Pixel-Perfect Structure-from-Motion with Featuremetric Refinement

Philipp Lindenberger[1*]   Paul-Edouard Sarlin[2*]   Viktor Larsson[2]   Marc Pollefeys[2,3]

Departments of [1]Mathematics and [2]Computer Science, ETH Zurich   [3]Microsoft

Abstract

Finding local features that are repeatable across multiple views is a cornerstone of sparse 3D reconstruction. The classical image matching paradigm detects keypoints per-image once and for all, which can yield poorly-localized features and propagate large errors to the final geometry. In this paper, we refine two key steps of structure-from-motion by a direct alignment of low-level image information from multiple views: we first adjust the initial keypoint locations prior to any geometric estimation, and subsequently refine points and camera poses as a post-processing. This refinement is robust to large detection noise and appearance changes, as it optimizes a featuremetric *error based on dense features predicted by a neural network. This significantly improves the accuracy of camera poses and scene geometry for a wide range of keypoint detectors, challenging viewing conditions, and off-the-shelf deep features. Our system easily scales to large image collections, enabling pixel-perfect crowd-sourced localization at scale. Our code is publicly available at* `github.com/cvg/pixel-perfect-sfm` *as an add-on to the popular SfM software COLMAP.*

1. Introduction

Mapping the world is an important requirement for spatial intelligence applications in augmented reality or robotics. Tasks like visual localization or path planning can benefit from accurate sparse or dense 3D reconstructions of the environment. These can be built from images using Structure-from-Motion (SfM), which associates observations across views to estimate camera parameters and 3D scene geometry. Sparse reconstruction based on matching local image features [10, 20, 22, 32, 48, 54, 56, 62] is the most common due to its scalability and its robustness to appearance changes introduced by varying devices, viewpoints, and temporal conditions found in crowdsourced scenarios [2, 28, 33, 39, 45, 47, 55].

SfM assumes that sparse interest points [10, 20, 22, 32, 48, 56, 59, 81, 88] can be reliably detected across views. It typi-
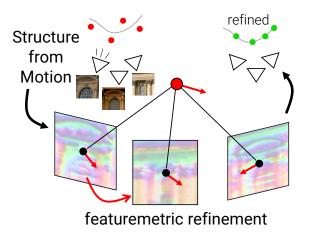

Figure 1: **From sparse to dense.** We improve the accuracy of sparse Structure-from-Motion by refining 2D keypoints, camera poses, and 3D points using the direct alignment of deep features. This featuremetric optimization leverages dense image information but can scale to scenes with thousands of images. Such refinement results in subpixel-accurate reconstructions, even in challenging conditions.

cally selects such points for each image independently and relies on these initial detections for the remainder of the reconstruction process. However, detecting keypoints from a single view is inherently inaccurate due to appearance changes and discrete image sampling [30]. The advent of convolutional neural network (CNNs) for detection has magnified this issue, as they generally do not retain local image information and instead favor global context.

Multi-view geometric optimization with bundle adjustment [4, 40, 79] is commonly used to refine cameras and points using reprojection errors. Dusmanu *et al.* [23] proposed to refine keypoint locations prior to SfM via an analogous geometric cost constrained with local optical flow. This can improve SfM, but has limited accuracy and scalability.

In this work, we argue that local image information is valuable throughout the SfM process to improve its accuracy. We adjust both keypoints and bundles, before and after reconstruction, by direct image alignment [18, 25, 49] in a learned feature space. Exploiting this locally-dense informa-
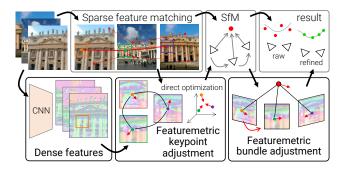

Figure 2: **Refinement pipeline.** Our refinement works on top of any SfM pipeline that is based on local features. We perform a two-stage adjustment of keypoints and bundles. The approach first refines the 2D keypoints only from tentative matches by optimizing a direct cost over dense feature maps. The second stage operates after SfM and refines 3D points and poses with a similar featuremetric cost.

tion is significantly more accurate than geometric optimization, while deep, high-dimensional features extracted by a CNN ensure wider convergence in challenging conditions. This formulation elegantly combines globally-discriminative sparse matching with locally-accurate dense details. It is applicable to both incremental [67, 72] and global [9, 12, 51] SfM irrespective of the types of sparse or dense features.

We validate our approach in experiments evaluating the accuracy of both 3D structure and camera poses in various conditions. We demonstrate drastic improvements for multiple hand-crafted and learned local features using off-the-shelf CNNs. The resulting system produces accurate reconstructions and scales well to large scenes with thousands of images. In the context of visual localization, it can, in addition to providing a more accurate map, also refine poses of single query images with minimal overhead.

For the benefit of the research community, we will release our code as an extension to COLMAP [67, 68] and to the popular localization toolbox hloc [60, 61]. We believe that our featuremetric refinement can significantly improve the accuracy of existing datasets [64] and push the community towards sub-pixel accurate localization at large scale.

2. Related work

Image matching is at the core of SfM and visual SLAM, which typically rely on sparse local features for their efficiency and robustness. The process i) detects a small number of interest points, ii) computes their visual descriptors, iii) matches them with a nearest neighbor search, and iv) verifies the matches with two-view epipolar estimation and RANSAC. The correspondences then serve for relative or absolute pose estimation and 3D triangulation. As keypoints are sparse, small inaccuracies in their locations can result in large errors for the estimated geometric quantities.

Differently, dense matching [13, 46, 58, 71, 74, 78, 80] considers all pixels in each image, resulting in denser and more accurate correspondences. It has been successful for constrained settings like optical flow [38, 73] or stereo depth estimation [86], but is not suitable for large-scale SfM due to its high computational cost due to many redundant correspondences. Several recent works [44, 57, 75, 92] improve the matching efficiency by first matching coarsely and subsequently refining correspondences using a local search. This is however limited to image pairs and thus cannot create point tracks required by SfM.

Our work combines the best of both paradigms by leveraging dense local information to refine sparse observations. It is inherently amenable to SfM as it can optimize all locations over multiple views in a track simultaneously.

Subpixel estimation is a well-studied problem in correspondence search. Common approaches either upsample the input images or fit polynomials or Gaussian distributions to local image neighborhoods [27,34,37,48,66]. With the widespread interest in CNNs for local features, solutions tailored to 2D heatmaps have been recently developed, such as learning fine local sub-heatmaps [36] or estimating subpixel corrections with regression [14, 77] or the soft-argmax [52, 89]. Cleaner heatmaps can also arise from aggregating predictions over multiple virtual views using data augmentation [20].

