Grounding Image Matching in 3D with MASt3R

Vincent Leroy, Yohann Cabon, and Jerome Revaud

Naver Labs Europe
<firstname>.<lastname>@naverlabs.com

Abstract. Image Matching is a core component of all best-performing algorithms and pipelines in 3D vision. Yet despite matching being fundamentally a 3D problem, intrinsically linked to camera pose and scene geometry, it is typically treated as a 2D problem. This makes sense as the goal of matching is to establish correspondences between 2D pixel fields, but also seems like a potentially hazardous choice. In this work, we take a different stance and propose to cast matching as a 3D task with DUSt3R, a recent and powerful 3D reconstruction framework based on Transformers. Based on pointmaps regression, this method displayed impressive robustness in matching views with extreme viewpoint changes, yet with limited accuracy. We aim here to improve the matching capabilities of such an approach while preserving its robustness. We thus propose to augment the DUSt3R network with a new head that outputs dense local features, trained with an additional matching loss. We further address the issue of quadratic complexity of dense matching, which becomes prohibitively slow for downstream applications if not treated carefully. We introduce a fast reciprocal matching scheme that not only accelerates matching by orders of magnitude, but also comes with theoretical guarantees and, lastly, yields improved results. Extensive experiments show that our approach, coined MASt3R, significantly outperforms the state of the art on multiple matching tasks. In particular, it largely outperforms the best published methods on the challenging Map-free localization dataset.

1 Introduction

Being able to establish correspondences between pixels across different images of the same scene, denoted as *image matching*, constitutes a core component of all 3D vision applications, spanning mapping [15, 63], localization [43, 78], navigation [16], photogrammetry [35,68] and autonomous robotics in general [67, 92]. State-of-the-art methods for visual localization, for instance, overwhelmingly rely upon image matching during the offline mapping stage, *e.g.* using COLMAP [81], as well as during the online localization step, typically using PnP [31]. In this paper, we focus on this core task and aim at producing, given two images, a list of pairwise correspondences, denoted as *matches*. In particular, we seek to output highly accurate and dense matches that are robust to viewpoint and illumination changes because these are, in the end, the limiting factor for real-world applications [38].

In the past, matching methods have traditionally been cast into a three-steps pipeline consisting of first extracting sparse and repeatable keypoints, then describing them with locally invariant features, and finally pairing the discrete set of keypoints by comparing their distance in the feature space. This pipeline has several merits: keypoint detectors are precise under low-to-moderate illumination and viewpoint changes, and the sparsity of keypoints makes the problem computationally tractable, enabling very precise matching in milliseconds whenever the images are viewed under similar conditions. This explains the success and persistence of SIFT [54] in 3D reconstruction pipelines like COLMAP [81].

Unfortunately, keypoint-based methods, by reducing matching to a bag-of-keypoint problem, discard the global geometric context of the correspondence task. This makes them especially prone to errors in situation with repetitive patterns or low-texture areas, which are in fact ill-posed for local descriptors. One way to remedy this is to introduce a global optimization strategy during the pairing step, typically leveraging some learned priors about matching, which SuperGlue and similar methods successfully implemented [53, 78]. However, leveraging global context during matching might be too late, if keypoints and their descriptors do not already encode enough information. For this reason, another direction is to consider dense holistic matching, *i.e.* avoiding keypoints altogether, and matching the entire image at once. This recently became possible with the advent of mechanism for global attention [100]. Such approaches, like LoFTR [87], thus consider images as a whole and the resulting set of correspondences is dense and more robust to repetitive patterns and low-texture areas [45, 72, 73, 87]. This led to new state-of-the-art results on the most challenging benchmarks, such as the Map-free localization benchmark [5].

Nevertheless, even a top-performing methods like LoFTR [87] score a relatively disappointing VCRE precision of 34% on the Map-free localization benchmark. We argue that this is because, so far, practically all matching approaches have been treating matching as a 2D problem in image space. In reality, the formulation of the matching task is intrinsically and fundamentally a 3D problem: pixels that correspond are pixels that observe the same 3D point. Indeed, 2D pixel correspondences and a relative camera pose in 3D space are two sides of the same coin, as they are directly related by the epipolar matrix [38]. Another evidence is that the current top-performer on the Map-free benchmark is DUSt3R [106], a method initially designed for 3D reconstruction rather than matching, and for which matches are only a by-product of the 3D reconstruction. Yet, correspondences obtained naively from this 3D output currently outperform all other keypoint- and matching-based methods on the Map-free benchmark.

