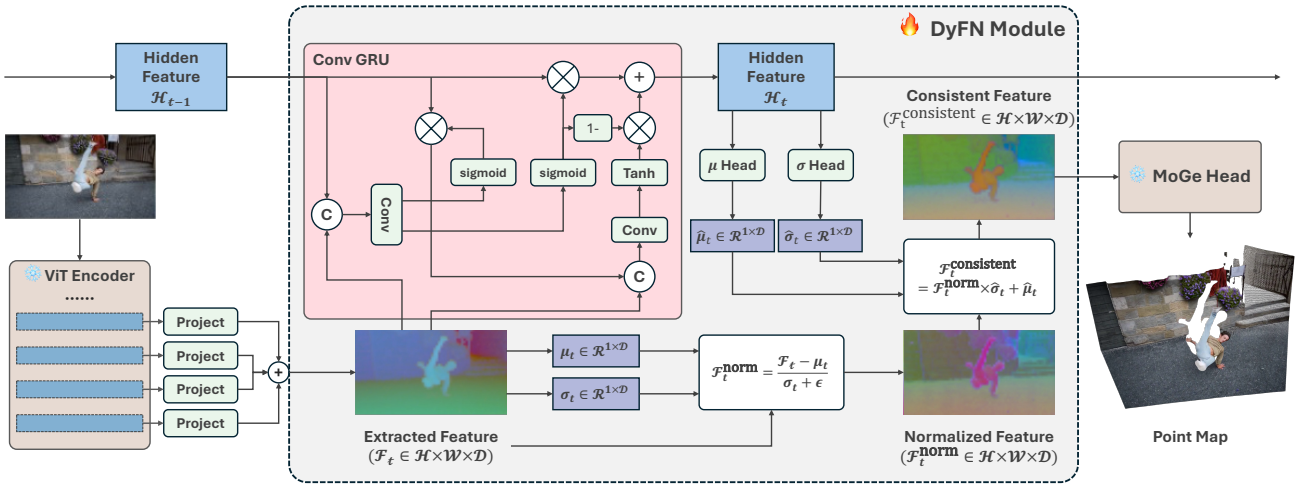

Figure 9. **Detailed network structures with ConvGRU.** We show the detailed structures when DyFN module use the ConvGRU as recurrent module to merge the historical information.

	Sintel		Scannet		KITTI		Bonn	
Method	Abs Rel↓	$\delta < 1.25$ ↑	Abs Rel↓	$\delta < 1.25$ ↑	Abs Rel↓	$\delta < 1.25$ ↑	Abs Rel↓	$\delta < 1.25$ ↑
DAV2	0.367	55.4	0.135	82.2	0.140	80.4	0.106	92.1
FlashDepth	0.265	64.2	0.101	90.3	0.103	89.5	**0.053**	**98.0**
DAV2 + DyFN	**0.242**	**64.8**	**0.087**	**93.2**	**0.093**	**93.3**	**0.053**	97.9
MoGe	0.216	65.3	0.117	84.7	0.076	96.0	0.074	95.5
MoGe + DyFN	**0.180**	**73.0**	**0.073**	**96.6**	**0.062**	**97.3**	**0.044**	**98.4**

Table 6. Qualitative results on streaming video depth estimation.

tains superior scale-shift consistency across long sequences, its performance ceiling is fundamentally constrained. A primary limitation is that the achievable accuracy remains bounded by the **per-frame aligned geometric fidelity** of the underlying monocular depth backbone. Since DyFN operates by stabilizing the existing feature representation, it does not leverage the redundant information across multiple continuous frames to resolve fundamental monocular ambiguities. Consequently, our method cannot inherently improve the geometric accuracy of any single frame beyond the backbone's original capability.

Future work will focus on extending the DyFN framework to better harness the structural cues present in continuous frames. By integrating multi-frame information within the recurrent structure, we aim to push past the conventional limits of single-image depth estimation, significantly enhancing the geometric fidelity and ambiguity resolution capacity of the resulting depth predictions.