Detections or local affine frames can be combined across multiple views with known poses in a least-squares geometric optimization [24, 79]. Dusmanu *et al.* [23] instead refine keypoints solely based on tentative matches, without assuming known geometry. This geometric formulation exhibits remarkable robustness, but is based on a local optical flow whose estimation for each correspondence is expensive and approximate. We unify both keypoint and bundle optimizations into a joint framework that optimizes a featuremetric cost, resulting in more accurate geometries and a more efficient keypoint refinement.

Direct alignment optimizes differences in pixel intensities by implicitly defining correspondences through the motion and geometry. It therefore does not suffer from geometric noise and is naturally subpixel accurate via image interpolation. Direct photometric optimization has been successfully applied to optical flow [8,49], visual odometry [18,25,26,42], SLAM [5, 69], multi-view stereo (MVS) [19, 21, 87], and pose refinement [70]. It generally fails for moderate displacements or appearances changes, and is thus not suitable for large-baseline SfM. One notable work by Woodford & Rosten [84] refines dense SfM+MVS models with a robust image normalization. It focuses on dense mapping with accurate initial poses and moderate appearance changes. Georgel *et al.* [29] instead estimate more accurate relative poses by elegantly combining photometric and geometric costs. They show that dense information can improve sparse estimation

but their approach ignores appearance changes. Differently, our work improves the entire SfM pipeline starting with tentative matches and addresses larger, challenging changes.

To improve on the weaknesses of photometric optimization, numerous recent works align multi-dimensional image representations. Examples of this *featuremetric* optimization include frame tracking with handcrafted [6, 53] or learned descriptors [17, 50, 82, 83, 85], optical flow [7, 11], MVS [90], and dense SfM in small scenes [76]. Closer to our work, PixLoc [63] learns deep features with a large basin of convergence for wide-baseline pose refinement. It improves the accuracy of sparse matching but is designed for single images and disregards the scalability to multiple images or large scenes. Here we extend this paradigm to other steps of SfM and propose an efficient algorithm that scales to thousands of images. We show that learning task-specific wide-context features is not necessary and demonstrate highly accurate refinements with off-the-shelf features.

In conclusion, our work is the first to apply robust featuremetric optimization to a large-scale sparse reconstruction problem and show significant benefits for visual localization.

3. Background

Given N images $\{\mathbf{I}_i\}$ observing a scene, we are interested in accurately estimating its 3D structure, represented as sparse points $\{\mathbf{P}_j \in \mathbb{R}^3\}$, intrinsic parameters $\{\mathbf{C}_i\}$ of the cameras, and the poses $\{(\mathbf{R}_i, \mathbf{t}_i) \in \mathbf{SE}(3)\}$ of the images, represented as rotation matrices and translation vectors.

A typical SfM pipeline performs geometric estimation from correspondences between sparse 2D keypoints $\{\mathbf{p}_u\}$ observing the same 3D point from different views, collectively called a track. Association between observations is based on matching local image descriptors $\{\mathbf{d}_u \in \mathbb{R}^D\}$, but the estimated geometry relies solely on the location of the keypoints, whose accuracy is thus critical. Keypoints are detected from local image information for each image individually, without considering multiple views simultaneously. Subsequent steps of the pipeline discover additional information about the scene, such as its its geometry or its multi-view appearance. Two approaches leverage this information to reduce the detection noise and refine the keypoints.

Global refinement: Bundle adjustment [79] is the gold standard for refining structure and poses given initial estimates. It minimizes the total geometric error

$$E_{\mathrm{BA}} = \sum_{j} \sum_{(i,u) \in \mathcal{T}(j)} \left\| \Pi\left(\mathbf{R}_i \mathbf{P}_j + \mathbf{t}_i, \mathbf{C}_i\right) - \mathbf{p}_u \right\|_\gamma \ , \quad (1)$$

where $\mathcal{T}(j)$ is the set the images and keypoints in track j, $\Pi(\cdot)$ projects to the image plane, and $\|\cdot\|_\gamma$ is a robust norm [31]. This formulation implicitly refines the keypoints while ensuring their geometric consistency. It however ignores the uncertainty of the initial detections and thus re-

quires many observations to reduce the geometric noise. Operating on an existing reconstruction, it cannot recover observations arising from noisy keypoints that are matched correctly but discarded by the geometric verification.

Track refinement: To improve the accuracy of the keypoints prior to any geometric 3D estimation, Dusmanu *et al.* [23] optimize their locations over tentative tracks formed by raw, unverified matches. They exploit the inherent structure of the matching graph to discard incorrect matches without relying on geometric constraints. Given two-view dense flow fields $\{\mathbf{T}_{v \to u}\}$ between the neighborhoods of matching keypoints u and v, this *keypoint adjustment* optimizes, for each tentative track j, the multi-view cost

$$E_{\mathrm{KA}}^j = \sum_{(u,v) \in \mathcal{M}(j)} \left\| \mathbf{p}_v + \mathbf{T}_{v \to u}[\mathbf{p}_v] - \mathbf{p}_u \right\|_\gamma \ , \quad (2)$$

where $\mathcal{M}(i)$ denotes the set of matches that forms the track and $[\cdot]$ is a lookup with subpixel interpolation. A deep neural network is trained to regress the flow of a single point from two input patches and the flow field is interpolated from a sparse grid. This dramatically improves the keypoint accuracy, but some errors remain as the regression and interpolation are only approximate.

Both bundle and keypoint adjustments are based on geometric observations, namely keypoint locations and flow, but do not account for their respective uncertainties. They thus require a large number of observations to average out the geometric noise and their accuracy is in practice limited.

4. Approach

Summarizing dense image information into sparse points is necessary to perform global data association and optimization at scale. However, refining geometry is an inherently local operation, which, we show, can efficiently benefit from locally-dense pixels. Given constraints provided by coarse but global correspondences or initial 3D geometry, the dense information only needs to be locally accurate and invariant but not globally discriminative. While SfM typically discards image information as early as possible, we instead exploit it in several steps of the process thanks to direct alignment. Leveraging the power of deep features, this translates into featuremetric keypoint and bundle adjustments that elegantly integrate into any SfM pipeline by replacing their geometric counterparts. Figure 2 shows an overview.

