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

Mateus Barbosa - Reviewer
Mohara Nascimento - Archaeologist
Veronika Treumova - Hacker
Mateus Barbosa - PhD Student

FlowMap

Reviewer - Mateus Barbosa

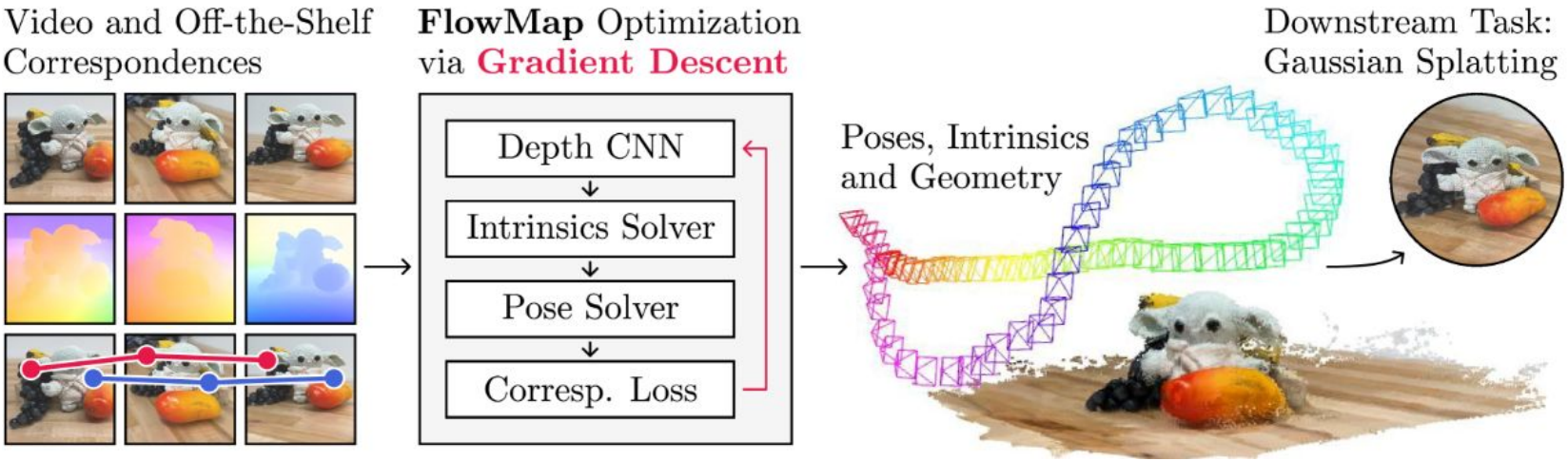
Introduction

Today, essentially all SotA 3D reconstruction methods are based on top of SfM methods like COLMAP.

Good results, but is not differentiable w.r.t. its free variables (camera poses, camera intrinsics and per-pixel depths)

Flowmap is an end-to-end differentiable method that solves for precise camera poses, camera intrinsics, and per-frame dense depth of a video sequence.

Flowmap



FlowMap: High-Quality Camera Poses, Intrinsics, and Depth via Gradient Descent (2024)

Supervision via Camera-Induced Flow

Given a video sequence, the goal is to supervise per-frame estimates of depth, intrinsics, and pose using known correspondences.

That will be done using the optical flow induced by the camera movement through the scene.

The known correspondences are derived from two sources: 1) dense optical flow between adjacent frames and 2) sparse point tracks which span longer windows.

Optical flow loss

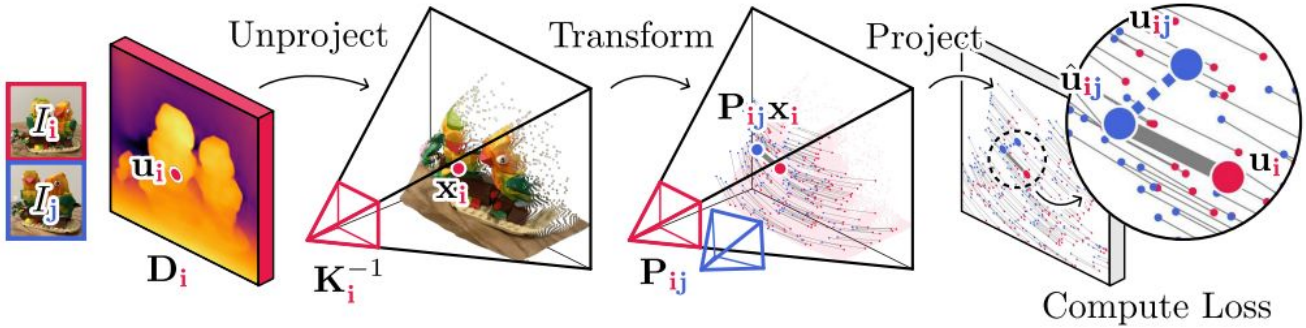
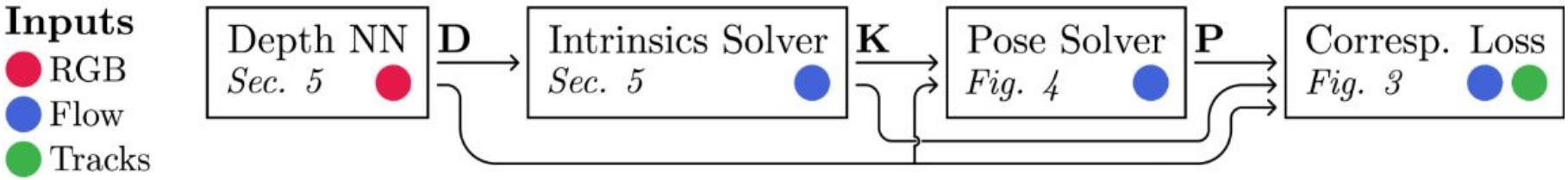


Fig. 3: Camera-Induced Flow Loss. To use a known correspondence $\mathbf{u}_{ij}$ to compute a loss $\mathcal{L}$, we unproject $\mathbf{u}_i$ using the corresponding depth map $\mathbf{D}_i$ and camera intrinsics $\mathbf{K}_i$, transform the resulting point $\mathbf{x}_i$ via the relative pose $\mathbf{P}_{ij}$, reproject the transformed point to yield $\hat{\mathbf{u}}_{ij}$, and finally compute $\mathcal{L} = \|\hat{\mathbf{u}}_{ij} - \mathbf{u}_{ij}\|$.

Parametrizing depth, pose and camera intrinsics



Depth neural network

Depth is parametrized as a neural network that maps an RGB frame to the corresponding per-pixel depth.

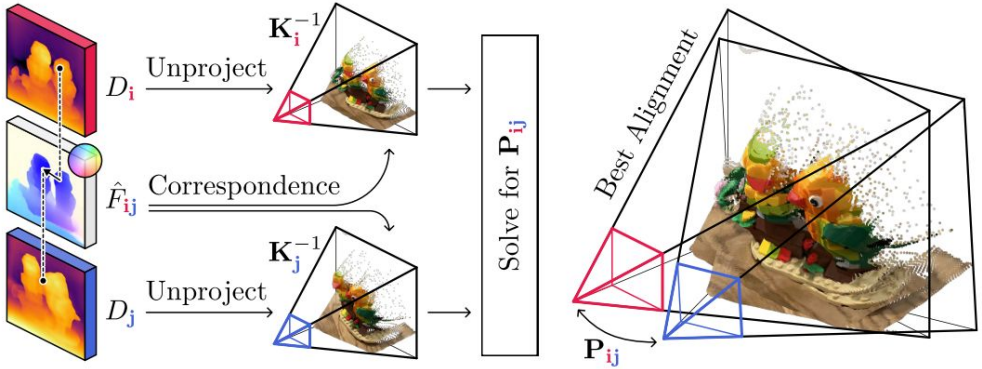
This ensures that similar patches have similar depths, allowing FlowMap to integrate geometry cues across frames: if a patch receives a depth gradient from one frame, the weights of the depth network are updated, and hence the depths of all similar video frame patches are also updated.

