FlowMap: High-Quality Camera Poses, Intrinsics, and Depth via Gradient Descent —Supplemental Material—

Anonymous ECCV 2024 Submission

Paper ID #8115

Table of Contents

1 Additional Ablation Studies

In Tab. 1 we report *per-scene* as well as mean results of the ablation study. We further plot per-scene *qualitative* ablation studies in Fig. 1. As reported in the main paper, we find that our proposed reparameterizations of depth, pose, and focal length significantly outperform free-variable parameterizations.

1.1 Free-Variable Focal Length

Studying the per-scene ablation studies, we find that focal length exhibits a unique behavior relative to free-variable parameterizations of pose and depth.

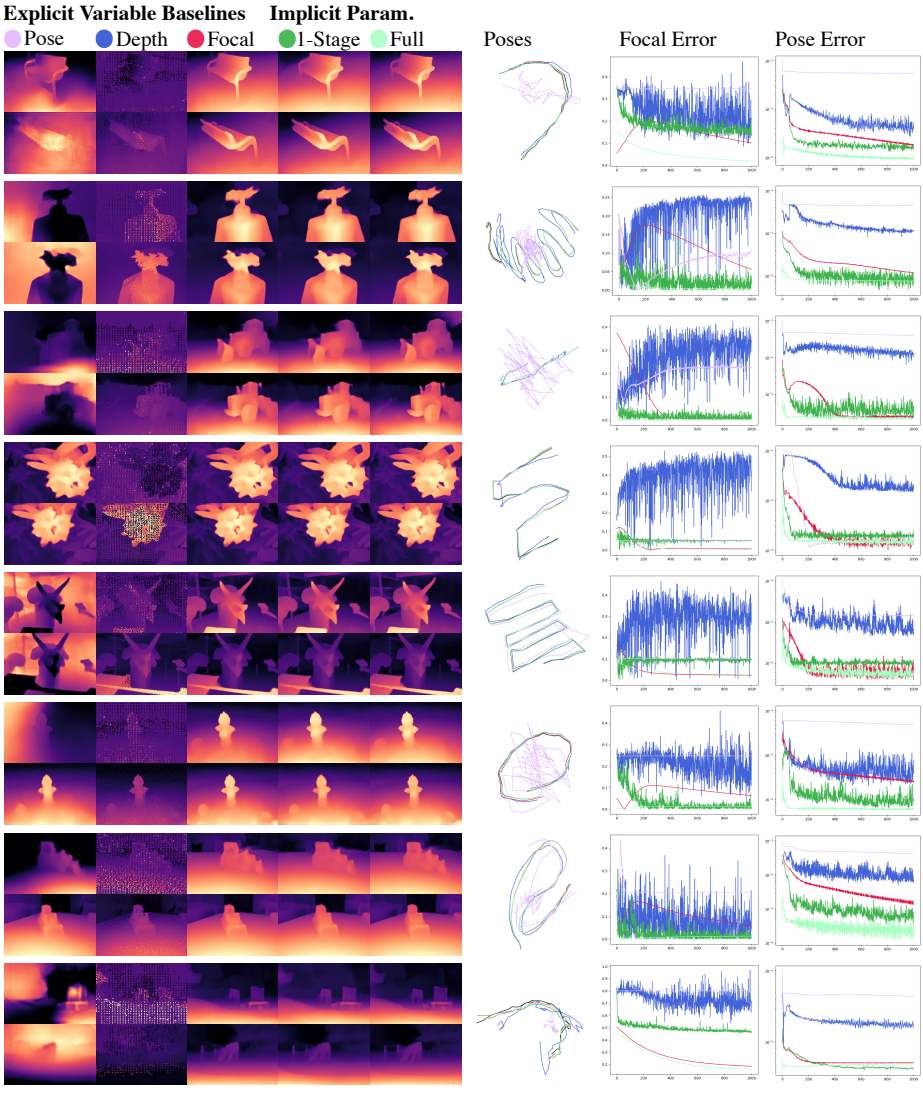

Fig. 1: Pose and Geometry Convergence for Free-Variable vs. Proposed Parameterizations. We plot poses, depths, focal lengths, and pose error (ATE) obtained with our proposed parameterizations ("Full") vs. those obtained with free-variable parameterizations at various optimization steps. With our proposed reparameterizations ("Full") as a baseline, we ablate either depth, focal length, or poses as free-variable optimizations and plot the resulting optimizations' pose and depth estimates. For instance, "Depth" corresponds to making the depth an explicit free-variable in the optimization. Using pose-as-variable and depth-as-variable often lead to "hollow-face" geometry, where the geometry is effectively inverted but still mostly satisfies the optical flow constraints. We also show results from a single-stage FlowMap pipeline, which only uses the implicit parameterization of intrinsics rather than switching to regressed intrinsics halfway through optimization.

	Caterpillar (T&T)			Bonsai (MipNeRF 360)			Hydrant (CO3D)			Horns (LLFF)		
Method	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
FlowMap	28.13	0.850	0.086	31.63	0.954	0.033	31.01	0.910	0.051	30.32	0.946	0.036
FlowMap (Single Stage)	27.45	0.829	0.098	32.25	0.959	0.029	29.38	0.872	0.071	30.19	0.941	0.035
Focal Var.	28.09	0.849	0.088	30.01	0.935	0.045	28.58	0.850	0.077	30.74	0.947	0.034
Depth Var.	17.18	0.456	0.416	12.84	0.356	0.606	19.46	0.388	0.309	26.28	0.827	0.100
Pose Var.	13.95	0.381	0.554	13.63	0.373	0.598	14.96	0.244	0.594	15.27	0.506	0.349

	Playground (T&T)			Kitchen (MipNeRF 360)			Bench (CO3D)			Flower (LLFF)		
