Reliev3R: Relieving Feed-forward 3D Reconstruction from Multi-View Geometric Annotations

Supplementary Material

A. Training Setup for Baselines

We introduce the training setup of baseline methods in this section for better depicting advantages of Reliev3R.

CUT3R, VGGT, π^3, and MapAnything [2, 5, 6, 8] . These works are representatives as the edge of FFRMs. Each of these FFRM is trained on 15-17 diverse datasets with multi-view geometric annotations, which include approximately 200k scenes in total.

$\pi^{3\dagger}$ [8]. For a comprehensive comparison between Reliev3R and SOTA FFRMs, we train π^3 [8] from scratch with the same training set as Reliev3R, *i.e*. the DL3DV-10K [3] dataset. The result is denoted as $\pi^{3\dagger}$. ***The parameter size of $\pi^{3\dagger}$ is controlled approximate to Reliev3R as 450M.*** The training for $\pi^{3\dagger}$ is directed with the same learning rate schedule as Reliev3R, and optimized for 30k steps until a full convergence on the validation set, which is randomly sampled as 2% of the DL3DV-10K dataset. Other setups, such as training objectives, follow π^3 [8] by default.

MVDUSt3R [4]. MVDUSt3R is one of the pioneering works to expand the input of DUSt3R [7] from two views to multiple views. MVDUSt3R is trained by finetuning DUSt3R on the subset of its training data, which is relatively small compared to the training data of SOTA FFRMs. As a result, MVDUSt3R is hindered with overfitting issue when performing a zero-shot evaluation on unseen datasets such as DL3DV-benchmark [3], which explains the catastrophic performance of MVDUSt3R in the paper.

FLARE [10]. Similar to MVDUst3R [4], FLARE is one of the early works to expand the input of DUSt3R [7] to multiple views. Differently, FLARE is capable to directly predict the camera pose along the forward pass of the model. ***We adopt the direct output from FLARE model as its camera pose prediction, instead of following DUSt3R [7] to post process the predicted point maps with a PnP solver [1] to estimate camera poses.*** As shown in the paper (both figures and tables), we find that the directly predicted camera pose of FLARE has an inconsistent scale with the point maps. This is because, to handle the unreliability in reprojection when camera pose is learned from scratch, FLARE proposes a learnable reprojector conditioned on the learned camera poses, without further constraint on the scale consistency between the reprojected point maps and camera positions. The unreliability of reprojection is addressed by Reliev3R with a trigonometry-based reprojection loss, which is capable to produce meaningful gradient for optimization given any random camera poses.

Table 1. We train π^3 from-scratch (random initial weights) in a fully-supervised manner on a down-sampled mixture of synthetic datasets, including ASE dataset, TartanAir dataset, MVS-Synth dataset, Blended-MVS dataset and *etc*. following VGGT. The training runs for 20k iterations with a total batch size of 128. The result is denoted as 'Synth Fully-Sup'. in the table. Then we subsequently finetune this checkpoint on DL3DV dataset for 20k steps using the weak supervision proposed by Reliev3R. The result is denoted as '+Reliev3R' in the table. We report the same metrics as Table 1 in the paper, which are evaluated on DL3DV-benchmark dataset. We also evaluate camera pose estimation for RayZer on this benchmark with its recently open-sourced checkpoint.

Method	Point Map		Camera Pose		Depth Map	
	rel ↓	τ ↑	ATE ↓	AUC ↑	rel ↓	τ ↑
RayZer	×	×	0.786	0.362	×	×
Synth Fully-Sup.	0.277	0.475	0.042	17.075	0.208	0.357
+Reliev3R	**0.137**	**0.667**	**0.033**	**34.240**	**0.106**	**0.679**

AnyCam [9]. Since there is not a proper baseline to be compared with Reliev3R as a weakly-supervised FFRM, we introduce AnyCam as a less rigorous baseline. AnyCam is one of the few works to explore training a feed-forward 3D model without multi-view geometric annotations. More concretely, AnyCam focuses on estimating both intrinsics and poses of cameras given a group of input views followed by view-wise metric depth maps and optical flows. By back-projecting the input metric depth maps into the world coordinate with predicted camera parameters, a registered point cloud can be produced to be compared with FFRMs. For a fair comparison, we condition the camera pose prediction of AnyCam on the ground truth focal length of cameras.