Intrinsics as a function of depth and optical flow

Camera intrinsics is solved by considering a set of reasonable candidates.
Loss function is calculated considering the pose calculated by K_k.
Intrinsics K is computed via softmin-weighted sum of the candidates.

$$\mathbf{K} = \sum_k w_k \mathbf{K}_k \qquad w_k = \frac{\exp(-\mathcal{L}_k)}{\sum_l \exp(-\mathcal{L}_l)}$$

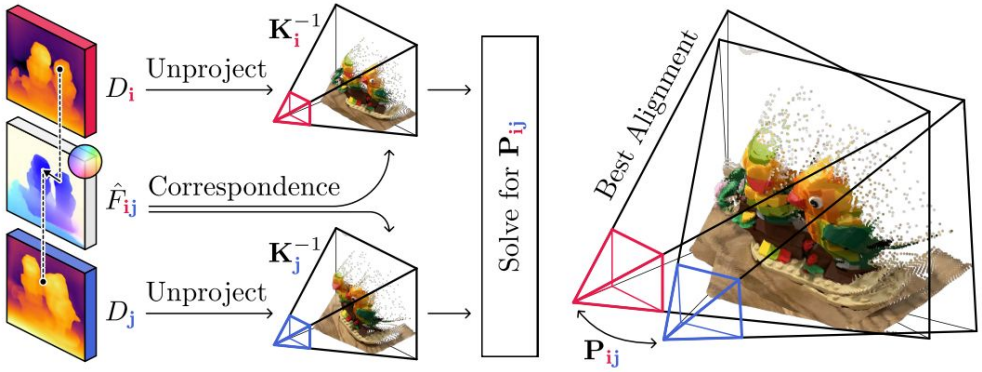
Pose as a function of depth, intrinsics and optical flow



Depth maps D_i and D_j are unprojected with their respective intrinsics to generate point clouds $\mathbf{X}_i$ and $\mathbf{X}_j$.
Points in the clouds are matched using the known optical flow, generating $\mathbf{X}_i^{\leftrightarrow}$ and $\mathbf{X}_j^{\leftrightarrow}$
The diagonal matrix $\mathcal{W}$ contains correspondence weights that can down-weight correspondences that are faulty due to occlusion or imprecise flow.

$$\mathbf{P}_{ij} = \underset{\mathbf{P}\in \mathrm{SE}(3)}{\arg\min} \|\mathcal{W}^{1/2}(\mathbf{X}_j^{\leftrightarrow} - \mathbf{P}\mathbf{X}_i^{\leftrightarrow})\|_2^2$$

Pose as a function of depth, intrinsics and optical flow



$$\mathbf{P}_{ij} = \underset{\mathbf{P}\in \mathrm{SE}(3)}{\arg\min} \|\mathcal{W}^{1/2}(\mathbf{X}_j^{\leftrightarrow} - \mathbf{P}\mathbf{X}_i^{\leftrightarrow})\|_2^2$$

$$\mathbf{R} = \mathbf{U}\mathbf{S}\mathbf{V}^T \text{ and } \mathbf{t} = (\mathbf{X} - \mathbf{R}\mathbf{X}')\mathbf{W}\mathbf{1}, \text{ where } \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathrm{SVD}(\Sigma_{\mathbf{X}'\mathbf{X}}),$$

$$\Sigma_{\mathbf{X}'\mathbf{X}} = \mathbf{X}\mathbf{K}\mathbf{W}\mathbf{K}\mathbf{X}'^T, \mathbf{K} = \mathbb{I} - \sqrt{\mathbf{w}}\sqrt{\mathbf{w}}^T, \text{ and } \mathbf{S} = \mathrm{diag}(1, ..., \det(\mathbf{U})\det(\mathbf{V})).$$

Novel view synthesis results

	MipNeRF 360 (3 scenes)					LLFF (7 scenes)				
Method	PSNR ↑	SSIM ↑	LPIPS ↓	Time (min.) ↓	ATE	PSNR ↑	SSIM ↑	LPIPS ↓	Time (min.) ↓	ATE
FlowMap	29.84	0.916	0.073	19.8	0.00055	27.23	0.849	0.079	7.5	0.00209
COLMAP	29.95	0.928	0.074	4.8	N/A	25.73	0.851	0.098	1.1	N/A
COLMAP (MVS)	31.03	0.938	0.060	42.5	N/A	27.99	0.867	0.072	13.4	N/A
DROID-SLAM*	29.83	0.913	0.066	0.6	0.00017	26.21	0.818	0.094	0.3	0.00074
NoPE-NeRF*	13.60	0.377	0.750	1913.1	0.04429	17.35	0.490	0.591	1804.0	0.03920
	Tanks & Temples (14 scenes)					CO3D (2 scenes)				
Method	PSNR ↑	SSIM ↑	LPIPS ↓	Time (min.) ↓	ATE	PSNR ↑	SSIM ↑	LPIPS ↓	Time (min.) ↓	ATE
FlowMap	27.00	0.854	0.101	22.3	0.00124	31.11	0.896	0.064	22.1	0.01589
COLMAP	26.74	0.848	0.130	5.5	N/A	25.17	0.750	0.190	12.6	N/A
COLMAP (MVS)	27.43	0.863	0.097	51.4	N/A	25.35	0.762	0.175	52.0	N/A
DROID-SLAM*	25.70	0.824	0.133	0.8	0.00122	25.97	0.790	0.139	0.8	0.01728
NoPE-NeRF*	13.38	0.449	0.706	2432.9	0.03709	14.97	0.400	0.770	2604.9	0.03648

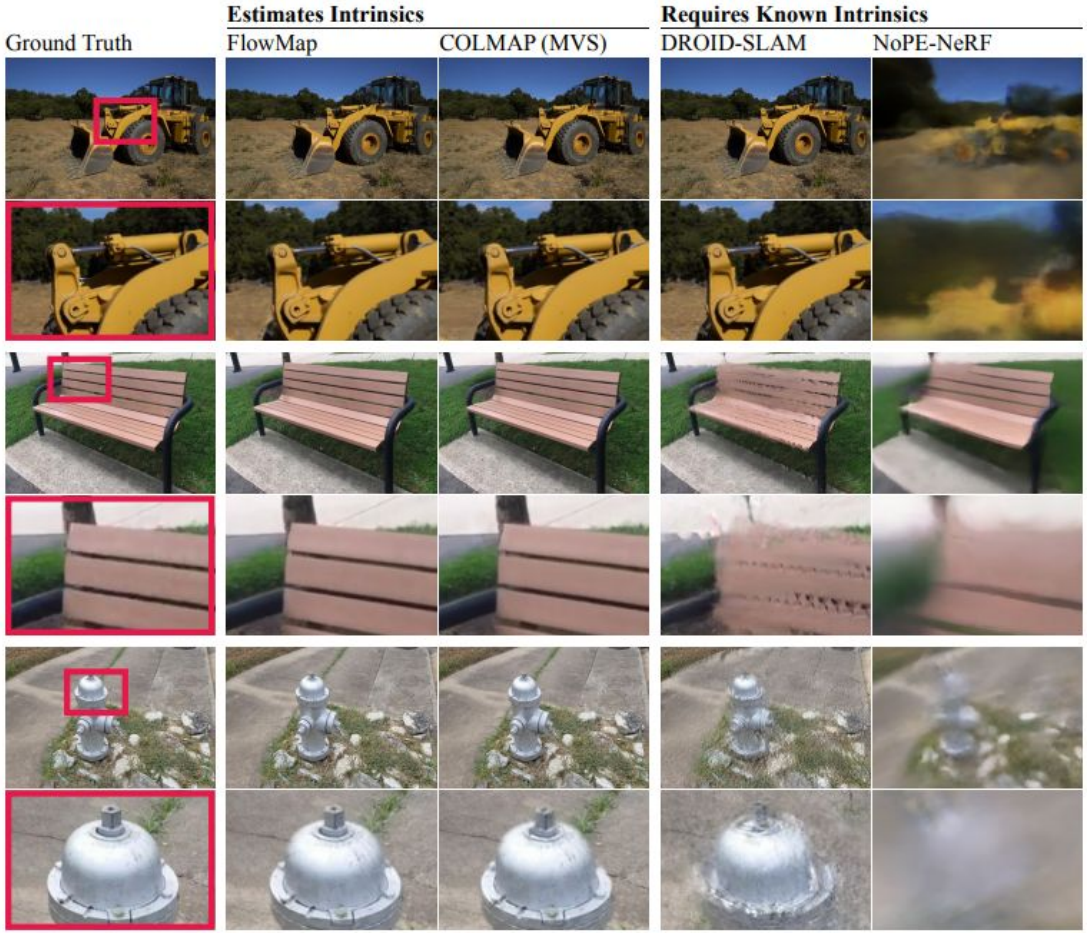
Comparison to other methods

Method	25 views		50 views		100 views		200 views		full	
	ATE↓	Reg.↑	ATE↓	Reg.↑	ATE↓	Reg.↑	ATE↓	Reg.↑	ATE↓	Reg.↑
COLMAP [46]	0.03840	44.4	0.02920	60.5	0.02640	85.7	0.01880	97.0	-	-
ACE-Zero [9]	0.11160	100.0	0.07130	100.0	0.03980	100.0	0.01870	100.0	0.01520	100.0
FlowMap [50]	0.10700	100.0	0.07310	100.0	0.04460	100.0	0.02420	100.0	N/A	66.7
VGGSfM [62]	0.05800	96.2	0.03460	98.7	0.02900	98.5	N/A	47.6	N/A	0.0
DF-SfM [20]	0.08110	99.4	0.04120	100.0	0.02710	99.9	N/A	33.3	N/A	76.2
MASt3R-SfM	0.03360	100.0	0.02610	100.0	0.01680	100.0	0.01300	100.0	0.01060	100.0

Method	MIP-360	LLFF	T&T	CO3Dv2
NoPE-NeRF [8]	0.04429	0.03920	0.03709	0.03648
DROID-SLAM [54]	0.00017	0.00074	0.00122	0.01728
FlowMap [50]	0.00055	0.00209	0.00124	0.01589
ACE-Zero [9]	0.00173	0.00396	0.00973	0.00520
MASt3R-SfM	0.00079	0.00098	0.00215	0.00538

Table 1: **Results on Tanks&Temples** in terms of ATE and overall registration rate (Reg.). For easier readability, we color-code ATE results as a linear gradient between worst and best ATE for a given dataset or split; and Reg results with linear gradient between 0% and 100%. **Left**: impact of the number of input views, regularly sampled from the full set. 'N/A' indicates that at least one scene did not converge. **Right**: ATE↓ on different datasets with the arbitrary splits defined in FlowMap [50].