We first introduce the featuremetric optimization in Section 4.1. We then describe our formulations of keypoint adjustment, in Section 4.2, and bundle adjustment, in Section 4.3, and analyze their efficiency.

4.1. Featuremetric optimization

Direct alignment: We consider the error between image intensities at two sparse observations: $\mathbf{r} = \mathbf{I}_i[\mathbf{p}_u] - \mathbf{I}_j[\mathbf{p}_v]$.

Local image derivatives implicitly define a flow from one point to the other through a gradient descent update:

$$\mathbf{T}_{v \to u}[\mathbf{p}_v] \propto -\frac{\partial \mathbf{I}_j}{\partial \mathbf{p}}[\mathbf{p}_v]^\top \mathbf{r} \ . \tag{3}$$

This flow can be efficiently computed at any location in a neighborhood around v, without approximate interpolation nor descriptor matching. It naturally emerges from the direct optimization of the photometric error, which can be minimized with second-order methods in the same way as the aforementioned geometric costs. Unlike the flow regressed from a black-box neural network [23], this flow can be made consistent across multiple view by jointly optimizing the cost over all pairs of observations.

Learned representation: SfM can handle image collections with unconstrained viewing conditions exhibiting large changes in terms of illumination, resolution, or camera models. The image representation used should be robust to such changes and ensure an accurate refinement in any condition. We thus turn to features computed by deep CNNs, which can exhibit high invariance by capturing a large context, yet retain fine local details. For each image $\mathbf{I}_i$, we compute a D-dimensional, L2-normalized feature map $\mathbf{F}_i \in \mathbb{R}^{W \times H \times D}$ at identical resolution. We use the same representations for keypoint and bundle adjustments, requiring a single forward pass per image. Our experiments show that multiple off-the-shelf dense local descriptors can result in highly accurate refinements. However, our formulation can also be applied to robust intensity representations, such as the normalized cross-correlation (NCC) over local image patches [84].

4.2. Keypoint adjustment

Once local features are detected, described, and matched, we refine the keypoint locations before geometrically verifying the tentative matches.

Track separation: Connected components in the matching graph define tentative tracks – sets of keypoints that are likely to observe the same 3D point, but whose observations have not yet been geometrically verified. Because a 3D point has a single projection on a given image plane, valid tracks cannot contain multiple keypoints detected in the same image. We can leverage this property to efficiently prune out most incorrect matches using the track separation algorithm introduced in [23]. This speeds up the subsequent optimization and reduces the noise in the estimation.

Objective: We then adjust the locations of 2D keypoints belonging to the same track j by optimizing its featuremetric consistency along tentative matches with the cost

$$E_{\text{FKA}}^j = \sum_{(u,v) \in \mathcal{M}(j)} w_{uv} \left\| \mathbf{F}_{i(u)}[\mathbf{p}_u] - \mathbf{F}_{i(v)}[\mathbf{p}_v] \right\|_\gamma, \tag{4}$$

where w_{uv} is the confidence of the correspondence (u, v), such as the similarity of its local feature descriptors $\mathbf{d}_u^\top \mathbf{d}_v$.

This allows the optimization to split tracks connected by weak correspondences, providing robustness to mismatches. The confidence is not based on the dense features since these are not expected to disambiguate correspondences at the global image level.

Efficiency : This direct formulation simply compares precomputed features on sparse points and is thus much more scalable than patch flow regression (Eq. 2), which performs a dense local correlation for each correspondence. All tracks are optimized independently, which is very fast in practice despite the sheer number of tentative matches.

Drift: Because of the lack of geometric constraints, the points are free to move anywhere on the underlying 3D surface of the scene. The featuremetric cost biases the updates towards areas with low spatial feature gradients and with better-defined features. This can result in a large drift if not accounted for. Keypoints should however remain repeatable w.r.t. unrefined detections to ensure the matchability of new images, such as for visual localization. It is thus critical to limit the drift, while allowing the refinement of noisier keypoints. For each track, we freeze the location of the keypoint $\bar{u}$ with highest connectivity, as in [23], and constrain the location $\mathbf{p}_u$ of each keypoint w.r.t. to its initial detection $\mathbf{p}_u^0$, such that $\left\| \mathbf{p}_u - \mathbf{p}_u^0 \right\| \leq K$.

Once all tracks are refined, the geometric estimation proceeds, typically using two-view epipolar geometric verification followed by incremental or global SfM.

4.3. Bundle adjustment

The estimated structure and motion can then be refined with a similar featuremetric cost. Here keypoints are implicitly defined by the projections of the 3D points into the 2D image planes, and only poses and 3D points are optimized.

Objective: We minimize for each track j the error between its observations and a reference appearance $\mathbf{f}^j$:

$$E_{\text{FBA}} = \sum_j \sum_{(i,u) \in \mathcal{T}(j)} \left\| \mathbf{F}_i \left[\Pi \left(\mathbf{R}_i \mathbf{P}_j + \mathbf{t}_i, \mathbf{C}_i \right) \right] - \mathbf{f}^j \right\|_\gamma \ . \tag{5}$$

The reference is selected at the beginning of the optimization and kept fixed from then on. This reduces the drift of the points significantly, as also noted in [5], but is more flexible than the common ray-based parametrization [25, 42, 84].

The reference is defined as the observation closest to the robust mean $\boldsymbol{\mu}$ over all initial observations $\mathbf{f}_u^j$ of the track:

$$\mathbf{f}^j = \operatorname*{argmin}_{\mathbf{f} \in \{\mathbf{f}_u^j\}} \left\| \boldsymbol{\mu}^j - \mathbf{f} \right\| \tag{6}$$

$$\text{with } \boldsymbol{\mu}^j = \operatorname*{argmin}_{\boldsymbol{\mu} \in \mathbb{R}^D} \sum_{\mathbf{f} \in \{\mathbf{f}_u^j\}} \left\| \mathbf{f} - \boldsymbol{\mu} \right\|_\gamma \ . \tag{7}$$

This ensures robustness to outlier observations and accounts for the unknown topology of the feature space.