B. Extending Reliev3R to Semi-Supervision

It would be a practical approach to utilize Reliev3R to generalize a FFRM, which is fully-supervised on synthetic data, to realistic data, where ground-truth is hard to access. As a proof of concept, we perform an experiment in Tab. 1. The FFRM trained merely on synthetic data performs bad on DL3DV because of domain gap between realistic data and synthetic data. Reliev3R succeeds to improve this FFRM, demonstrating the feasibility of the warm-up&finetune usage of Reliev3R. In Tab. 1, both 'Synth Fully-Sup.' and '+Reliev3R' are equipped with a FoV head for intrinsic prediction, where the former is supervised with ground-truth intrinsics and the latter is supervised with pseudo FoV predicted by MoGe2. To avoid overfitting on FoV, we introduce image

Table 2. Ablation study on different confidence weight α on DL3DV-benchmark [3] dataset. $\alpha = 1.0$ is adopted in the paper to train Reliev3R. Absolute relative error (rel) and inlier ratio at a relative threshold of 10% (τ) are reported to evaluate point maps and depth maps. Average aligned trajectory error (ATE) and area under curve at an error threshold of 30° (AUC) are reported to evluate camera pose estimation.

	Point Map		Camera Pose		Depth Map	
	rel ↓	τ ↑	ATE ↓	AUC ↑	rel ↓	τ ↑
$\alpha = 0.2$	0.143	0.607	0.022	42.116	0.137	0.606
$\alpha = 0.5$	0.137	0.626	0.023	46.703	0.129	0.628
$\alpha = 1.0$ (Reliev3R)	**0.122**	**0.663**	**0.018**	**49.426**	**0.115**	**0.657**
$\alpha = 2.0$	0.127	0.651	0.020	48.566	0.121	0.647

cropping as augmentation to FoV. The pseudo depth labels used to perform the experiment in Tab. 1 are generated by MoGe2 to keep consistency with pseudo intrinsics.

C. Ablation Study

In Eq.4 of the paper, we introduce an ambiguity-aware scale-invariant depth loss to regularize the shape of Reliev3R depth prediction while addressing the multi-view inconsistency in pseudo monocular depth labels. Intuitively, a large α encourages the training of Reliev3R to ignore multi-view inconsistency, and ends up with the confidence learned as a constant of 2 (the confidence value is restricted within $(0, 2)$). On the other hand, a small α results in degraded camera pose estimation and depth registration for disabling the depth shape regularization.

To evaluate the sensitivity of Reliev3R to α, we present ablation study on different α, as shown in Tab. 2. The results advocate the analysis above, and demonstrate that Reliev3R is robust against a large variation in α.

D. More Visualization for Reliev3R Prediction

We show the input, output and pseudo labels of Reliev3R in Fig. 1, Fig. 2, Fig. 3 and Fig. 4. It's observed that Reliev3R surpasses the pseudo monocular depth labels in case of multi-view depth consistency. And the learned confidence maps function well as we designed to mask out unreliable regions such as sky, contents in the far and edges of objects.

References

[1] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 1

[2] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*, 2025. 1

[3] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 1, 2

[4] Zhenggang Tang, Yuchen Fan, Dilin Wang, Hongyu Xu, Rakesh Ranjan, Alexander Schwing, and Zhicheng Yan. Mv-dust3r+: Single-stage scene reconstruction from sparse views in 2 seconds. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5283–5293, 2025. 1

[5] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 1

[6] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10510–10522, 2025. 1

[7] 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. 1

[8] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025. 1

[9] Felix Wimbauer, Weirong Chen, Dominik Muhle, Christian Rupprecht, and Daniel Cremers. Anycam: Learning to recover camera poses and intrinsics from casual videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16717–16727, 2025. 1

[10] Shangzhan Zhang, Jianyuan Wang, Yinghao Xu, Nan Xue, Christian Rupprecht, Xiaowei Zhou, Yujun Shen, and Gordon Wetzstein. Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21936–21947, 2025. 1