Results from: Duisterhof, B. et al. (2024). **MASt3R-SfM: a Fully-Integrated Solution for Unconstrained Structure-from-Motion**. *arXiv preprint arXiv:2409.19152*.



FlowMap: High-Quality Camera Poses, Intrinsics, and Depth via Gradient Descent (2024)

Camera parameter estimation

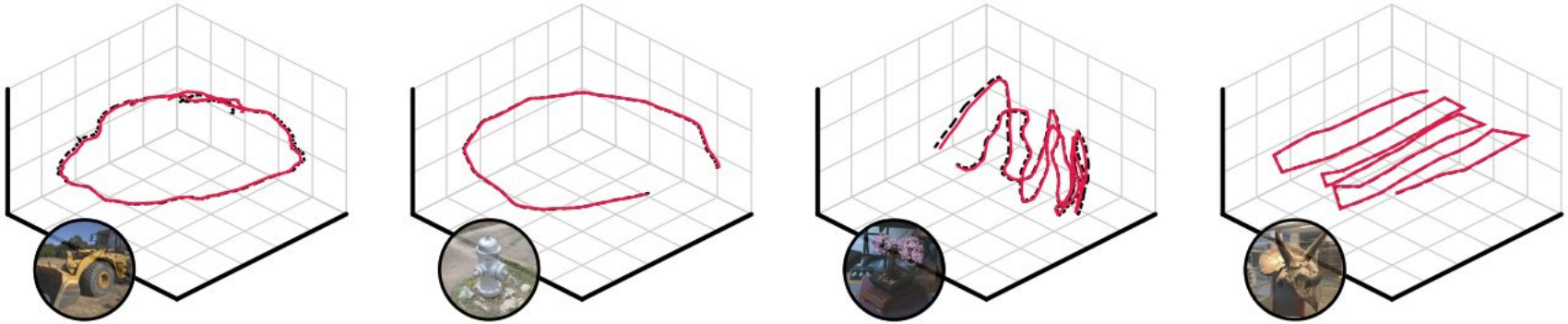


Fig. 6: Qualitative Pose Estimation Comparison. FlowMap (solid red) recovers camera poses that are very close to those of COLMAP (dotted black).

Camera parameter estimation

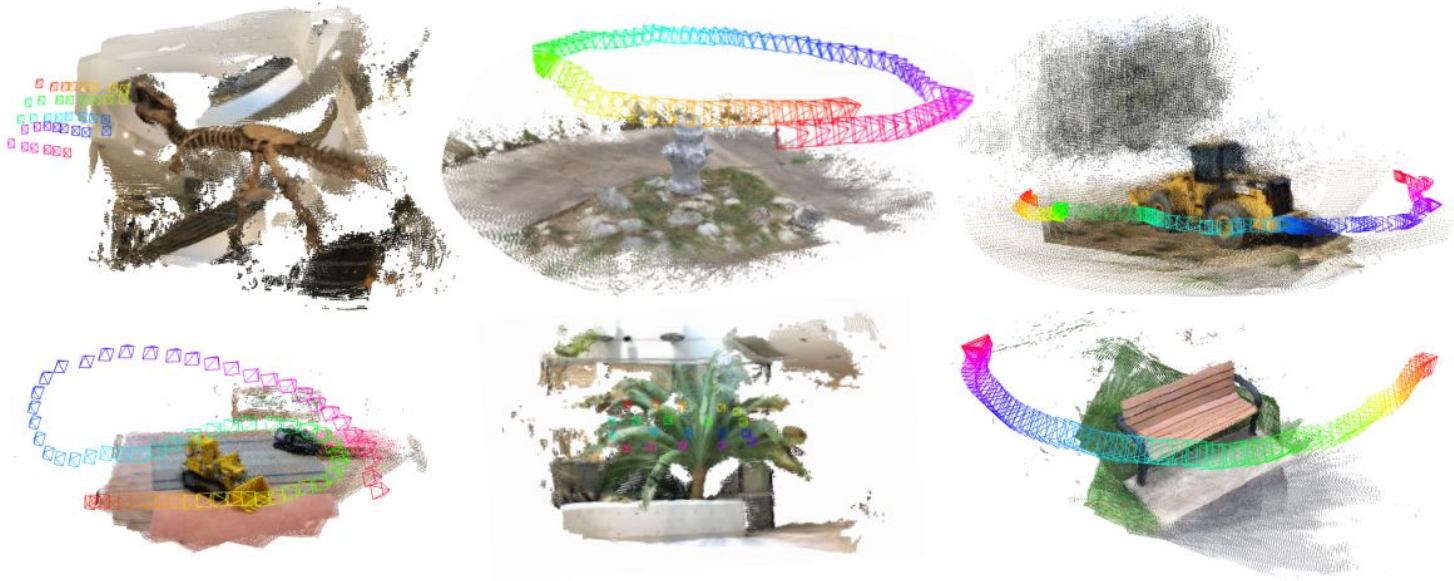


Fig. 7: Point Clouds Reconstructed by FlowMap. Unprojecting FlowMap depths using FlowMap's intrinsics and poses yields dense and consistent point clouds.

Large-scale robustness study

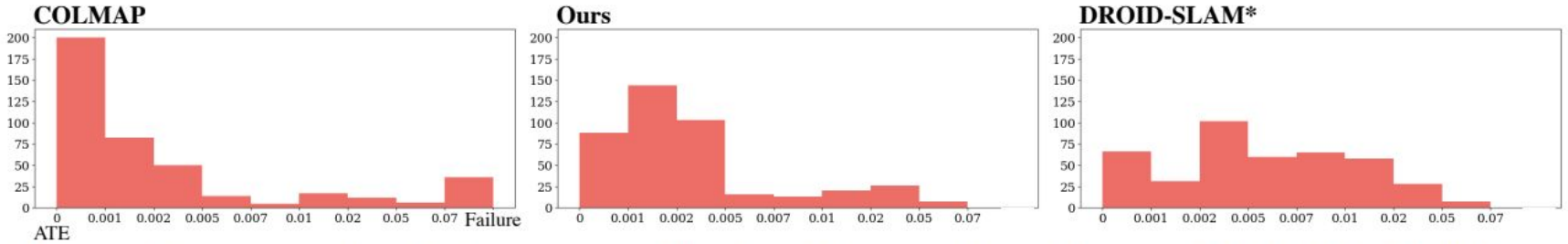
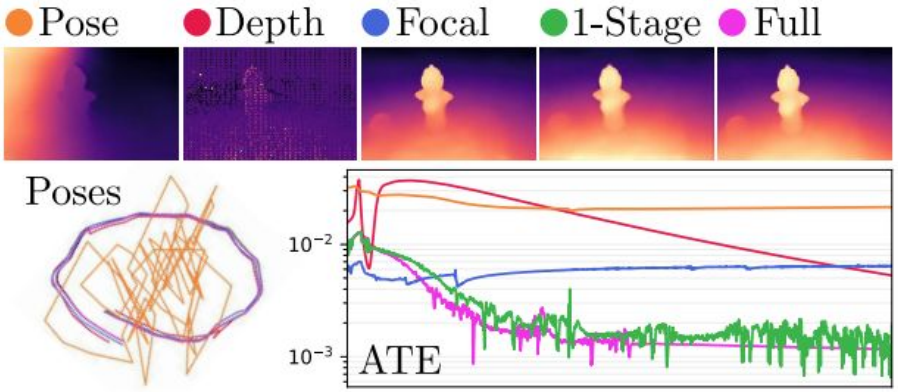


Fig. 8: Large-scale Robustness Study. We run FlowMap and DROID-SLAM on 420 CO3D scenes across 10 categories and plot mean ATEs with respect to CO3D's COLMAP-generated pose metadata. We also re-run COLMAP on the same data. Compared to DROID-SLAM, which requires ground-truth intrinsics, FlowMap produces notably lower ATEs. FlowMap's ATE distribution is similar to one obtained by re-running COLMAP, with most ATEs falling under 0.005 in both cases.