Method	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
FlowMap	26.31	0.797	0.132	32.60	0.960	0.029	33.52	0.933	0.039	30.23	0.917	0.030
FlowMap (Single Stage)	23.38	0.702	0.202	32.14	0.957	0.030	32.67	0.920	0.045	29.75	0.909	0.034
Focal Var.	26.57	0.809	0.112	30.57	0.945	0.036	33.14	0.927	0.041	30.25	0.916	0.030
Depth Var.	17.35	0.485	0.409	19.45	0.512	0.226	26.38	0.816	0.101	19.52	0.536	0.253
Pose Var.	14.46	0.395	0.591	14.53	0.325	0.455	16.23	0.491	0.439	29.82	0.910	0.033

	Mean		
Method	PSNR ↑	SSIM ↑	LPIPS ↓
FlowMap	30.47	0.908	0.054
FlowMap (Single Stage)	29.65	0.886	0.068
Focal Var.	29.75	0.897	0.058
Depth Var.	19.81	0.547	0.302
Pose Var.	16.61	0.453	0.452

Table 1: Ablations. We present per-scene and mean ablation results.

While pose and depth perform worse across the board, on some scenes, treating focal length as a free variable performs as well as our proposed focal length solver. The fact that explicitly optimizing focal length is more well-posed than explicitly optimizing depths and poses is perhaps unsurprising since doing so only involves a single scalar for the whole video. However, notably, on about one out of five scenes that we tried, the free-variable formulation performs noticeably worse than our implicit parameterization (about 2dB PSNR worse on downstream Gaussian Splats). We hypothesize that a free-variable optimization is tractable when the initialized focal length is "close enough" to ground truth, but may fall into local minima otherwise. More importantly, gradient-descent based optimization of the focal length makes it impossible to pre-train FlowMap or, more generally, use FlowMap in a generalizable setting. The proposed implicit focal length parameterization solves for the focal length *in the forward pass* and thus uniquely enables pre-training and generalization. Additionally, the explicit focal optimization typically takes significantly longer to converge; see Fig. 1 for convergence comparisons. Therefore, we use the implicit focal parameterization for pre-training, and for per-scene optimization, we combine the robustness of our implicit parameterization and the precision of the focal-variable optimization (as discussed in the main paper). Specifically, our full model first optimizes the fully implicit prediction, and in a second stage, optimizes the focal length as a free variable, initialized with the converged implicit estimate.

