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


Revisor - Mateus Barbosa
IMPA
mateus.barbosa@impa.br

Archaeologist - Mohara Nascimento
Instituto Tecgraf/PUC-Rio
mohara.civil@gmail.com

Hacker - Veronika Treumova
IMPA
veronika.treumova@impa.br

PhD Student - Mateus Barbosa
IMPA
mateus.barbosa@impa.br


## 1. Revisor

### 1.1. Overview

FlowMap is a method designed to recover high-quality camera poses, intrinsics, and dense depth maps for video sequences. Unlike conventional Structure-from-Motion (SfM) methods like COLMAP, FlowMap employs a gradient-descent-based approach that directly optimizes depth, intrinsics, and poses using optical flow and point track correspondences. This differentiable approach allows integration into deep learning pipelines and removes the reliance on precomputed camera parameters. The method utilizes feed-forward re-parameterizations to ensure consistency and robustness while maintaining differentiability.

Key contributions include:

- End-to-end differentiable formulation for camera and depth estimation.
- Dense per-frame depth maps instead of sparse 3D points.
- Achieves comparable quality to COLMAP in 3D reconstruction and novel view synthesis tasks in some scenarios.

### 1.2. Methodology

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.

FlowMap's pipeline is shown in Figure 1.

**Depth Neural Network.** Depth is parameterized as a neural network that maps an RGB frame to the corresponding per-pixel depth. They use the lightweight CNN version of MiDaS [5], pretrained with the publicly available weights trained on relative-depth estimation. The weights of the entire network are optimized during training.

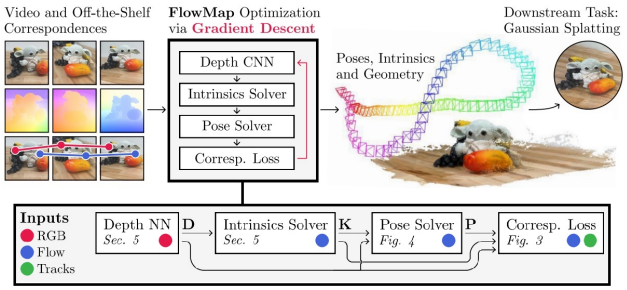


Figure 1. FlowMap's pipeline
Source: FlowMap[11]

**Intrinsics solver.** Camera intrinsics are solved considering a set of 60 focal length candidates $K_k$, that range uniformly from .5 to 2. A loss function $\mathcal{L}_k$ is calculated considering the pose calculated by $K_k$ and finally intrinsics $K$ is computed via a softmin-weighted sum of the candidates, as we can see below.

To make this approach computationally efficient, it is assumed that the intrinsics can be represented via a single $K$ that is shared across frames. Second, we assume that $K$ can be modeled via a single focal length with a principal point fixed at the image center. Finally, we only compute the soft selection losses on the first two frames of the sequence.

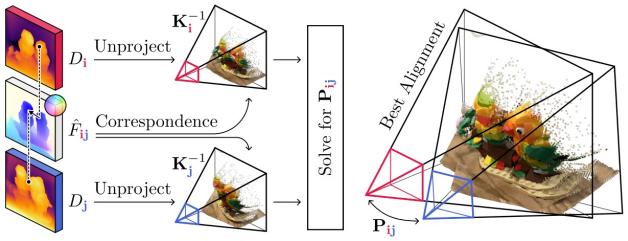


Figure 2. Pose solver
Source: FlowMap[11]

$$K = \sum_k w_k K_k \qquad w_k = \frac{\exp(-\mathcal{L}_k)}{\sum_l \exp(-\mathcal{L}_l)} \tag{1}$$

**Pose solver.** The relative poses between consecutive frames are solved using their depth maps, camera intrinsics, and optical flow. To do so, they first unproject their depth maps, then solve for the pose that best aligns the resulting point clouds, as shown in Figure 4

More formally, depth map alignment is cast as an orthogonal Procrustes problem, to draw upon this problem's differentiable, closed-form solution.The depth maps $D_i$ and $D_j$ are unprojected using their respective intrinsics $\mathbf{K}_i$ and $\mathbf{K}_j$ to generate two point clouds $\mathbf{X}_i$ and $\mathbf{X}_j$. Next, because the Procrustes formulation requires correspondence between points, the known optical flow between frames $i$ and $j$ is used to match points in $\mathbf{X}_i$ and $\mathbf{X}_j$. This yields $\mathbf{X}_i^{\leftrightarrow}$ and $\mathbf{X}_j^{\leftrightarrow}$, two filtered point clouds for which a one-to-one correspondence exists. The Procrustes formulation seeks the rigid transformation that minimizes the total distance between the matched points:

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

The diagonal matrix $\mathbf{W}$ contains correspondence weights that can down-weight correspondences that are faulty due to occlusion or imprecise flow. This weighted least-squares problem can be solved in closed form via a single singular value decomposition, which is both cheap and fully differentiable. As in FlowCam, these weights are predicted by concatenating corresponding per-pixel features and feeding them into a small MLP.