Ablations

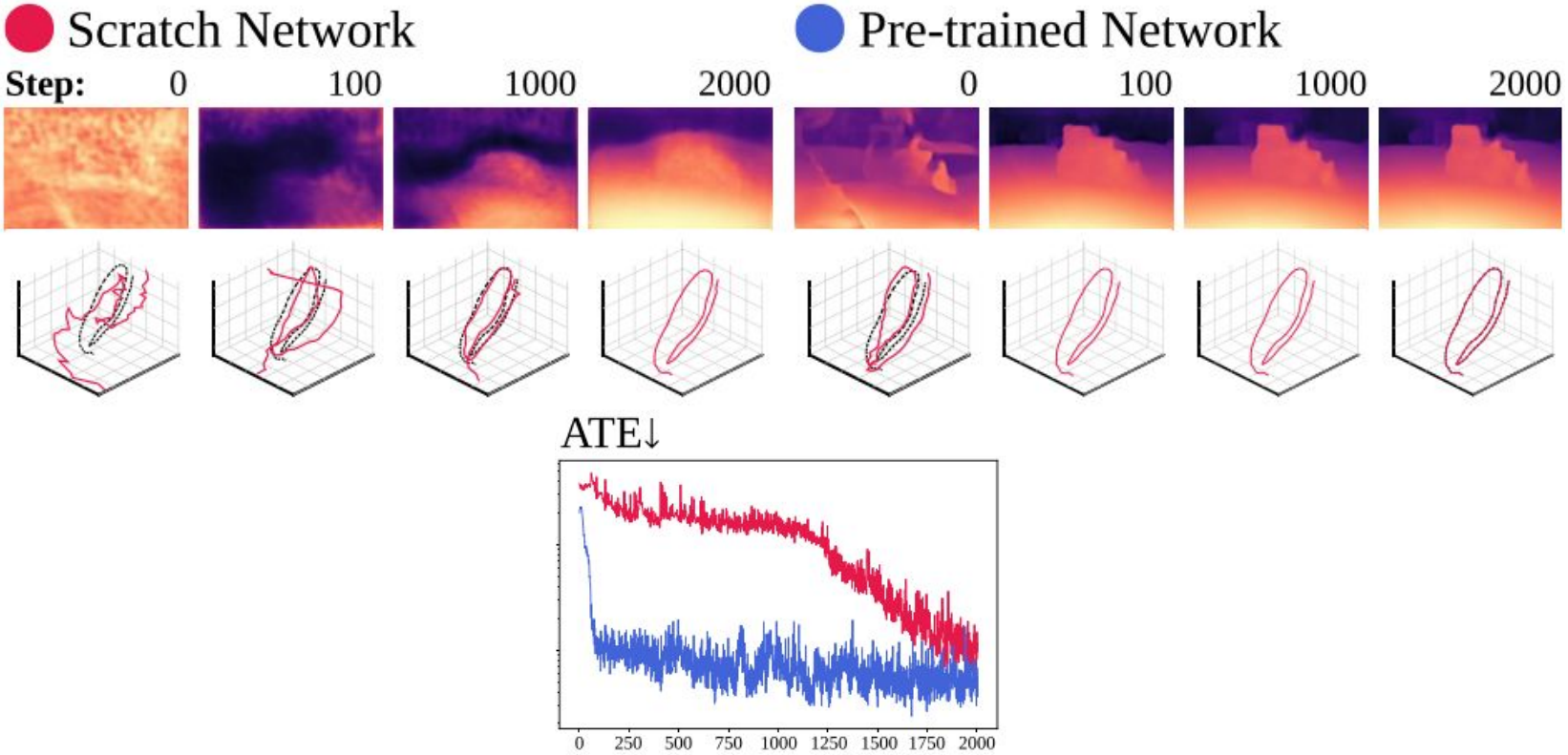
Why not free variables?

Are point tracks necessary?



Method	PSNR ↑	SSIM ↑	LPIPS ↓
FlowMap	27.70	0.863	0.089
Single Stage	26.66	0.842	0.112
Expl. Focal Length	25.15	0.788	0.141
Expl. Depth	8.84	0.168	0.684
Expl. Pose	16.00	0.533	0.495
No Point Tracks	25.83	0.822	0.122

Pre-trained depth networks?



Strengths

- Novel view synthesis results on par with COLMAP
- End-to-end differentiable design
- Closed form solution to pose and intrinsics estimation
- New approach that differs from traditional SfM methods

Weaknesses

- Compared to COLMAP:
  - is slower
  - requires significantly more GPU memory
  - pose and intrinsics are less accurate and less robust
- Results compared to other contemporary methods are lackluster
- Is constrained to work on frame sequences with significant overlap (i.e., videos).
- Important details of the paper are hidden away
- Dependent of many off-the-shelf components

In conclusion

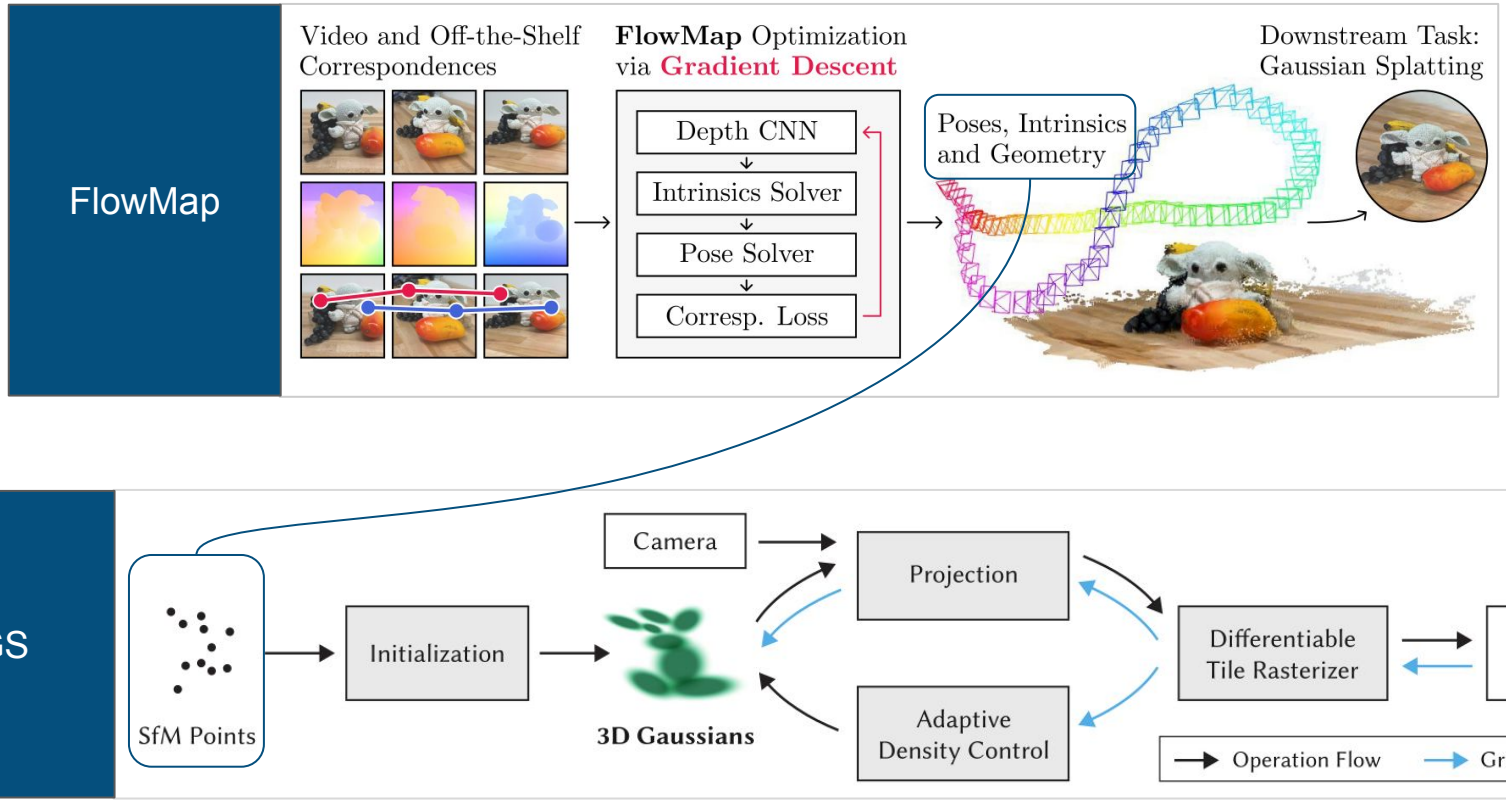
The paper should be rejected.

- Achievements aren’t as impressive as the authors seem to suggest.
- NVS results are on-par with COLMAP only in limited scenarios.
- Other concurrent paper seem to perform better on a wider set of scenarios.
- Methods and implementation aren’t sufficiently well explained.

FlowMap

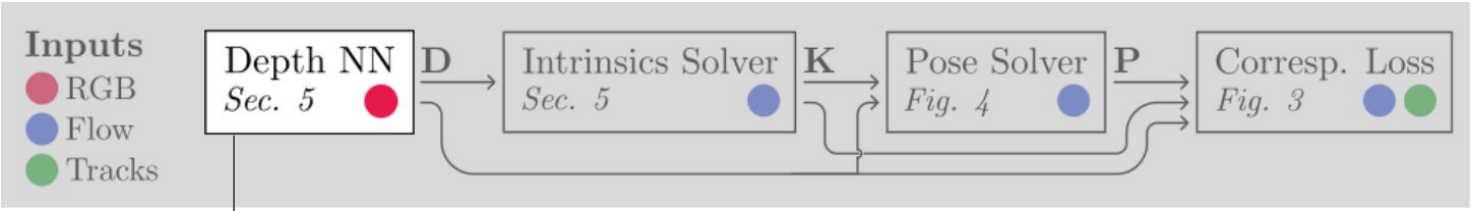
Archaeologist - Mohara Nascimento

Summary



FlowMap: High-Quality Camera Poses, Intrinsics, and Depth via Gradient Descent (2024)