1.2 Correspondence MLP

We don't formally ablate the correspondence MLP, as we do not claim it as a contribution. It was introduced in FlowCam [2], which ran an in-depth ablation study that confirmed its necessity.

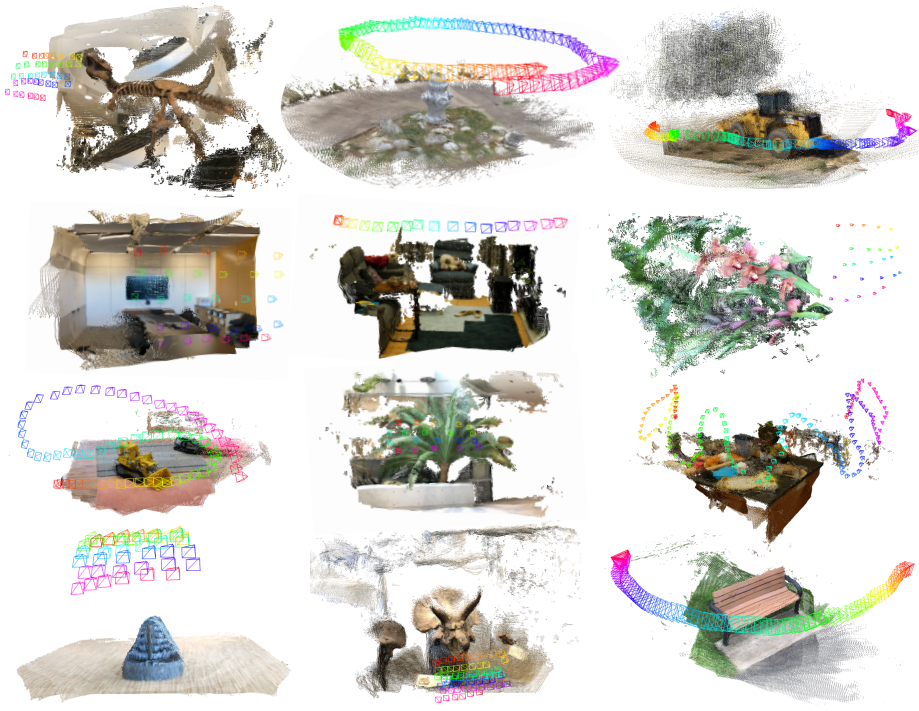

Fig. 3: Additional Point Clouds Here we plot additional point clouds across the Tanks and Temples, LLFF, Mip-NeRF360, and CO3D datasets.

2 Additional Results

2.1 Pre-Trained Depth vs. Fine-Tuned Depth vs. High-Resolution Fine-Tuned Depth.

In Fig. 4, we compare the depths produced by FlowMap's initialization to the depths produced after FlowMap optimization. We additionally compare these results to a MiDaS CNN fine-tuned at a significantly higher resolution. We find that per-scene fine-tuning leads to high-quality depth predictions. This is illustrated by Fig. 3, which demonstrates FlowMap's ability to generate high-quality, consistent depths. However, it is worth noting that FlowMap's off-the-shelf depths are slightly blurry. To investigate whether this is a limitation of our loss or the architecture of the depth-predicting CNN, we also perform optimization at a higher resolution. We find that this leads to crisp depth maps, demonstrating that blurry depth maps are a result of insufficient capacity of the MiDaS backbone and not a limitation of our camera-induced flow loss. Notably, the poses barely change in this fine-tuning stage. It is likely that replacing the MiDaS depth predictor with a more powerful depth backbone would lead to sharper depth without high-resolution fine-tuning.