In this paper, we point out that, while DUSt3R [106] can indeed be used for matching, it is relatively imprecise, despite being extremely robust to viewpoint changes. To remedy this flaw, we propose to attach a second head that regresses dense local feature maps, and train it with an InfoNCE loss. The resulting architecture, called MASt3R for "Matching And Stereo 3D Reconstruction" outperforms DUSt3R on multiple benchmarks. To get pixel-accurate matches, we propose a coarse-to-fine matching scheme during which matching is performed at

several scales. Each matching step involves extracting reciprocal matches from dense feature maps which, perhaps counter-intuitively, is by far more time consuming than computing the dense feature maps themselves. Our proposed solution is a faster algorithm for finding reciprocal matches that is almost two orders of magnitude faster while improving the pose estimation quality.

To summarize, we claim three main contributions. First, we propose MASt3R, a 3D-aware matching approach building on the recently released DUSt3R framework. It outputs local feature maps that enable highly accurate and extremely robust matching. Second, we propose a coarse-to-fine matching scheme associated with a fast matching algorithm, enabling to work with high-resolution images. Third, MASt3R significantly outperform the state-of-the-art on several absolute and relative pose localization benchmarks.

2 Related works

Keypoint-based matching has been a cornerstone of computer vision. Matching is carried out in three distinct stages: keypoint detection, locally invariant description and nearest-neighbor search in descriptor space. Departing from the former handcrafted methods like SIFT [54,76], modern approaches have been shifting towards learning-based data-driven schemes for detecting keypoints [9,62,101,121], describing them [7, 34, 39, 93] or both at the same time [11, 21, 55, 56, 74, 102]. Overall, keypoint-based approaches are predominant in many benchmarks [7, 8, 37, 46, 80], underscoring their enduring value in tasks requiring high precision and speed [19, 80]. One notable issue, however, is they reduce matching to a local problem, *i.e.* discarding its holistic nature. SuperGlue and similar approaches [53, 78] thus propose to perform global reasoning in the last pairing step leveraging stronger priors to guide matching, yet leaving the detection and description local. While successful, it is still limited by the local nature of keypoints and their inability to remain invariant to strong viewpoint changes.

Dense matching. In contrast to keypoint-based approaches, semi-dense [12, 17, 45, 48, 87, 90] and dense approaches [28–30, 60, 97–99, 126] offer a different paradigm for establishing image correspondences, considering all possible pixel associations. Very reminiscent of optical flow approaches [23, 42, 44, 84, 85, 91], they are usually employing coarse-to-fine schemes to decrease computational complexity. Overall, these methods aim to consider matching from a global perspective, at the cost of increased computational resources. Dense matching has proven effective in scenarios where detailed spatial relationships and textures are critical for understanding scene geometry, leading to top performance on many benchmarks [4–6, 61, 78, 87] that are especially challenging for keypoints due to extreme changes in viewpoint or illumination. These approaches still cast matching as a 2D problem, which limits their usage for visual localization.

Camera Pose estimation techniques vary widely, but the most successful strategies, for speed, accuracy and robustness trade-off, are fundamentally based on pixel matching [77, 81, 109]. The constant improvement of matching methods has fostered the introduction of more challenging camera pose estimation bench-

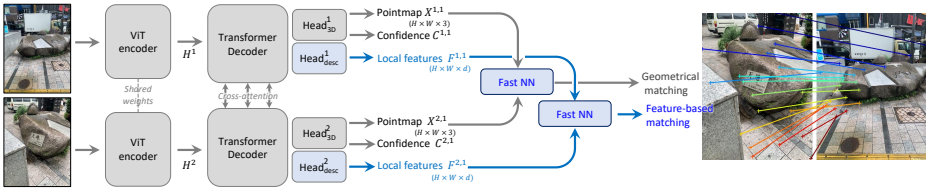

Fig. 1: Overview of the proposed approach. Given two input images to match, our network regresses for each image and each input pixel a 3D point, a confidence value and a local feature. Plugging either 3D points or local features into our fast reciprocal NN matcher (3.3) yields robust correspondences. Compared to the DUSt3R framework which we build upon, our contributions are highlighted in blue.