Previous and basement works



Depth is parameterized via a neural network

- FlowMap uses the lightweight CNN version of MiDaS in their depth network

INPUT

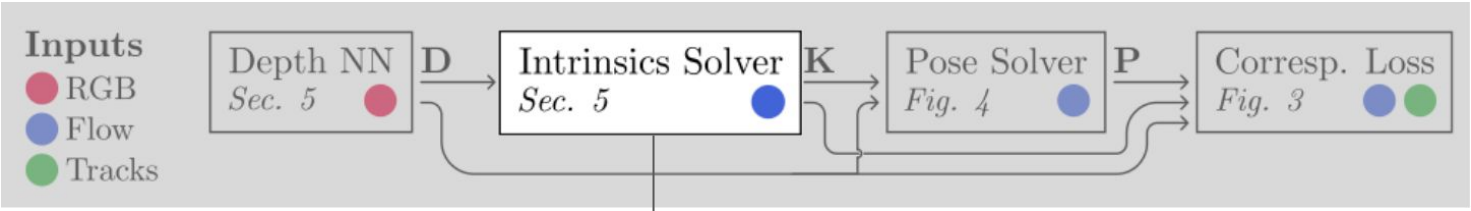
RGB images and depth annotation

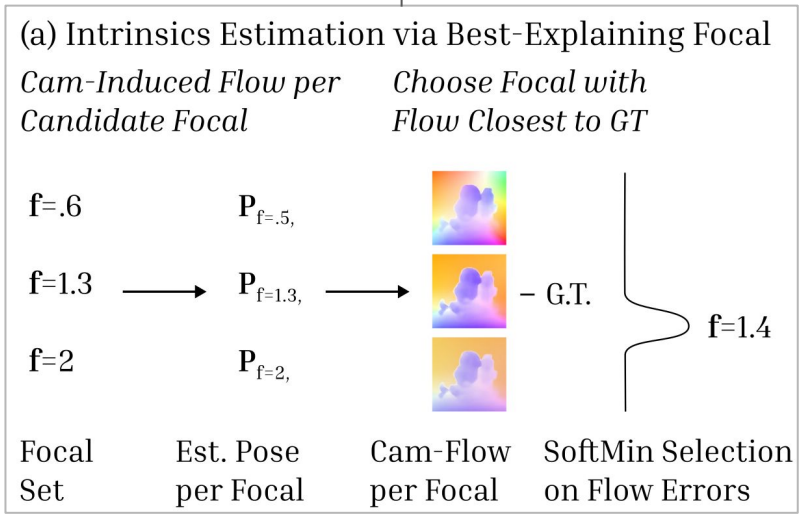
OUTPUT

Inverse depth maps

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).

Previous and basement works



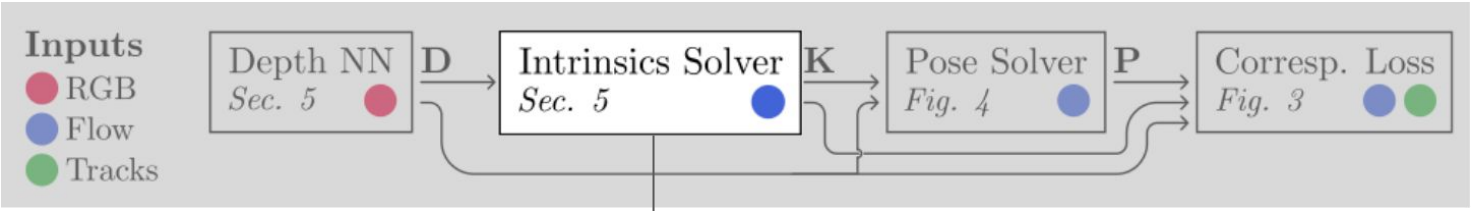


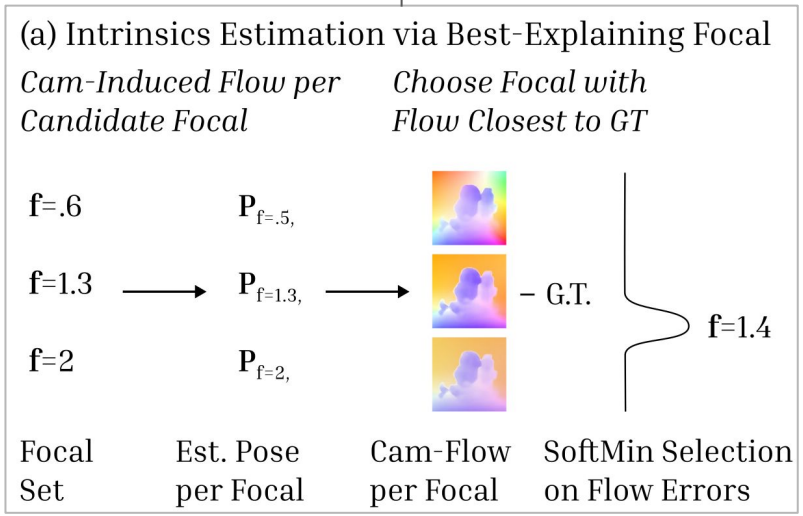
During per-scene optimization

RAFT to compute the optical flow

Teed, Z., Deng, J.: **RAFT: Recurrent all-pairs field transforms for optical flow**. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020).