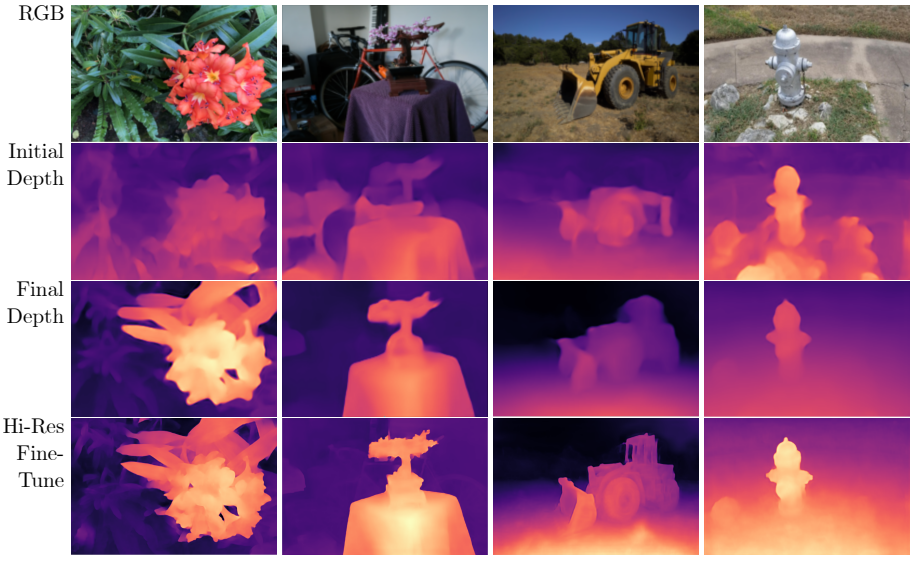

Fig. 4: Depth Estimates Before and After Optimization. The depth prediction neural network can either be randomly initialized or pre-trained, though pre-trained depth networks lead to much faster convergence. In the second row, we show the output of the depth prediction neural network after pre-training it on a dataset consisting of CO3D, KITTI, and RealEstate10k. These estimates converge to high-quality depth within only a few hundred FlowMap optimization steps. We see that the quality of the initial, pre-trained depth predictions is not critical to achieve accurate reconstructions. Although we estimate geometry at a lower resolution during optimization to manage memory constraints, we can quickly fine-tune at high-resolution for more detailed depth maps if necessary (bottom row).

2.2 Additional Point Clouds and Qualitative Pose Reconstructions

In Fig. 3, we display 12 additional point clouds plus estimated camera poses across popular datasets and scenes across the LLFF, Tanks and Temples, Mip-NeRF 360, and CO3D datasets. FlowMap robustly recovers camera poses and scene geometry across these diverse, challenging, and real-world sequences.

2.3 Failure Cases

We ran FlowMap on 25 more scenes and observed 3 soft failures. Some of these failures include the Tanks-and-Temples Auditorium scene (our model struggles with rotation-dominant trajectories), the LLFF Leaves scene (our model falls into a "hollow-face minimum"), and the Tanks-and-Temples Lighthouse scene (this video features a large lens flare which degrades the optical flow). Future extensions to FlowMap could use an occlusion-aware formulation to avoid hollow-face minima.

3 Implementation Details

3.1 Procrustes Solver Details

Our pose solver is the one introduced in FlowCam [2]; see [2] for details. The only difference is that instead of selecting 1000 random points for the Procrustes estimation, we fix the points (uniformly spaced throughout the image) when performing per-scene overfitting. We find that fixing the points used for the pose solver allows the network to better overfit confidence weights and subsequently yields better poses.

3.2 Intrinsics Solver Details

For the intrinsics solver, we assume a pinhole camera estimate and discretize a set of 60 candidate focal lengths between .5 and 2 (in resolution-independent units). We use a softmin on the flow error maps, as discussed in the main paper. We scale the error maps by a temperature factor of 10 and weight the error maps by the flow confidence weights. See Fig. 6 for illustration.

3.3 Depth NN (MiDaS) details

For our depth network, we use the lightweight CNN version of MiDaS [1], pre-trained with the publicly available weights trained on relative-depth estimation.