References

[1] Nicolas Ballas, Yao Li, Chris Pal, and Aaron Courville. Delving deeper into convolutional networks for learning video representations. In *International Conference on Learning Representations (ICLR)*, 2016. 3, 5

[2] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 1, 3

[3] Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3.1 – a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023. 3, 4

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 3

[5] Alexey Bochkovskiy, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *The Thirteenth International Conference on Learning Representations*, 2025. 6

[6] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conf. on Computer Vision (ECCV)*, pages 611–625, 2012. 6, 2

[7] Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy. Must3r: Multi-view network for stereo 3d reconstruction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1050–1060, 2025. 3

[8] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. *arXiv preprint arXiv:2501.12375*, 2025. 3

[9] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22831–22840, 2025. 8

[10] Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Ttt3r: 3d reconstruction as test-time training. *arXiv preprint arXiv:2509.26645*, 2025. 3, 6

[11] Gene Chou, Wenqi Xian, Guandao Yang, Mohamed Abdelfattah, Bharath Hariharan, Noah Snavely, Ning Yu, and Paul Debevec. Flashdepth: Real-time streaming video depth estimation at 2k resolution. *arXiv preprint arXiv:2504.07093*, 2025. 2, 3, 6, 8

[12] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 6, 8, 2

[13] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10786–10796, 2021. 3

[14] Michael Fonder and Marc Van Droogenbroeck. Mid-air: A multi-modal dataset for extremely low altitude drone flights. In *Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2019. 2

[15] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pages 241–258. Springer, 2024. 3

[16] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 6, 2

[17] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*, 2024. 3

[18] Yuliang Guo, Sparsh Garg, S Mahdi H Miangoleh, Xinyu Huang, and Liu Ren. Depth any camera: Zero-shot metric depth estimation from any camera. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26996–27006, 2025. 3

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 3

[20] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997. 3

[21] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10579–10596, 2024. 3

[22] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. *arXiv preprint arXiv:2409.02095*, 2024. 2, 3, 6

[23] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Dynamicstereo: Consistent dynamic depth from stereo videos. *CVPR*, 2023. 2

[24] Bingxin Ke, Dominik Narnhofer, Shengyu Huang, Lei Ke, Torben Peters, Katerina Fragkiadaki, Anton Obukhov, and Konrad Schindler. Video depth without video models. *arXiv preprint arXiv:2411.19189*, 2024. 3

[25] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024. 3, 6

[26] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 3

[27] Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[28] Riku Murai, Eric Dexheimer, and Andrew J Davison. Mast3r-slam: Real-time dense slam with 3d reconstruction priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16695–16705, 2025. 3

[29] Simon Niklaus, Long Mai, Jimei Yang, and Feng Liu. 3d ken burns effect from a single image. *ACM Transactions on Graphics*, 38(6):184:1–184:15, 2019. 2

[30] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 3

[31] E. Palazzolo, J. Behley, P. Lottes, P. Giguère, and C. Stachniss. ReFusion: 3D Reconstruction in Dynamic Environments for RGB-D Cameras Exploiting Residuals. In *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS)*, 2019. 6, 2

[32] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10106–10116, 2024. 1, 3

[33] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler, 2025. 3

[34] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3

[35] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 1, 3

[36] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 3

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[38] Jiahao Shao, Yuanbo Yang, Hongyu Zhou, Youmin Zhang, Yujun Shen, Vitor Guizilini, Yue Wang, Matteo Poggi, and Yiyi Liao. Learning temporally consistent video depth from video diffusion priors, 2024. 2, 3, 6

[39] Yang-Tian Sun, Xin Yu, Zehuan Huang, Yi-Hua Huang, Yuan-Chen Guo, Ziyi Yang, Yan-Pei Cao, and Xiaojuan Qi. Unigeo: Taming video diffusion for unified consistent geometry estimation. *arXiv preprint arXiv:2505.24521*, 2025. 3

[40] Cheng Tan, Zhangyang Gao, Lirong Wu, Yongjie Xu, Jun Xia, Siyuan Li, and Stan Z Li. Temporal attention unit: Towards efficient spatiotemporal predictive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18770–18782, 2023. 3

[41] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 3

[42] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. *arXiv preprint arXiv:2408.16061*, 2024. 3

[43] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. *arXiv preprint arXiv:2503.11651*, 2025. 3, 6

[44] Qiang Wang, Shizhen Zheng, Qingsong Yan, Fei Deng, Kaiyong Zhao, and Xiaowen Chu. Irs: A large naturalistic indoor robotics stereo dataset to train deep models for disparity and surface normal estimation, 2021. 2

[45] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. *arXiv preprint arXiv:2501.12387*, 2025. 2, 3, 6

[46] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5261–5271, 2025. 1, 2, 4, 6

[47] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *arXiv preprint arXiv:2507.02546*, 2025. 1, 3, 6

[48] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 3

[49] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020. 2

[50] Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. *Advances in Neural Information Processing Systems*, 35:3502–3516, 2022. 3

[51] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela

Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17969–17980, 2023.

[52] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 3

[53] Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21924–21935, 2025. 3

[54] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 1, 3, 6

[55] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 1, 3, 6

[56] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 204–213, 2021. 3

[57] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9043–9053, 2023. 1, 3

[58] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024. 3, 6

[59] Yang Zheng, Adam W. Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J. Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *ICCV*, 2023. 2