Previous and basement works



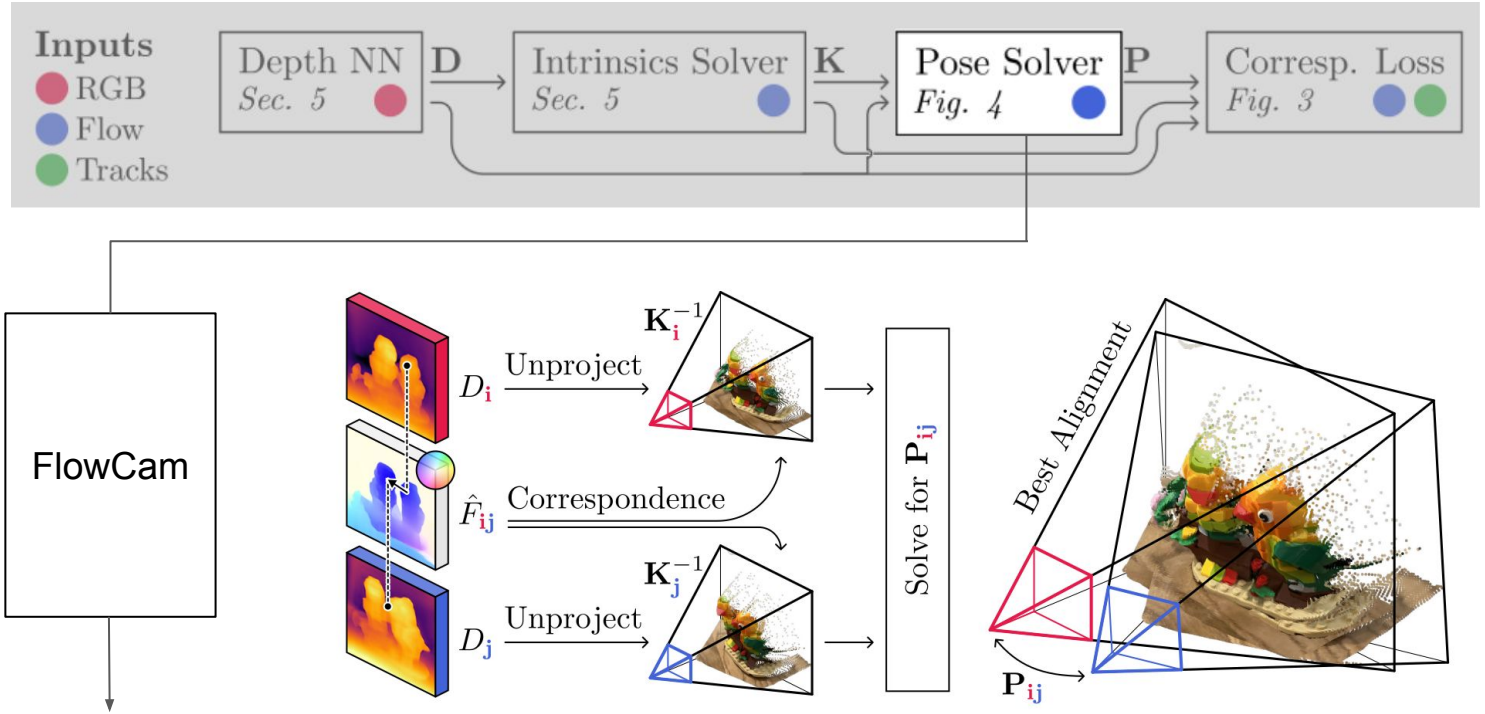


During pre-training optimization

GMFlow to compute the optical flow

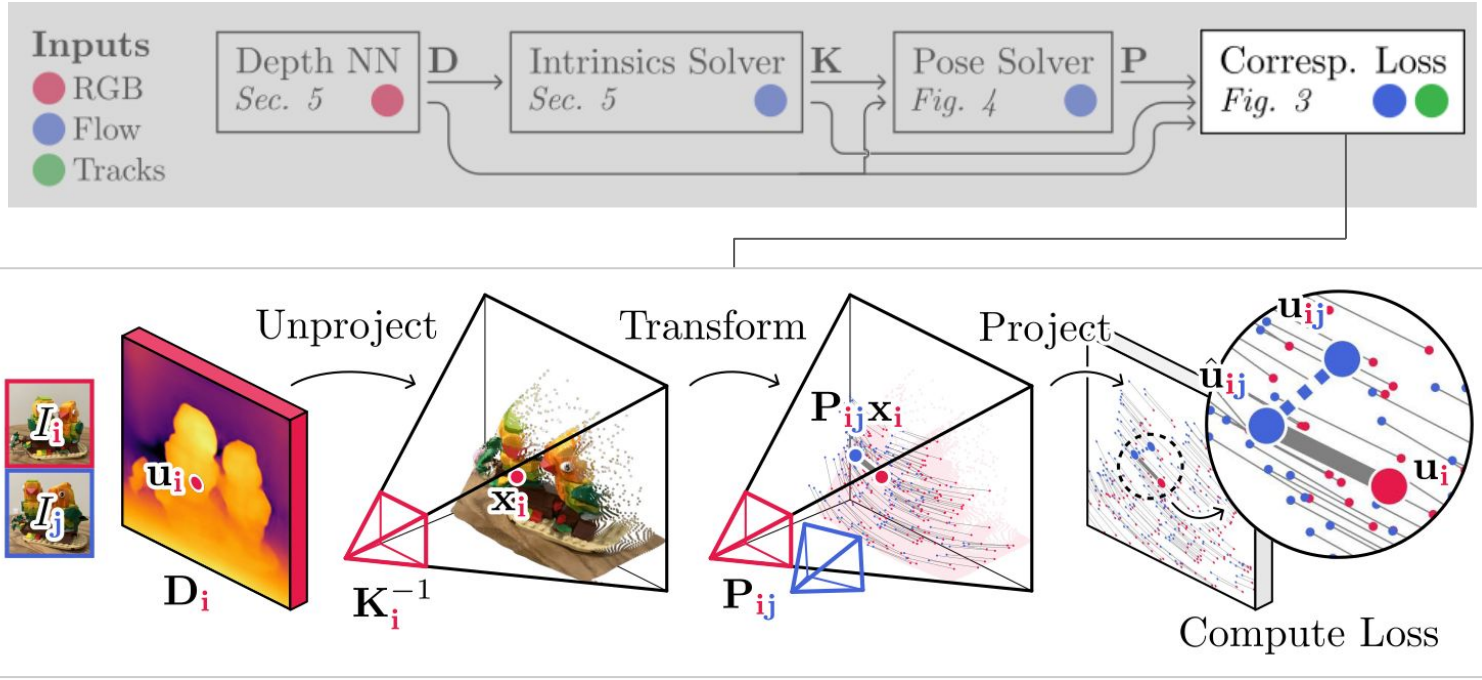
TXu, H., Zhang, J., Cai, J., Rezatofighi, H., Tao, D.: **GMFlow: Learning optical flow via global matching**. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8121–8130 (2022).

Previous and basement works



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).

Previous and basement works

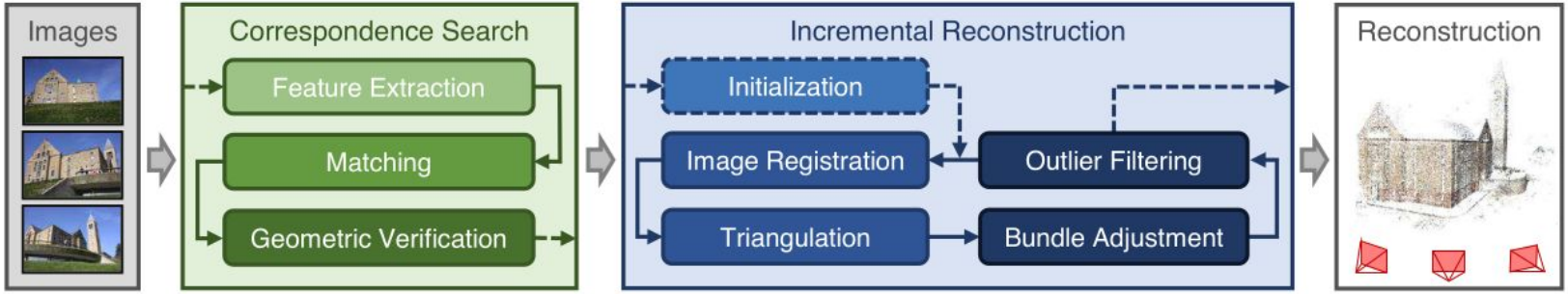


- Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., Rupprecht, C. **CoTracker: It is better to track together** (2023).

Comparative and concurrent works

COLMAP

Schonberger, J.L., Frahm, J.M. S**tructure-from-motion revisited**. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4104–4113 (2016).



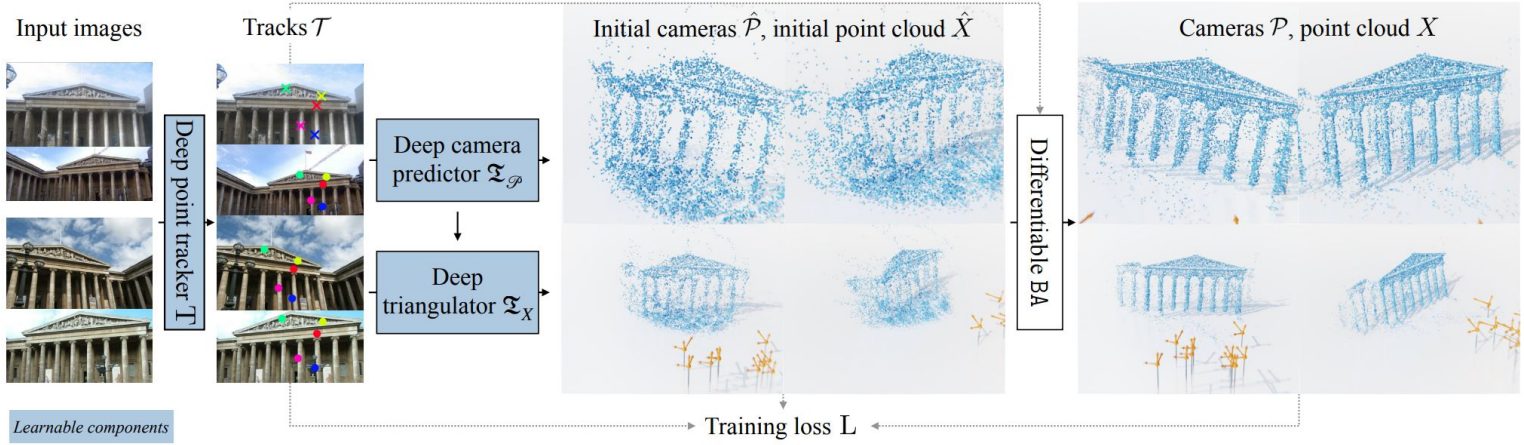
- Isolated pre-processing step x end-to-end differentiable, and can be embedded in deep learning pipeline
- Sparse 3D points x Dense per-frame depth estimates
- Depth, intrinsics and camera poses: free variables x parameterized

COLMAP FlowMap

Comparative and concurrent works

VGGSfM

Wang, J., Karaev, N., Rupprecht, C., Novotny, D.: Visual geometry grounded deep structure from motion. arXiv preprint arXiv:2312.04563 (2023).

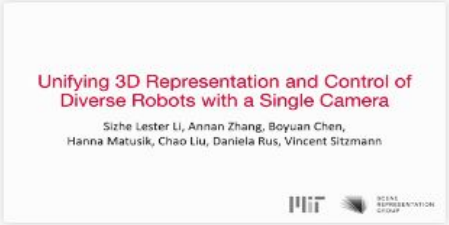


Key differences are that their method is **fully supervised** with camera poses, point clouds, and intrinsics; requires large-scale, multi-stage training; **solves only for sparse depth**; and is built around the philosophy of **making each part of the conventional SfM pipeline differentiable**.

Cited by

- **lg-slam: Instant gaussian slam (2024)** - In Related Work section.
- **EG4D: Explicit Generation of 4D Object without Score Distillation (2024)** - In Future Work section, as a possible alternative to incorporate camera pose technics in 4D reconstruction.
- **D-NPC: Dynamic Neural Point Clouds for Non-Rigid View Synthesis from Monocular Video (2024)** - Brief citation.
- **MASt3R-SfM: a Fully-Integrated Solution for Unconstrained Structure-from-Motion (2024)** - Comparing results.
- **Seamless Augmented Reality Integration in Arthroscopy: A Pipeline for Articular Reconstruction and Guidance (2024)** - Comparing results.

Authors Group

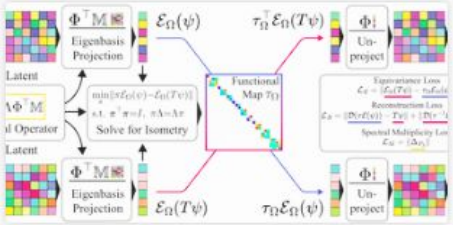


Neural Jacobian Fields: Unifying 3D Representation and Control of...

ArXiv · July 2024 · Li et al.