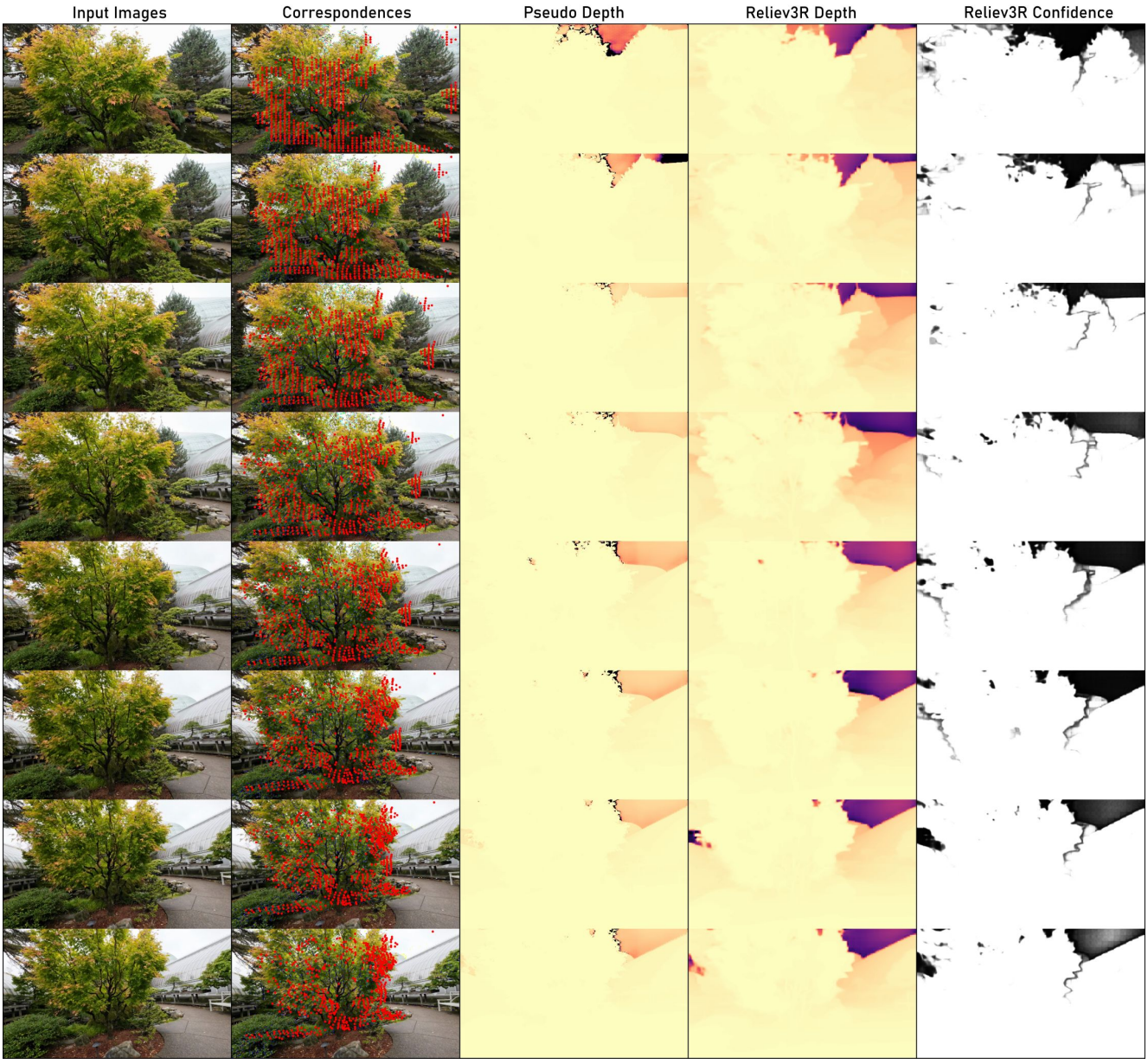

Figure 1. Visualization of input (multi-view images), output (multi-view depth maps and confidence maps) and pseudo labels (multi-view correspondences, monocular depth maps) of Reliev3R.