**Correspondence Loss.** Consider a 2D pixel at coordinate $\mathbf{u}_i \in \mathbb{R}^2$ in frame $i$ of the video sequence. Using frame $i$'s estimated depth $D_i$ and intrinsics $\mathbf{K}_i$, we can compute the pixel's 3D location $\mathbf{x}_i \in \mathbb{R}^3$. Then, using the estimated relative pose $\mathbf{P}_{ij}$ between frames $i$ and $j$, we can transform

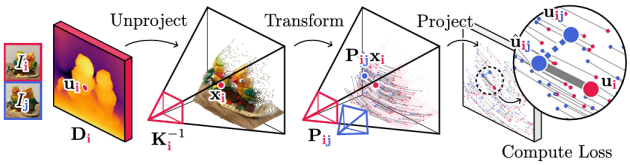


Figure 3. Camera-induced Flow Loss
Source: FlowMap[11]

this location into frame $j$'s camera space. Finally, we can project the resulting point $\mathbf{P}_{ij}\mathbf{x}_i$ onto frame $j$'s image plane to yield an implied correspondence $\hat{\mathbf{u}}_{ij}$. This correspondence can be compared to the known correspondence $\mathbf{u}_{ij}$ to yield a loss $\mathcal{L}$, as illustrated in Figure 3.

$$\mathcal{L} = \|\hat{u}_{ij} - u_{ij}\| \tag{3}$$

### 1.3. Strengths

- Fully differentiable design suitable for deep learning frameworks.
- Presents a closed form solution to pose and intrinsics estimation
- Demonstrates competitive performance with state-of-the-art methods like COLMAP, particularly for novel view synthesis.

### 1.4. Weaknesses

- Higher computational demands, requiring significant GPU memory.
- Performance is highly dependent on high-quality optical flow and point tracking.
- Limited applicability to unstructured image collections, as it primarily supports continuous video sequences.
- Lacks robustness in long sequences due to cumulative drift in pose estimation.

### 1.5. Evaluation

FlowMap presents an alternative to traditional SfM techniques, particularly for scenarios requiring differentiability.

That being said, its achievements aren't as impressive as the authors suggest, its results are on-par with COLMAP only in limited scenarios with other concurrent methods performing better on a wider set of scenarios, and its methods and implementation aren't sufficiently well explained. Therefore, I suggest the paper should be rejected.

## 2. Archaeologist

### 2.1. Previous and basement works

The main goal of FlowMap is to replace COLMAP as the input for 3DGS, with the advantage of being end-to-end differentiable. Intending to achieve this, the method receives video and off-the-shelf correspondences and calculates poses, intrinsics and geometry through specific neural networks. The works used to enable this will be discussed below.

**Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer[5].** This work presents the MiDaS neural network, which is used in a lightweight version for FlowMap. MiDaS is a depth network that produces inverse depth maps, receiving RGB images and depth annotations as input. The paper uses a zero-shot cross-dataset transfer protocol, which states that the test datasets were not used in the training.

**RAFT: Recurrent all-pairs field transforms for optical flow.[12]** and **GMFlow: Learning optical flow via global matching.[14]** The RAFT and GMFlow are deep networks that capture optical flow from input images (video frames). RAFT uses a encoder to extract per-pixel features, after that, generate 4D correlations with a correlation layer, relates this with a context encoder and then produces the optical flow. While GMFlow extracts the features, use a Transformer for feature enhancement, compute a feature matching and flow propagation, to then generate the optical flow. These two deep networks are used for FlowMap to generate optical flow of a image in a specific focal set and camera pose. This flow is compared with the ground truth to calculate the intrinsics of the camera. RAFT is applied in the per-scene optimization and GMFlow in the pre-training.

**FlowCam: Training generalizable 3d radiance fields without camera poses via pixel-aligned scene flow.[10].** FlowCam is a method that reconstructs camera poses with an existent optical flow, following the descripted in the Figure 4. FlowMap uses a pose solver based in this.

**CoTracker: It is Better to Track Together.[4].** CoTracker is used in FlowMap to obtain point tracks. This method tracks 2D points from videos with RGB frames. To achieve this, the authors apply an attention mechanism to share information between tracked points and also add context to them.