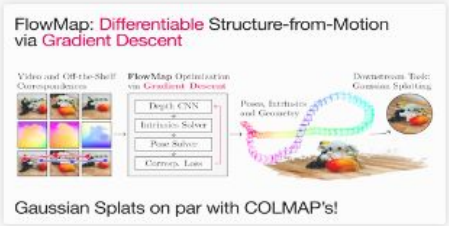
Diffusion Forcing: Next-token Prediction Meets Full-Sequence...

NeurIPS · July 2024 · Chen et al.



Neural Isometries: Taming Transformations for Equivariant ML

NeurIPS · June 2024 · Mitchel et al.

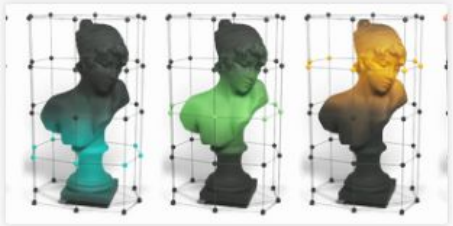


FlowMap: High-Quality Camera Poses, Intrinsics, and Depth via...

ArXiv · April 2024 · Smith, Charatan et al.

pixelSplat: 3D Gaussian Splats from Image Pairs for Scalable...

CVPR · Dec. 2023 · Charatan · ✨Best Pa...



Variational Barycentric Coordinates

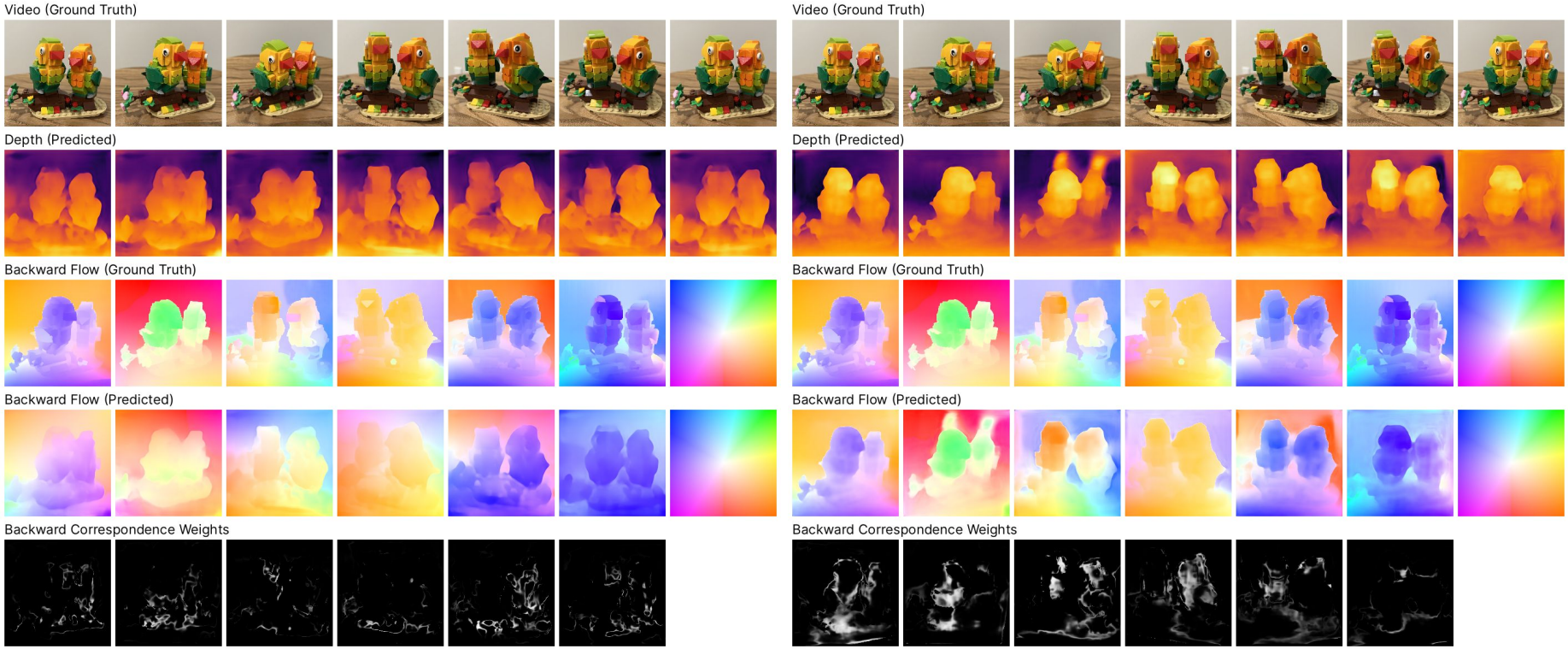
SIGGRAPH Asia · Oct. 2023 · Dodik · ✨Jo...

Scene Representation Group - MIT

FlowMap

Hacker - Veronika

Flowmap

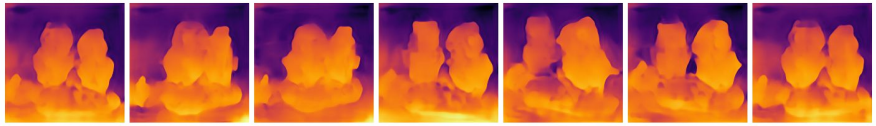


FlowMap: High-Quality Camera Poses, Intrinsics, and Depth via Gradient Descent (2024)

Without point tracks

Video (Ground Truth)

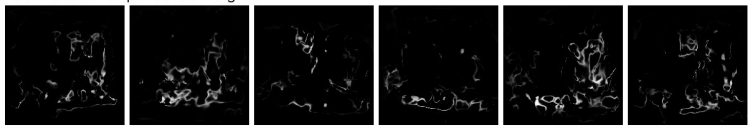
Depth (Predicted)



Backward Flow (Ground Truth)

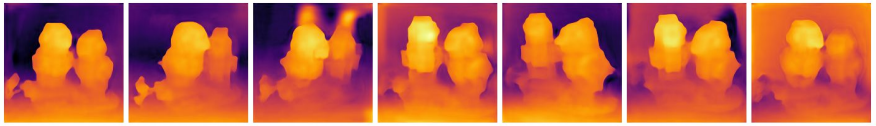
Backward Flow (Predicted)

Backward Correspondence Weights



Video (Ground Truth)

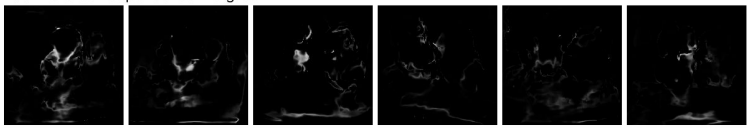
Depth (Predicted)



Backward Flow (Ground Truth)

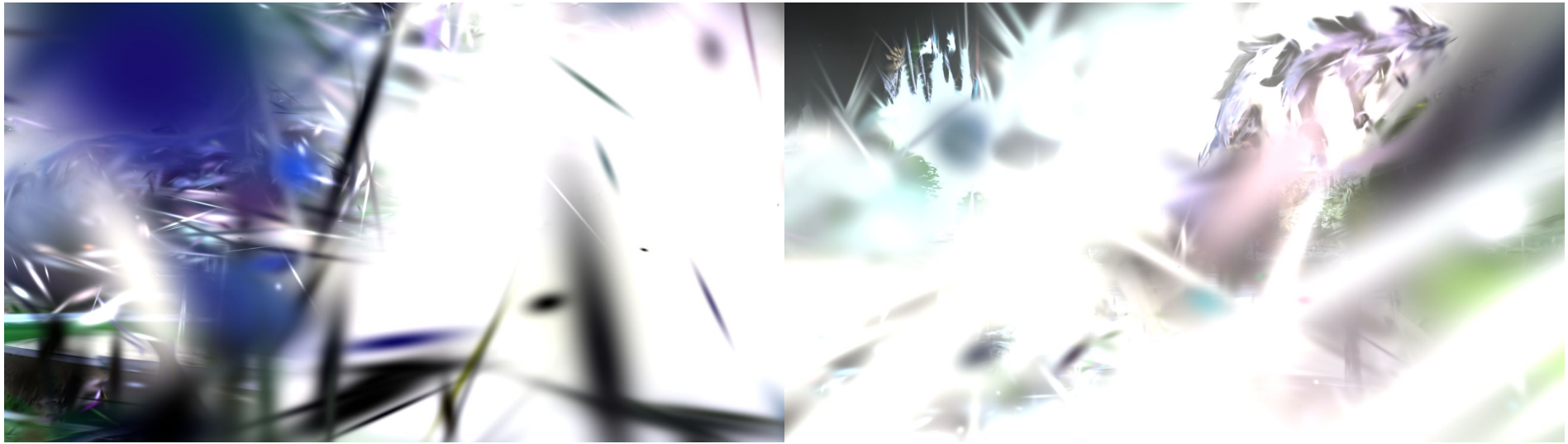
Backward Flow (Predicted)

Backward Correspondence Weights



3dgs

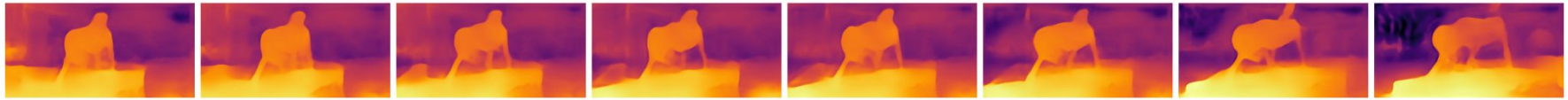
3dgs



3dgs

Video (Ground Truth)