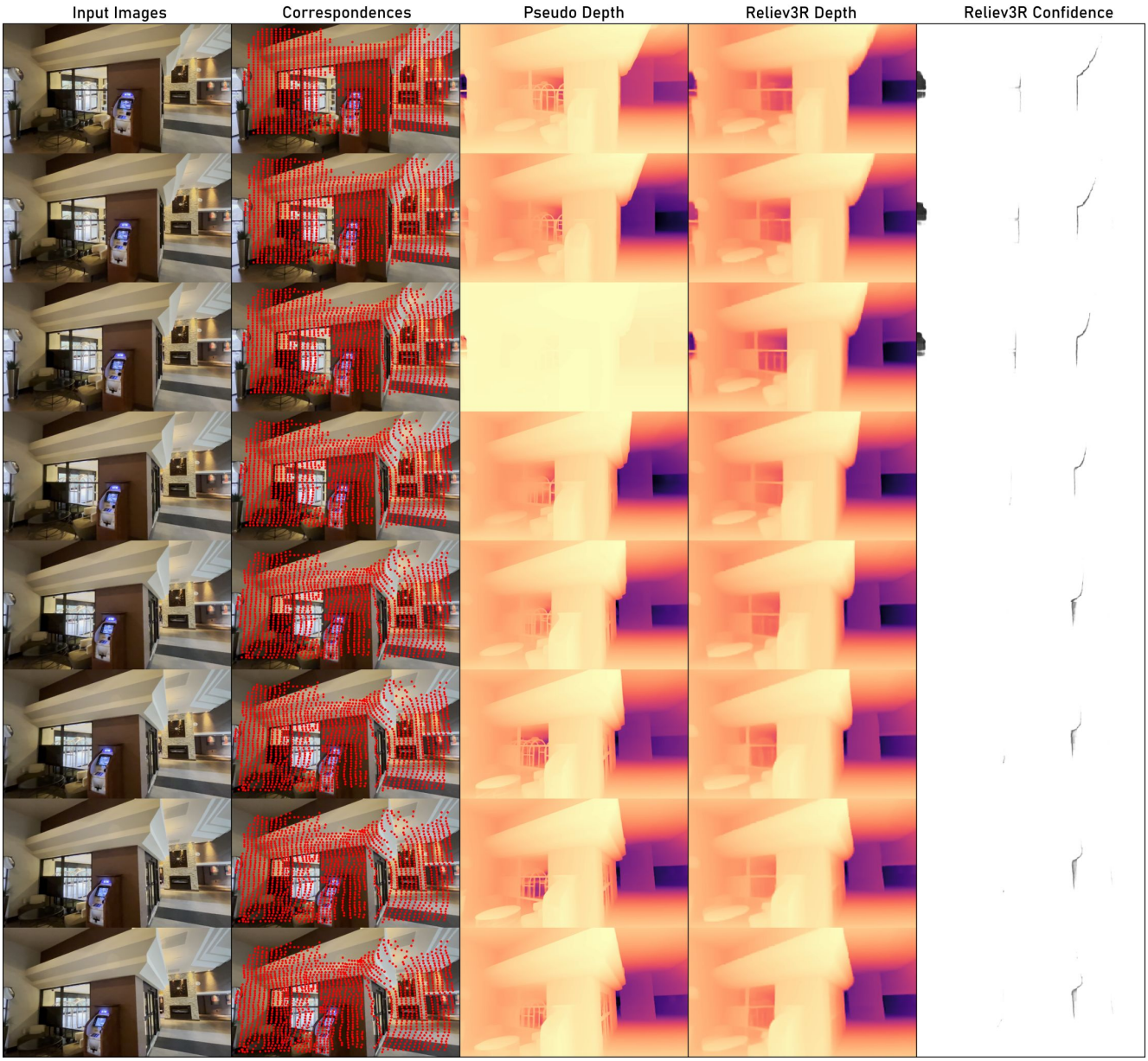

Figure 2. Visualization of input (multi-view images), output (multi-view depth maps and confidence maps) and pseudo labels (multi-view correspondences, monocular depth maps) of Reliev3R.