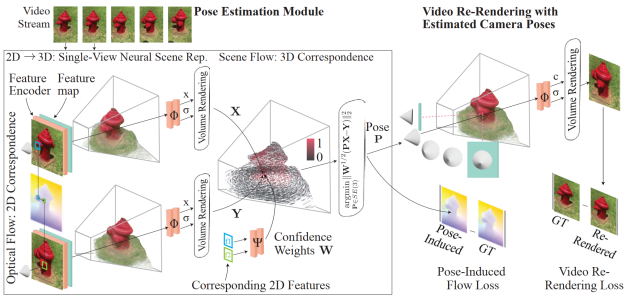


Figure 4. FlowCam Method Overview [10]

### 2.2. Comparative and concurrent works

**Structure-from-motion revisited.[8].** This paper is responsible for the classical COLMAP, applied in the most 3D reconstruction, being the major concurrent to the FlowMap. The main difference between these two are the pipeline, FlowMap is end-to-end differentiable, while CLOMAP is not. The FlowMap authors argue that this the main advantage of their method compared COLMAP, defending that their technique can be used in a end-to-end deep learning pipeline, however, this is not done in the paper. Otherwise, the FlowMap is just input in the 3DGS in the same way than COLMAP.

**Visual Geometry Grounded Deep Structure From Motion.[13].** Shortly called VGGSfM, this work presents an end-to-end differentiable method, also concurrent to COLMAP[8], but using the principles of these last. The VGGSfM overview can be seen in the Figure 5.

### 2.3. Current works

The FlowMap paper is not published and is also very new, then there is no much works based on it. The papers that cite FlowMap, uses it just as related or comparative work [2, 3, 6, 9].

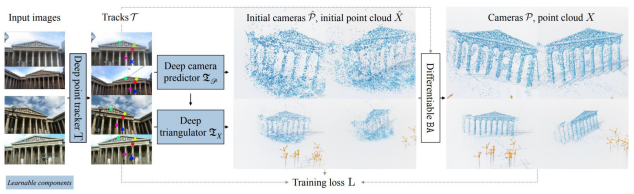


Figure 5. VGGSfM Method Overview [13]

## 3. Code and experiments

## 4. PhD Project

Because of its dependence on optical flow or point tracks to find correspondences, FlowMap can only process continuous video. Additionally, cumulative drift poses challenges for long video sequences.

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

So we propose to incorporate SuperPoint[1] to extract features off of any pair of images, and SuperGlue[7] to extract correspondences between them, as shown in Figure 6. Then, FlowMap might be extended to support unstructured collections of images.

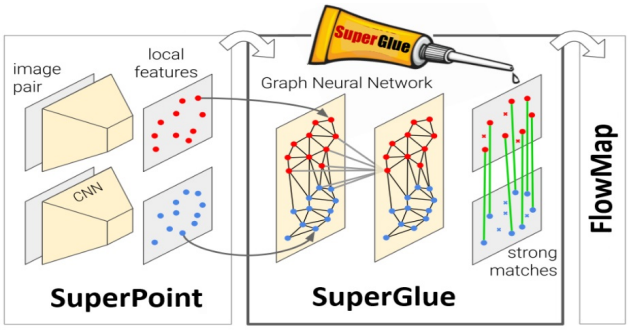


Figure 6. SuperPoint-SuperGlue-FlowMap

## 5. Conclusion

FlowMap introduces a novel framework for end-to-end differentiable 3D reconstruction, challenging the dominance of traditional SfM methods like COLMAP. While its innovative approach yields significant benefits, challenges related to efficiency, robustness, and generalization remain. Addressing these issues in future research might make FlowMap a practical alternative for scalable 3D vision applications.

## References


[1] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 224–236, 2018. 4

[2] Bardienus Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion, 2024. 3

[3] Moritz Kappel, Florian Hahlbohm, Timon Scholz, Susana Castillo, Christian Theobalt, Martin Eisemann, Vladislav Golyanik, and Marcus Magnor. D-npc: Dynamic neural point clouds for non-rigid view synthesis from monocular video, 2024. 3

[4] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. 2024. 3

[5] 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*, 2020. 1, 3

[6] F. Aykut Sarikamis and A. Aydin Alatan. Ig-slam: Instant gaussian slam, 2024. 3

[7] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4938–4947, 2020. 4

[8] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[9] Hongchao Shu, Mingxu Liu, Lalithkumar Seenivasan, Suxi Gu, Ping-Cheng Ku, Jonathan Knopf, Russell Taylor, and Mathias Unberath. Seamless augmented reality integration in arthroscopy: A pipeline for articular reconstruction and guidance, 2024. 3

[10] Cameron Smith, Yilun Du, Ayush Tewari, and Vincent Sitzmann. Flowcam: Training generalizable 3d radiance fields without camera poses via pixel-aligned scene flow. 2023. 3

[11] Cameron Smith, David Charatan, Ayush Tewari, and Vincent Sitzmann. Flowmap: High-quality camera poses, intrinsics, and depth via gradient descent. *arXiv preprint arXiv:2404.15259*, 2024. 1, 2

[12] Zachary Teed and Jia Deng. RAFT: recurrent all-pairs field transforms for optical flow. *CoRR*, abs/2003.12039, 2020. 3

[13] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Visual geometry grounded deep structure from motion, 2023. 3

[14] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. 2022. 3