3.4 Correspondence Weight MLP

The correspondence weight MLP is a three-layer MLP with ReLU activations and 128 hidden units per layer. It takes as input two corresponding image features and outputs a per-correspondence weight between 0 and 1, regularized by a sigmoid. Here we use intermediate feature maps from the depth network as the image features. These weights are used in the weighted Procrustes pose solver.

4 Experiment Details

4.1 Image Resolution

To manage computational cost (our current implementation loads the entire video into memory), we use a resolution of 128x192 for network optimization, and 672x1024 for downstream Gaussian Splatting, the optical flow, and point track predictions.

4.2 Hyperparameters

We train for 1000 steps using Adam and use a learning rate of 1e-4. For the pose-as-variable experiments, we choose Euler angles as the parameterization of the rotation matrix.

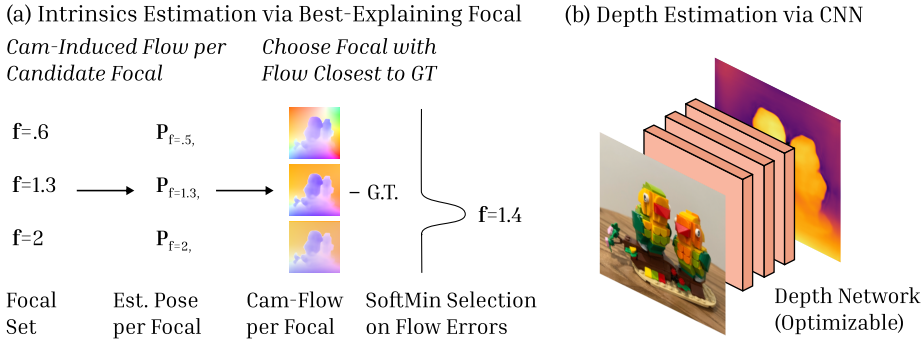

Fig. 6: In (a) we illustrate our implicit focal length formulation, which considers a set of candidate focal lengths, assigns each one an error score, and softly selects the focal length with the lowest error. To calculate the error score for a focal length, we use that focal length to estimate a pose, and then compare the resulting pose-induced optical flow to the ground truth optical flow. In (b) we illustrate that we parameterize depth via the output of a monocular depth prediction CNN.

4.3 Pre-Training Details

Before performing per-scene fine-tuning, we found it useful to learn a large-scale prior for better initialization. We use the same FlowMap loss formulation but train it on datasets of videos (instead of optimizing on a single scene). We use videos from CO3D, RealEstate-10K, and KITTI for pretraining. Note that we only use the raw videos from these datasets (no intrinsics, poses, or sparse geometry).

5 Limitations

Wile our method is much faster than MVS COLMAP, it is about 30 percent slower than COLMAP at its highest quality setting (on long sequences, about 20 minutes for our method vs. 14 minutes for COLMAP). It additionally requires signficantly more GPU memory than COLMAP does. Our method's pose and intrinsics predictions are less accurate and robust than COLMAP's, as measured by ATE, though after Gaussian Splatting with fine-tuning of camera parameters, we often perform on par with COLMAP.

Our method further depends on correspondences estimated by point tracks and optical flow. While existing methods for computing point tracks and optical flow are robust, failures sometimes occur, and these failures can affect FlowMap's accuracy if they are significant. On the other hand, FlowMap will directly improve alongside advancements in these domains.

Finally, our method is constrained to work on frame sequences with significant overlap (i.e., videos) and fails when input sequences contain significant scene motion. The latter limitation is shared with COLMAP, though we hope

that our method may serve as a step towards novel methods that address this shortcoming.

References

1. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020) 6
2. Smith, C., Du, Y., Tewari, A., Sitzmann, V.: Flowcam: Training generalizable 3d radiance fields without camera poses via pixel-aligned scene flow. Advances in Neural Information Processing Systems (NeurIPS) (2023) 3, 6