SfM features ↳ Refinement	ETH3D indoor						ETH3D outdoor					
	Accuracy (%)			Completeness (%)			Accuracy (%)			Completeness (%)		
	1cm	2cm	5cm	1cm	2cm	5cm	1cm	2cm	5cm	1cm	2cm	5cm
SIFT [48]	75.62	85.04	92.45	0.21	0.87	3.61	57.64	71.92	85.23	0.06	0.34	2.45
↳ Patch Flow	80.99	89.06	**95.06**	0.24	0.97	**3.88**	64.79	78.90	90.04	**0.08**	0.41	**2.76**
↳ **ours**	**82.82**	**89.77**	94.77	**0.25**	**0.96**	3.75	**68.43**	**80.73**	**91.28**	**0.08**	**0.42**	2.75
SuperPoint [20]	75.76	85.61	93.38	0.59	2.21	8.89	50.45	65.07	80.26	0.10	0.55	3.92
↳ Patch Flow	85.77	91.57	95.85	0.72	2.51	**9.59**	64.94	77.65	88.86	0.15	0.77	4.93
↳ **ours**	**89.33**	**93.58**	**96.58**	**0.74**	**2.53**	9.51	**71.27**	**82.58**	**92.08**	**0.16**	**0.83**	**5.06**
D2-Net [22]	47.18	64.94	83.37	0.47	1.87	7.07	20.87	34.55	56.53	0.03	0.19	1.78
↳ Patch Flow	79.10	86.64	93.26	**1.45**	**4.53**	**12.95**	57.34	70.71	84.12	**0.21**	**1.06**	**6.02**
↳ **ours**	**82.49**	**88.83**	**94.35**	1.36	4.13	11.80	**65.71**	**77.95**	**89.22**	**0.21**	1.01	5.63
R2D2 [56]	66.30	79.21	90.00	0.53	2.06	8.62	49.32	66.10	83.10	0.11	0.55	3.63
↳ Patch Flow	77.94	85.82	92.48	0.66	**2.32**	**9.07**	64.14	78.10	90.18	**0.16**	0.71	**4.09**
↳ **ours**	**80.67**	**87.61**	**93.42**	**0.67**	2.31	8.95	**67.77**	**80.85**	**91.91**	**0.16**	**0.73**	**4.09**

SuperPoint - raw

SuperPoint - refined

correct/incorrect @ 1cm

Table 1: **3D sparse triangulation.** Our refinement yields significantly more accurate and complete point clouds than the common geometric SfM pipeline. It is more effective than the existing Patch Flow [23], especially at 1cm or with SIFT.

Efficiency: Compared to the keypoint adjustment (Eq. 4), using a reference feature reduces the number of residuals from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$. On the other hand, all tracks need to be updated simultaneously because of the interdependency caused by the camera poses. To accelerate the convergence, we form a reduced camera system based on the Schur complement and use embedded point iterations [40]. The refinement generally converges within a few camera updates.

4.4. Implementation

Dense extractor: Our refinement can work with any off-the-shelf CNN that produces feature maps that are locally discriminative. These should be of the same resolution as the input (stride 1) to enable subpixel accuracy. The radius of convergence, or context, of such features depends on the amount of noise in the keypoints. Most detectors like SIFT have at most a few pixels of error, while others like D2-Net exhibit a much larger detection noise. In our experiments, we use S2DNet [30] for dense feature extraction, as it computes fine features very efficiently in only 4 convolutions, but also produce, if required, deeper features with a larger context. These can then be combined into a multi-level optimization scheme [25, 63, 82] that sequentially refines based on coarse to fine features. The convergence can thus be adjusted depending on the detector and on the image resolution. We show in Section 5.4 that other dense features work well too.

Optimization: The optimization problems of both keypoint and bundle adjustments are solved with the Levenberg-Marquardt [43] algorithm implemented using Ceres [3]. Feature maps are stored as collections of 16×16 patches centered around the initial keypoint detections. We thus constrain points to move at most $K = 8$ pixels. The feature lookup is implemented as bicubic interpolation. We use the Cauchy loss γ with a scale of 0.25. The robust mean in Eq. 7

is computed with iteratively reweighted least squares [35].

Simultaneously storing all high-dimensional feature patches incurs high memory requirements during BA. We dramatically increase its efficiency by exhaustively precomputing patches of feature distances and directly interpolate an approximate cost $\bar{E}_{ij} = \left\| \mathbf{F}_i - \mathbf{f}^j \right\|_\gamma \left[\mathbf{p}_{ij} \right]$. To improve the convergence, we store and optimize its spatial derivatives $\partial \bar{E}_{ij} / \partial \mathbf{p}_{ij}$. This reduces the residual size from D to 3 with no loss of accuracy. See Supplemental C for more details.

Run time and memory: S2DNet can extract 3-5 dense feature maps per second and both featuremetric adjustments run in less than 5 minutes for 100 images. As these features are 128-dimensional, the memory consumption can be a bottleneck. We believe that much fewer dimensions are actually required for refinement, and retraining a compact feature extractor would improve the efficiency of the optimization.

5. Experiments

We evaluate our featuremetric refinement on various SfM tasks with several handcrafted and learned local features and show substantial improvements for all of them. We first evaluate its accuracy on the tasks of triangulation and camera pose estimation in Sections 5.1 and 5.2, respectively. We then assess in Section 5.3 the impact of the refinement on two-view and multi-view pose estimation for end-to-end reconstruction in challenging conditions. Lastly, Section 5.4 analyzes the validity and scalability of our design decisions through an ablation study.

5.1. 3D triangulation

We first evaluate the accuracy of the refined 3D structure given known camera poses and intrinsics.

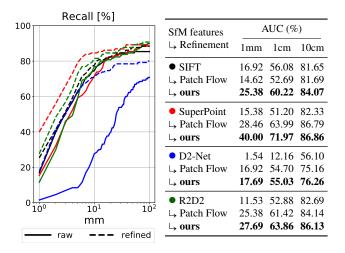**Evaluation:** We use the ETH3D benchmark [70], which

Table 2: **Camera pose estimation.** We plot the cumulative translation error and report its AUC. Our refinement improves the accuracy of the query camera poses for all local features, even when for SIFT, whose detections are already well-localized. It is generally more accurate than Patch Flow.