marks, such as Aachen Day-Night, InLoc, CO3D or Map-free [5, 71, 89, 123], all featuring strong viewpoint and/or illumination changes. The most challenging of them is undoubtedly Map-free [5], a localization dataset for which a single reference image is provided but no map, with viewpoint changes up to 180°.

Grounding matching in 3D thus becomes a crucial necessity in these challenging conditions where classical 2D-based matching utterly falls short. Leveraging priors about the physical properties of the scene in order to improve accuracy or robustness has been widely explored in the past, but most previous works settle for leveraging epipolar constraints for fully-supervised [22,36,64,103,114] or semi-supervised learning of correspondences without any fundamental change [10, 40, 49,105,112,116,118,124]. Toft *et al.* [94], on its part, propose to improve keypoint descriptors by rectifying images with perspective transformations obtained from an off-the-shelf monocular depth predictor. Recently, diffusion for pose [104] or rays [119], although not matching approaches strictly speaking, show promising performance by incorporating 3D geometric constraints into their pose estimation formulation. Finally, the recent DUSt3R [106] explore the possibility of recovering correspondences from the *a-priori* harder task of 3D reconstruction from uncalibrated images. Despite not being trained explicitly for matching, this approach yields promising results, topping the Map-free leaderboard [5]. Our contribution is to pursue this idea, by regressing local features and explicitly training them for pairwise matching.

3 Method

Given two images I^1 and I^2, respectively captured by two cameras C^1 and C^2 with unknown parameters, we wish to recover a set of pixel correspondences $\{(i, j)\}$ where i, j are pixels $i = (u_i, v_i), j = (u_j, v_j) \in \{1, \dots, W\} \times \{1, \dots, H\}$, W, H being the respective width and height of the images. We assume they have the same resolution for the sake of simplicity, yet without loss of generality. The final network can handle pairs of variable aspect ratios.

Our approach, illustrated in Fig. 1, aims at jointly performing 3D scene reconstruction and matching given two input images. It is based on the DUSt3R framework recently proposed by Wang *et al.* [106], which we first review in

Sec. 3.1 before presenting our proposed matching head and its corresponding loss in Sec. 3.2. We then introduce an optimized matching scheme specially devised to deal with dense feature maps in 3.3, that we use for coarse-to-fine matching in Sec. 3.4.

3.1 The DUSt3R framework

DUSt3R [106] is a recently proposed approach that jointly solves the calibration and 3D reconstruction problems from images alone. A transformer-based network predicts a *local* 3D reconstruction given two input images, in the form of two dense 3D point-clouds $X^{1,1}$ and $X^{2,1}$, denoted as *pointmaps* in the following. A pointmap $X^{a,b} \in \mathbb{R}^{H\times W\times 3}$ represents a dense 2D-to-3D mapping between each pixel $i = (u, v)$ of the image I^a and its corresponding 3D point $X^{a,b}_{u,v} \in \mathbb{R}^3$ expressed in the coordinate system of camera C^b. By regressing two pointmaps $X^{1,1}, X^{2,1}$ expressed in the *same* coordinate system of camera C^1, DUSt3R effectively solves the joint calibration and 3D reconstruction problem. In the case where more than two images are provided, a second step of global alignment merges all pointmaps in the same coordinate system. Note that, in this paper, we do not make use of this step and restrict ourselves to the binocular case. We now explain the inference in more details.