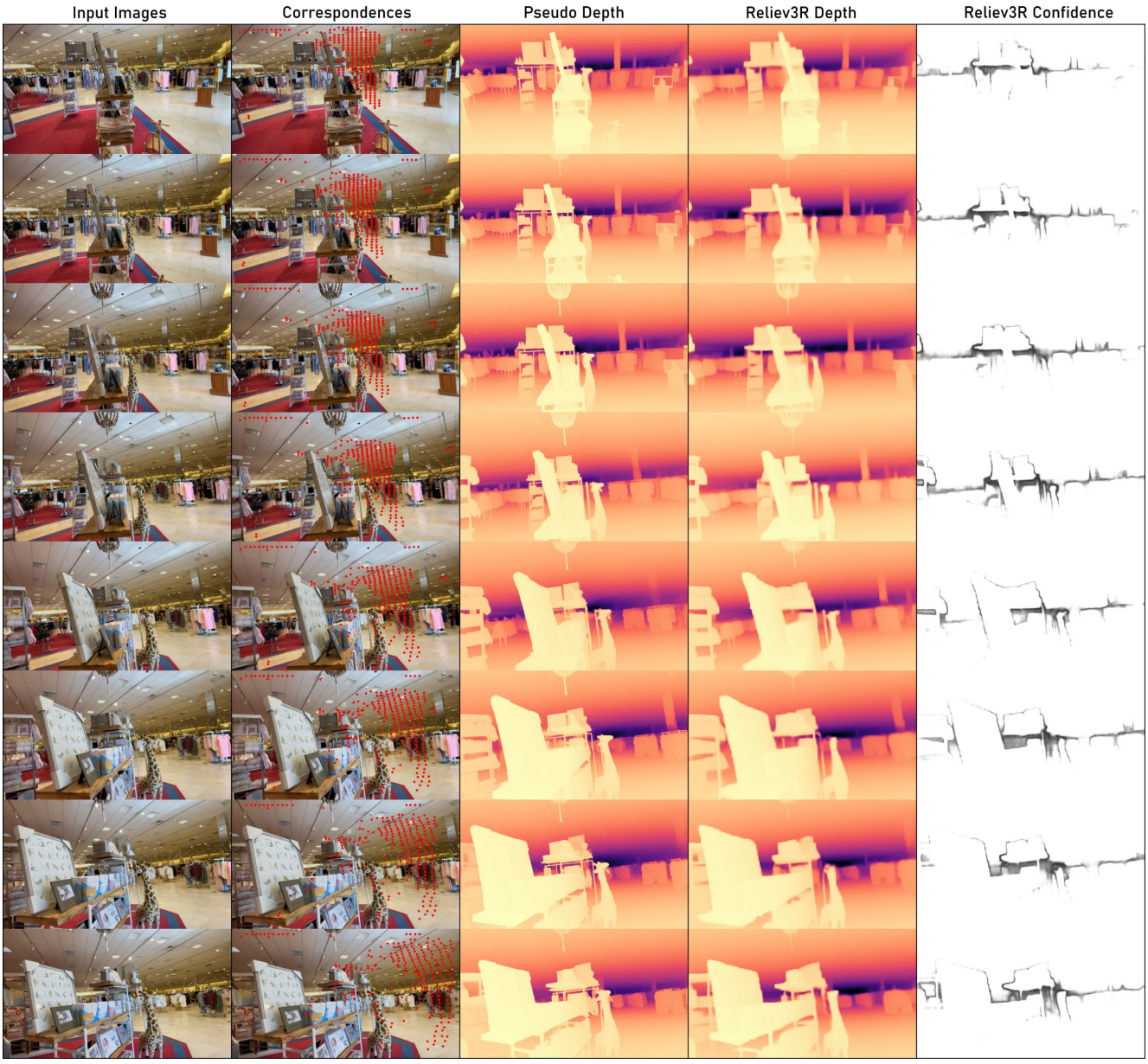

Figure 3. Visualization of input (multi-view images), output (multi-view depth maps and confidence maps) and pseudo labels (multi-view correspondences, monocular depth maps) of Reliev3R.