SfM features ↳ Refinement	AUC (%)		
	1mm	1cm	10cm
● SIFT	16.92	56.08	81.65
↳ Patch Flow	14.62	52.69	81.69
↳ **ours**	**25.38**	**60.22**	**84.07**
● SuperPoint	15.38	51.20	82.33
↳ Patch Flow	28.46	63.99	86.79
↳ **ours**	**40.00**	**71.97**	**86.86**
● D2-Net	1.54	12.16	56.10
↳ Patch Flow	16.92	54.70	75.16
↳ **ours**	**17.69**	**55.03**	**76.26**
● R2D2	11.53	52.88	82.69
↳ Patch Flow	25.38	61.42	84.14
↳ **ours**	**27.69**	**63.86**	**86.13**

SfM features (# keypoints) ↳ Refinement	Task 1: Stereo AUC@K°		Task 2: Multiview AUC@5°@N		
	5°	10°	5	10	25
SuperPoint+SuperGlue (2k)	58.78	71.01	63.02	77.36	86.76
↳ **ours**	**65.89**	**76.51**	**68.87**	**82.09**	**89.73**
SIFT (2k)	38.09	48.05	25.12	50.82	77.28
↳ **ours**	**40.59**	**50.87**	**28.01**	**53.59**	**79.49**
D2-Net (4k)	16.83	22.40	16.52	33.07	49.35
↳ **ours**	**25.89**	**33.32**	**21.33**	**40.69**	**57.93**

Table 3: **End-to-end SfM.** The proposed refinement improves the accuracy of poses estimated by epipolar geometry (stereo) or a complete SfM pipeline (multiview) with crowd-sourced imagery. Improvements are substantial for both standard (SIFT) and recent (SuperGlue) matching configurations, especially when few images N observe the scene.

5.2. Camera pose estimation

We now evaluate the impact of our refinement on the task of camera pose estimation from a single image.

Evaluation: We again follow the setup of [23] based on the ETH3D benchmark. For each scene, 10 images are randomly selected as queries. For each of them, the remaining images, excluding the 2 most covisible ones, are used to triangulate a sparse 3D partial model. Each query is then matched against its corresponding partial model and the resulting 2D-3D matches serve to estimate its absolute pose using LO-RANSAC+PnP [15] followed by geometric refinement. We compare the 130 estimated query poses to their ground truth and report the area under the cumulative translation error curve (AUC) up to 1mm, 1cm, and 10cm.

Baselines: Patch Flow performs multi-view optimization over each partial model independently as well as over the matches between each query and its partial model. Similarly, we first refine each partial model as in Section 5.1. We then adjust the query keypoints using its tentative matches, estimate an initial pose, and refine it with featuremetric BA.

Results: The AUC and its cumulative plot are shown in Table 2. Our refinement substantially improves the localization accuracy for all local features, including SIFT, for which Patch Flow does not show any benefit. At all error thresholds, featuremetric optimization is consistently more accurate than its geometric counterparts. The accuracy of SuperPoint is raised far higher than other detectors, despite the high sparsity of the 3D models that it produces. This shows how more accurate keypoint detections can result in much more accurate visual localization.

5.3. End-to-end Structure-from-Motion

While the previous experiments precisely quantify the accuracy of the refinement, they do not contain any variations

The benchmark is composed of 13 indoor and outdoor scenes and provides images with millimeter-accurate camera poses and highly-accurate ground truth dense reconstructions obtained with a laser scanner. We follow the protocol introduced in [23], in which a sparse 3D model is triangulated for each scene using COLMAP [67] with fixed camera poses and intrinsics. Following the original benchmark setup, we report the accuracy and completeness of the reconstruction, in %, as the ratio of triangulated and ground-truth dense points that are within a given distance of each other.

Baselines: We evaluate our featuremetric refinement with the hand-crafted local features SIFT [48] and the learned ones SuperPoint [20], D2-Net [22], and R2D2 [56], using the associated publicly available code repositories. We compare our approach to the geometric optimization of [23], referred here as Patch Flow. We re-compute the numbers provided in the original paper using the code provided by the authors.

Results: Table 1 shows that our approach results in significantly more accurate and complete 3D reconstructions compared to the traditional geometric SfM. It is more accurate than Patch Flow, especially at the strict threshold of 1cm, and exhibits similar completeness. The improvements are consistent across all local features, both indoors and outdoors. The gap with Patch Flow is especially large for SIFT, which already detects well-localized keypoints. This confirms that our featuremetric optimization better captures low-level image information and yields a finer alignment. Patch Flow is more complete for larger thresholds as it partly solves a different problem by increasing the keypoint repeatability with its large receptive field, while we focus on their localization.

of appearance or camera models. We thus turn to crowd-sourced imagery and evaluate the benefits of our featuremetric optimization in an end-to-end reconstruction pipeline.

Evaluation: We use the data, protocol, and code of the 2020 Image Matching Challenge [1,41]. It is based on large collections of crowd-sourced images depicting popular landmarks around the world. Pseudo ground truth poses are obtained with SfM [67] and used for two tasks. The stereo task evaluates relative poses estimated from image pairs by decomposing their epipolar geometry. This is a critical step of global SfM as it initializes its global optimization. The multiview task runs incremental SfM for small subsets of images, making the SfM problem much harder, and evaluates the final relative poses within each subset. For each task, we report the AUC of the pose error at the threshold of $5°$, where the pose error is the maximum of the angular errors in rotation and translation. As the evaluation server accepts at most correspondences, we cannot evaluate our method using the test data. We instead test on a subset of the publicly available validation scenes, and tune the RANSAC and matching parameters on the remaining scenes. More details on this setup are provided in the Supplemental.

Baselines: We evaluate our refinement in combination with SIFT [48], D2-Net [22], and SuperPoint+SuperGlue [20,62]. We limit the number of detected keypoints to 2k for computational reasons, but increase this number to 4k for D2-Net as it otherwise performs poorly. In the stereo task, we adjust the keypoints using the entire exhaustive tentative match graph (4950 pairs per scene). We use LO-DEGENSAC [15, 16] for match verification, the ratio test for SIFT, and the mutual check for SIFT and D2-Net. In the multiview task, we adjust keypoints for each subset independently, considering only the matches between images in the subset, and run our bundle adjustment after SfM.

Results: Table 3 summarizes the results. For stereo, our featuremetric keypoint adjustment significantly improves the accuracy of the two-view epipolar geometries across all local features and despite the challenging conditions. In multiview setting, it also improves the accuracy of the SfM poses, especially for small sets of images. Featuremetric optimization is particularly effective in this situation, as geometric optimization cannot fully suppress the detection noise due to the small number of observations. We visualize tracks of a 5-image reconstruction in Figure 4 and highlight the accuracy of the refined SfM model.