Depth (Predicted)



Backward Flow (Ground Truth)

Backward Flow (Predicted)

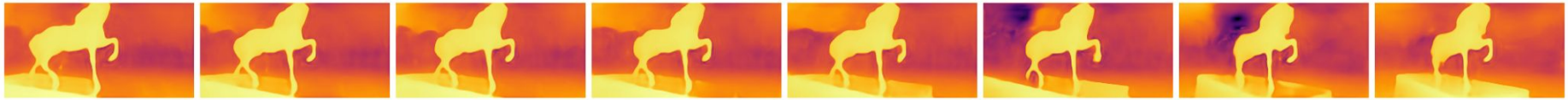
Backward Correspondence Weights

FlowMap: High-Quality Camera Poses, Intrinsics, and Depth via Gradient Descent (2024)

3dgs

Video (Ground Truth)

Depth (Predicted)



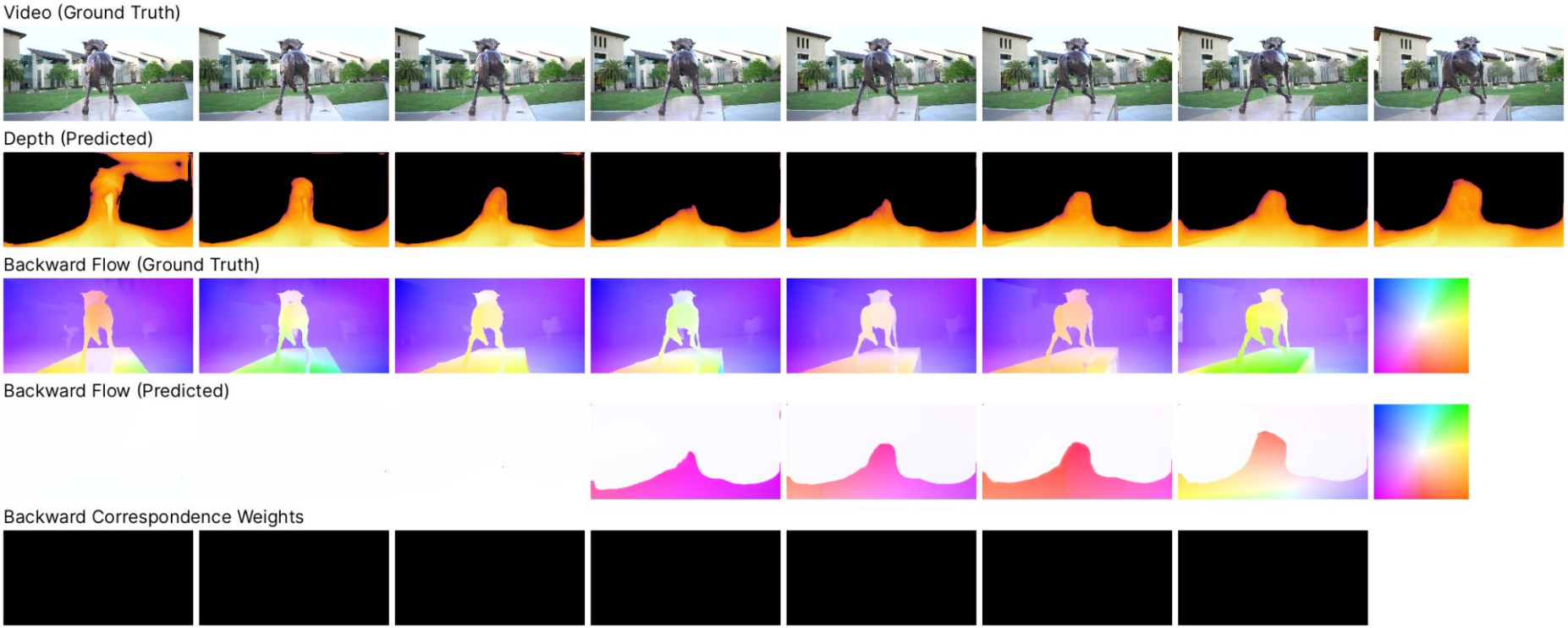
Backward Flow (Ground Truth)

Backward Flow (Predicted)

Backward Correspondence Weights

FlowMap: High-Quality Camera Poses, Intrinsics, and Depth via Gradient Descent (2024)

3dgs


Video (Ground Truth)
Depth (Predicted)
Backward Flow (Ground Truth)
Backward Flow (Predicted)
Backward Correspondence Weights

FlowMap

PhD Student - Mateus Barbosa

Problem

Because of its dependence on optical flow or point tracks to find correspondences, Flowmap can only process continuous video.

The authors suggest that leveraging unstructured correspondences might be used to overcome this limitation.

Other attempts

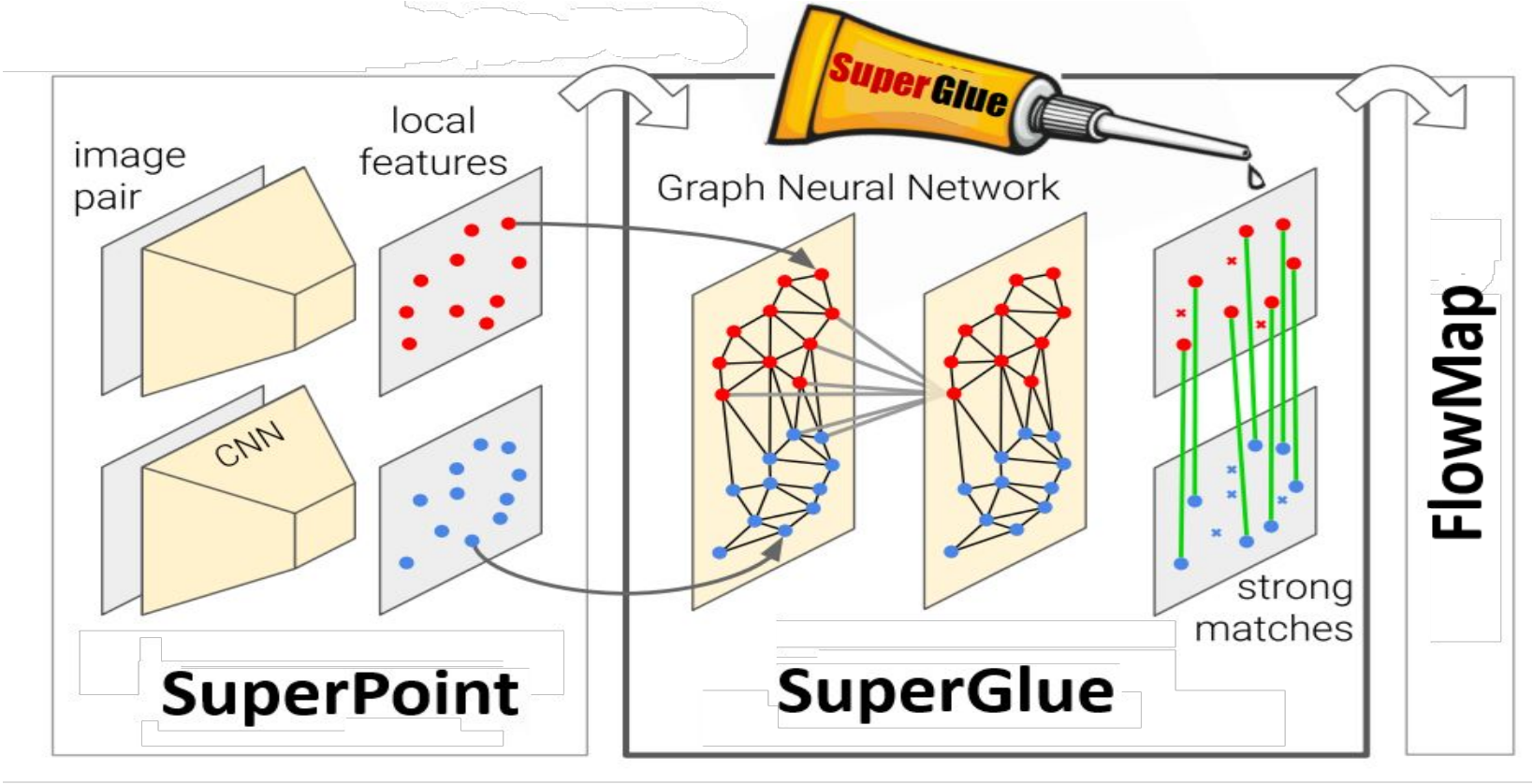
Schonberger, J. L., & Frahm, J. M. (2016). **Structure-from-motion revisited**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Duisterhof, B. et al. (2024). **MASt3R-SfM: a Fully-Integrated Solution for Unconstrained Structure-from-Motion**. *arXiv preprint arXiv:2409.19152*.

He, X. et al. (2024). **Detector-free structure from motion**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Wang, J. et al. (2024). **VGGSfM: Visual Geometry Grounded Deep Structure From Motion**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Project: incorporate SuperPoint and SuperGlue to find correspondences



Obrigado!

Dúvidas?