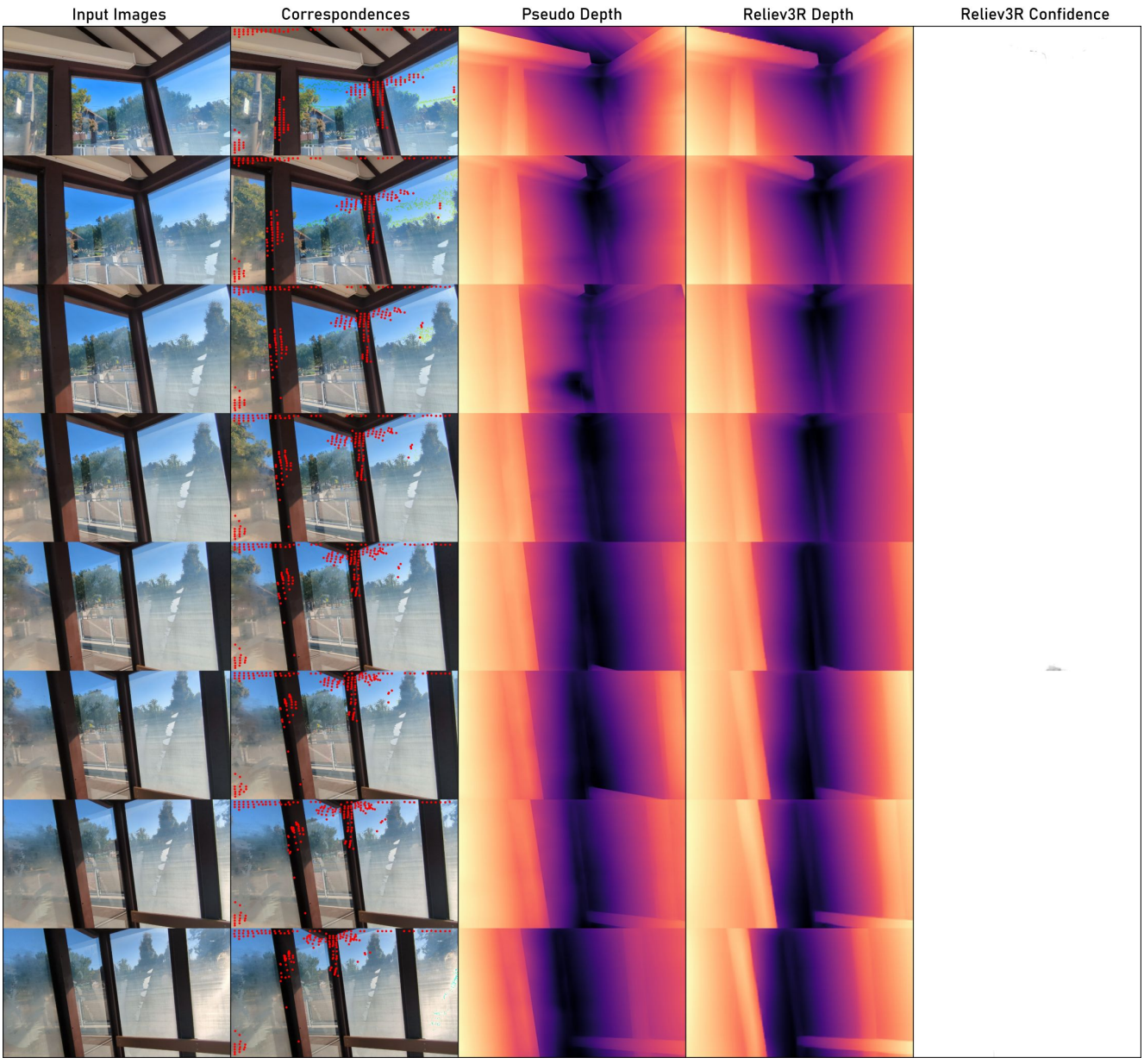

Figure 4. Visualization of input (multi-view images), output (multi-view depth maps and confidence maps) and pseudo labels (multi-view correspondences, monocular depth maps) of Reliev3R.