5.4. Additional insights

Ablation study: Table 4 shows the performance of several variants of our featuremetric optimization on ETH3D in terms of triangulation (scene *Facade* only) and localization (all scenes). We compare both types of adjustments, minor tweaks, and different image representations, including

SuperPoint ↳ Refinement	Acc. (%)		Compl. (%)		track length	AUC
	1cm	2cm	1cm	2cm		1cm
KA vs. BA						
unrefined	18.42	32.23	0.06	0.49	4.17	51.20
↳ Patch Flow [23]	37.00	55.18	0.15	0.93	**5.24**	63.53
↳ F-KA	36.85	54.48	0.15	0.90	5.02	69.84
↳ F-BA	43.65	62.44	0.18	1.06	4.17	67.61
↳ **F-KA+BA (full)**	**46.46**	**65.41**	**0.19**	**1.14**	5.02	**71.97**
bonus						
w/ F-BA drift	47.93	66.52	0.20	1.17	5.02	64.51
Patch Flow + F-BA	46.30	65.22	0.19	1.13	5.24	-
higher resolution	47.67	65.39	0.21	1.21	5.12	-
dense feats						
photometric BA [84]	28.43	45.87	0.11	0.72	4.17	-
VGG-16 ImageNet	36.86	54.99	0.15	0.90	4.61	-
DSIFT [46]	38.78	56.46	0.16	0.96	4.73	-
PixLoc [63]	29.49	46.60	0.12	0.74	4.48	-

Table 4: **Ablation study on ETH3D.** i) Featuremetric keypoint and bundle adjustments (KA and BA) both largely improve the triangulation and localization accuracy. Patch Flow produces a longer track length because of its larger receptive field but is less accurate. ii) Letting the BA drift by updating reference features or increasing the image resolution both improve the triangulation, at the expense of poorer localization and increased run time, respectively. iii) Different image representations are better than the unrefined detections but S2DNet (our default) works best.
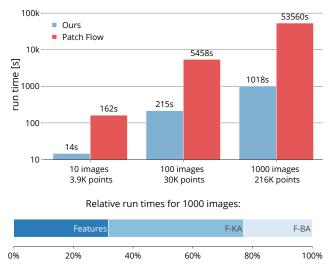

Figure 3: **Run-times.** We show the duration, in logarithmic scale, of the refinement for varying numbers of images. Our refinement is more than ten times faster than Patch Flow [23], whose run-time is dominated by the computation of the pairwise flow, which scales quadratically. Thanks to our precomputed cost patches, the featuremetric BA is fast. The KA amounts for the majority of the refinement time.

NCC-normalized intensity patches with fronto-parallel warping. Our final configuration, based on on the dense features of S2DNet [30], performs best across all metrics. We will now show that it is also fairly efficient.
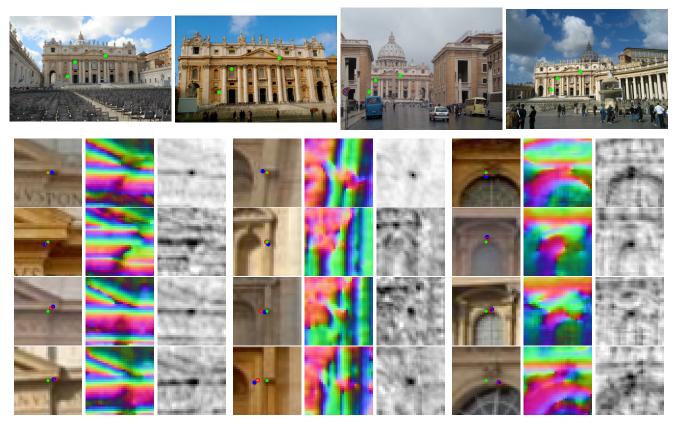
Figure 4: **Refined SfM tracks.** We show patches centered around reprojections of 3x 3D points observed in 4 images of the *St. Peter's Square* scene. Deep features and their correlation maps with a reference are robust to scale or illumination changes, yet preserve local details required for fine alignment. Points refined with our approach (in green) are consistent across multiple views while those of a standard SfM pipeline (in red) are misaligned because the initial keypoint detections (in blue) are noisy.

Scalability: We run SfM on subsets of images of the Aachen Day-Night dataset [64, 65, 91]. Figure 3 shows the run times of the refinement for subsets of 10, 100 and 1000 images. The featuremetric refinement is an order of magnitude faster than Patch-Flow [23]. Precomputing distance maps reduces the peak memory requirement of the bundle adjustment from 80 GB to less than 10GB for 1000 images. As storing feature maps only requires 50 GB of disk space, this refinement can easily run on a desktop PC. We thus refined the entire Aachen Day-Night v1.1 model, composed of 7k images, in less than 2 hours. Scene partitioning [67] could further reduce the peak memory. See Supplemental D for more details.

6. Conclusion

In this paper we argue that the recipe for accurate large-scale Structure-from-Motion is to perform an initial coarse estimation using sparse local features, which are by necessity globally-discriminative, followed by a refinement using locally-accurate dense features. Since the dense feature only need to be locally-discriminative, they can afford to capture much lower-level texture, leading to more accurate corre-spondences. Through extensive experiments we show that this results in more accurate camera poses and structure; in challenging conditions and for different local features.

While we optimize against dense feature maps, we keep the sparse scene representation of SfM. This ensures not only that the approach is scalable but also that the resulting 3D model is compatible with downstream applications, *e.g.* mapping for visual localization. Since our refinement works well even with few observations, as it does not need to average out the keypoint detection noise, it has the potential to achieve more accurate results using fewer images.

We thus believe that our approach can have a large impact in the localization community as it can improve the accuracy of the ground truth poses of standard benchmark datasets, of which many are currently saturated. Since this refinement is less sensitive to under-sampling, it enables benchmarking for crowd-sourced scenarios beyond densely-photographed tourism landmarks.

References

[1] CVPR 2020 Image Matching Challenge. https://www.cs.ubc.ca/research/image-matching-challenge/. Accessed March 1, 2021. 7

[2] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building Rome in a day. *Communications of the ACM*, 54(10):105–112, 2011. 1

[3] Sameer Agarwal, Keir Mierle, and Others. Ceres solver. http://ceres-solver.org. 5

[4] Sameer Agarwal, Noah Snavely, Steven M Seitz, and Richard Szeliski. Bundle adjustment in the large. In *ECCV*, 2010. 1

[5] Hatem Alismail, Brett Browning, and Simon Lucey. Photometric bundle adjustment for vision-based SLAM. In *ACCV*, 2016. 2, 4

[6] Hatem Alismail, Brett Browning, and Simon Lucey. Robust tracking in low light and sudden illumination changes. In *3DV*, 2016. 3

[7] Epameinondas Antonakos, Joan Alabort-i Medina, Georgios Tzimiropoulos, and Stefanos P Zafeiriou. Feature-based lucas–kanade and active appearance models. *IEEE Transactions on Image Processing*, 2015. 3

[8] Simon Baker, Ralph Gross, and Iain Matthews. Lucas-kanade 20 years on: A unifying framework. *IJCV*, 56, 2003. 2

[9] Daniel Barath, Dmytro Mishkin, Ivan Eichhardt, Ilia Shipachev, and Jiri Matas. Efficient initial pose-graph generation for global sfm. In *CVPR*, 2021. 2

[10] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded up robust features. In *ECCV*, 2006. 1

[11] Che-Han Chang, Chun-Nan Chou, and Edward Y Chang. CLKN: Cascaded Lucas-Lanade networks for image alignment. In *CVPR*, 2017. 3

[12] Avishek Chatterjee and Venu Madhav Govindu. Efficient and robust large-scale rotation averaging. In *CVPR*, 2013. 2

[13] Christopher B Choy, JunYoung Gwak, Silvio Savarese, and Manmohan Chandraker. Universal correspondence network. In *NIPS*, 2016. 2

[14] Peter Hviid Christiansen, Mikkel Fly Kragh, Yury Brodskiy, and Henrik Karstoft. UnsuperPoint: End-to-end unsupervised interest point detector and descriptor. *arXiv:1907.04011*, 2019. 2

[15] Ondřej Chum, Jiří Matas, and Josef Kittler. Locally optimized RANSAC. In *Joint Pattern Recognition Symposium*, pages 236–243. Springer, 2003. 6, 7

[16] Ondrej Chum, Tomas Werner, and Jiri Matas. Two-view geometry estimation unaffected by a dominant plane. In *CVPR*, 2005. 7

[17] Ronald Clark, Michael Bloesch, Jan Czarnowski, Stefan Leutenegger, and Andrew J. Davison. LS-Net: Learning to solve nonlinear least squares for monocular stereo. In *ECCV*, 2018. 3

[18] Jan Czarnowski, Stefan Leutenegger, and Andrew J. Davison. Semantic texture for robust dense tracking. In *ICCV Workshops*, 2017. 1, 2

[19] Amaël Delaunoy and Marc Pollefeys. Photometric bundle adjustment for dense multi-view 3d modeling. In *CVPR*, 2014. 2

[20] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *CVPR Workshop on Deep Learning for Visual SLAM*, 2018. 1, 2, 5, 6, 7

[21] Frédéric Devernay and Olivier D Faugeras. *Computing differential properties of 3-D shapes from stereoscopic images without 3-D models*. 1994. 2

[22] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-Net: A trainable CNN for joint detection and description of local features. In *CVPR*, 2019. 1, 5, 6, 7

[23] Mihai Dusmanu, Johannes L. Schönberger, and Marc Pollefeys. Multi-View Optimization of Local Feature Geometry. In *ECCV*, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[24] Ivan Eichhardt and Daniel Barath. Optimal multi-view correction of local affine frames. In *BMVC*, 2019. 2

[25] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *TPAMI*, 2017. 1, 2, 4, 5

[26] Jakob Engel, Thomas Schöps, and Daniel Cremers. LSD-SLAM: Large-scale direct monocular SLAM. In *ECCV*, 2014. 2

[27] Wolfgang Förstner and Eberhard Gülch. A fast operator for detection and precise location of distinct points, corners and centres of circular features. In *Proc. ISPRS intercommission conference on fast processing of photogrammetric data*, 1987. 2

[28] Jan-Michael Frahm, Pierre Fite-Georgel, David Gallup, Tim Johnson, Rahul Raguram, Changchang Wu, Yi-Hung Jen, Enrique Dunn, Brian Clipp, Svetlana Lazebnik, et al. Building Rome on a cloudless day. In *ECCV*, 2010. 1

[29] P. Georgel, Selim Benhimane, and Nassir Navab. A unified approach combining photometric and geometric information for pose estimation. In *BMVC*, 2008. 2

[30] Hugo Germain, Guillaume Bourmaud, and Vincent Lepetit. S2DNet: Learning accurate correspondences for sparse-to-dense feature matching. In *ECCV*, 2020. 1, 5, 7

[31] Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. *Robust statistics: the approach based on influence functions*. Wiley, 1986. 3

[32] Christopher G Harris, Mike Stephens, et al. A combined corner and edge detector. In *Alvey vision conference*, 1988. 1

[33] Jared Heinly, Johannes L Schonberger, Enrique Dunn, and Jan-Michael Frahm. Reconstructing the World* in Six Days *(as Captured by the Yahoo 100 Million Image Dataset). In *CVPR*, 2015. 1

[34] Heiko Hirschmüller, Peter R Innocent, and Jon Garibaldi. Real-time correlation-based stereo vision with reduced border errors. *IJCV*, 2002. 2

[35] Paul W Holland and Roy E Welsch. Robust regression using iteratively reweighted least-squares. *Communications in Statistics-theory and Methods*, 6(9):813–827, 1977. 5

[36] Danying Hu, Daniel DeTone, and Tomasz Malisiewicz. Deep ChArUco: Dark ChArUco Marker Pose Estimation. In *CVPR*, 2019. 2

[37] Andres Huertas and Gerard Medioni. Detection of intensity changes with subpixel accuracy using laplacian-gaussian masks. *TPAMI*, 1986. 2

[38] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, 2017. 2

[39] Arnold Irschara, Christopher Zach, Jan-Michael Frahm, and Horst Bischof. From structure-from-motion point clouds to fast location recognition. In *CVPR*, 2009. 1

[40] Yekeun Jeong, David Nister, Drew Steedly, Richard Szeliski, and In-So Kweon. Pushing the envelope of modern methods for bundle adjustment. *TPAMI*, 2011. 1, 5

[41] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image Matching across Wide Baselines: From Paper to Practice. *IJCV*, 2020. 7

[42] Christian Kerl, Jürgen Sturm, and Daniel Cremers. Dense visual slam for RGB-D cameras. In *IROS*, 2013. 2, 4

[43] Kenneth Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics*, 2(2):164–168, 1944. 5

[44] Xinghui Li, Kai Han, Shuda Li, and Victor Prisacariu. Dual-resolution correspondence networks. In *NeurIPS*, 2020. 2

[45] Yunpeng Li, Noah Snavely, Dan Huttenlocher, and Pascal Fua. Worldwide pose estimation using 3D point clouds. In *ECCV*, 2012. 1

[46] Ce Liu, Jenny Yuen, and Antonio Torralba. SIFT Flow: Dense correspondence across scenes and its applications. *TPAMI*, 2010. 2, 7

[47] Liu Liu, Hongdong Li, and Yuchao Dai. Efficient global 2D-3D matching for camera localization in a large-scale 3D map. In *ICCV*, 2017. 1

[48] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 1, 2, 5, 6, 7

[49] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, 1981. 1, 2

[50] Zhaoyang Lv, Frank Dellaert, James M. Rehg, and Andreas Geiger. Taking a deeper look at the inverse compositional algorithm. In *CVPR*, 2019. 3

[51] Daniel Martinec and Tomas Pajdla. Robust rotation and translation estimation in multiview reconstruction. In *CVPR*, 2007. 2

[52] Yuki Ono, Eduard Trulls, Pascal Fua, and Kwang Moo Yi. LF-Net: Learning local features from images. In *NeurIPS*, 2018. 2

[53] Seonwook Park, Thomas Schöps, and Marc Pollefeys. Illumination Change Robustness in Direct Visual SLAM. In *ICRA*, 2017. 3

[54] Rémi Pautrat, Viktor Larsson, Martin R Oswald, and Marc Pollefeys. Online invariance selection for local feature descriptors. In *ECCV*, 2020. 1

[55] Filip Radenovic, Johannes L Schonberger, Dinghuang Ji, Jan-Michael Frahm, Ondrej Chum, and Jiri Matas. From dusk till dawn: Modeling in the dark. In *CVPR*, 2016. 1

[56] Jerome Revaud, Philippe Weinzaepfel, César De Souza, Noe Pion, Gabriela Csurka, Yohann Cabon, and Martin Humenberger. R2D2: Repeatable and reliable detector and descriptor. In *NeurIPS*, 2019. 1, 5, 6

[57] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. Efficient neighbourhood consensus networks via submanifold sparse convolutions. In *ECCV*, 2020. 2

[58] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. In *NeurIPS*, 2018. 2

[59] Edward Rosten and Tom Drummond. Machine learning for high-speed corner detection. In *ECCV*, 2006. 1

[60] Paul-Edouard Sarlin. Visual localization made easy with hloc. https://github.com/cvg/Hierarchical-Localization/. 2

[61] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, 2019. 2

[62] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 1, 7

[63] Paul-Edouard Sarlin, Ajaykumar Unagar, Måns Larsson, Hugo Germain, Carl Toft, Viktor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik Kahl, and Torsten Sattler. Back to the Feature: Learning robust camera localization from pixels to pose. In *CVPR*, 2021. 3, 5, 7

[64] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Fredrik Kahl, and Tomas Pajdla. Benchmarking 6DOF outdoor visual localization in changing conditions. In *CVPR*, 2018. 2, 8

[65] Torsten Sattler, Tobias Weyand, Bastian Leibe, and Leif Kobbelt. Image retrieval for image-based localization revisited. In *BMVC*, 2012. 8

[66] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 2002. 2

[67] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 2, 6, 7, 8

[68] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016. 2

[69] Thomas Schops, Torsten Sattler, and Marc Pollefeys. BAD SLAM: Bundle Adjusted Direct RGB-D SLAM. In *CVPR*, June 2019. 2

[70] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, 2017. 2, 5

[71] Xi Shen, François Darmon, Alexei A Efros, and Mathieu Aubry. RANSAC-Flow: generic two-stage image alignment. In *ECCV*, 2020. 2

[72] Noah Snavely, Steven M Seitz, and Richard Szeliski. Modeling the world from internet photo collections. *IJCV*, 2008. 2

[73] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *CVPR*, 2018. 2

[74] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with Transformers. *CVPR*, 2021. 2

[75] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. InLoc: Indoor Visual Localization with Dense Matching and View Synthesis. *TPAMI*, 2019. 2

[76] Chengzhou Tang and Ping Tan. BA-Net: Dense bundle adjustment network. In *ICLR*, 2019. 3

[77] Jiexiong Tang, Hanme Kim, Vitor Guizilini, Sudeep Pillai, and Rares Ambrus. Neural outlier rejection for self-supervised keypoint learning. In *ICLR*, 2020. 2

[78] Engin Tola, Vincent Lepetit, and Pascal Fua. DAISY: An efficient dense descriptor applied to wide-baseline stereo. *TPAMI*, 2009. 2

[79] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment — a modern synthesis. In *International workshop on vision algorithms*, 1999. 1, 2, 3

[80] Prune Truong, Martin Danelljan, and Radu Timofte. GLU-Net: Global-local universal network for dense flow and correspondences. In *CVPR*, 2020. 2

[81] Michał J Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. In *NeurIPS*, 2020. 1

[82] Lukas Von Stumberg, Patrick Wenzel, Qadeer Khan, and Daniel Cremers. GN-Net: The Gauss-Newton loss for multi-weather relocalization. *RA-L*, 5(2):890–897, 2020. 3, 5

[83] Lukas Von Stumberg, Patrick Wenzel, Nan Yang, and Daniel Cremers. LM-Reloc: Levenberg-Marquardt based direct visual relocalization. In *3DV*, 2020. 3

[84] Oliver J Woodford and Edward Rosten. Large scale photometric bundle adjustment. In *BMVC*, 2020. 2, 4, 7

[85] Binbin Xu, Andrew J. Davison, and Stefan Leutenegger. Deep probabilistic feature-metric tracking. *RA-L*, 6(1):223–230, 2021. 3

[86] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth inference for unstructured multi-view stereo. *ECCV*, 2018. 2

[87] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent MVSNet for high-resolution multi-view stereo depth inference. In *CVPR*, 2019. 2

[88] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. LIFT: Learned invariant feature transform. In *ECCV*, 2016. 1

[89] Baosheng Yu and Dacheng Tao. Heatmap regression via randomized rounding. *arXiv:2009.00225*, 2020. 2

[90] Zehao Yu and Shenghua Gao. Fast-MVSNet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In *CVPR*, 2020. 3

[91] Zichao Zhang, Torsten Sattler, and Davide Scaramuzza. Reference Pose Generation for Long-term Visual Localization via Learned Features and View Synthesis. *IJCV*, 2020. 8

[92] Qunjie Zhou, Torsten Sattler, and Laura Leal-Taixe. Patch2Pix: Epipolar-guided pixel-level correspondences. In *CVPR*, 2021. 2