Both images are first encoded in a Siamese manner with a ViT encoder [24], yielding two representations H^1 and H^2, with $H^i = \text{Encoder}(I^i)$. Then, two intertwined decoders process these representations jointly, exchanging information via cross-attention to 'understand' the spatial relationship between viewpoints and the global 3D geometry of the scene. The new representations augmented with this spatial information are denoted as $H'^1, H'^2 = \text{Decoder}(H^1, H^2)$. Finally, two prediction heads regress the final pointmaps and confidence maps from the concatenated representations output by the encoder and decoder:

$$X^{1,1}, C^1 = \text{Head}^1_{\text{3D}}([H^1, H'^1]), \tag{1}$$

$$X^{2,1}, C^2 = \text{Head}^2_{\text{3D}}([H^2, H'^2]). \tag{2}$$

Regression loss. DUSt3R is trained in a fully-supervised manner using a simple regression loss

$$\ell_{\text{regr}}(v, i) = \left\| \frac{1}{z} X^{v,1}_i - \frac{1}{\hat{z}} \hat{X}^{v,1}_i \right\|, \tag{3}$$

where $v \in \{1, 2\}$ is the view and i is a pixel for which the ground-truth 3D point $\hat{X}^{v,1} \in \mathbb{R}^3$ is defined. In the original formulation, normalizing factors $z, \hat{z}$ are introduced to make the reconstruction invariant to scale. These are simply defined as the mean distance of all valid 3D points to the origin.

Metric predictions. In this work, we note that scale invariance is not necessarily desirable, as some potential use-cases like map-free visual localization necessitates metric-scale predictions. Therefore, we modify the regression loss to ignore normalization for the predicted pointmaps when the ground-truth pointmaps are known to be metric. That is, we set $z := \hat{z}$ whenever ground-truth is metric,

so that $\ell_{\text{regr}}(v,i) = ||X_i^{v,1} - \hat{X}_i^{v,1}||/\hat{z}$ in this case. As in DUSt3R [106], the final confidence-aware regression loss is defined as

$$\mathcal{L}_{\text{conf}} = \sum_{v\in\{1,2\}} \sum_{i\in\mathcal{V}^v} C_i^v \ell_{\text{regr}}(v,i) - \alpha \log C_i^v. \tag{4}$$

3.2 Matching prediction head and loss

To obtain reliable pixel correspondences from pointmaps, a standard solution is to look for reciprocal matches in some invariant feature space [27, 83, 106, 110]. While such a scheme works remarkably well with DUSt3R's regressed pointmaps (*i.e.* in a 3-dimensional space) even in presence of extreme viewpoint changes, we note that the resulting correspondences are rather imprecise, yielding suboptimal accuracy. This is a rather natural result as (i) regression is inherently affected by noise, and (ii) because DUSt3R was never explicitly trained for matching.

Matching head. For these reasons, we propose to add a second head that outputs two dense feature maps D^1 and $D^2 \in \mathbb{R}^{H\times W\times d}$ of dimensional d:

$$D^1 = \text{Head}^1_{\text{desc}}([H^1, H'^1]), \tag{5}$$
$$D^2 = \text{Head}^2_{\text{desc}}([H^2, H'^2]). \tag{6}$$

We implement the head as a simple 2-layers MLP interleaved with a non-linear GELU activation function [41]. Lastly, we normalize each local feature to unit norm. More details can be found in the supplementary material.

Matching objective. We wish to encourage each local descriptor from one image to match with at most a single descriptor from the other image that represents the same 3D point in the scene. To that aim, we leverage the infoNCE [65] loss over the set of ground-truth correspondences $\hat{\mathcal{M}} = \{(i,j)|\hat{X}_i^{1,1} = \hat{X}_j^{2,1}\}$:

$$\mathcal{L}_{\text{match}} = -\sum_{(i,j)\in\hat{\mathcal{M}}} \log \frac{s_\tau(i,j)}{\sum_{k\in\mathcal{P}^1} s_\tau(k,j)} + \log \frac{s_\tau(i,j)}{\sum_{k\in\mathcal{P}^2} s_\tau(i,k)}, \tag{7}$$

$$\text{with } s_\tau(i,j) = \exp\left[-\tau D_i^{1\top} D_j^2\right]. \tag{8}$$

Here, $\mathcal{P}^1 = \{i|(i,j) \in \hat{\mathcal{M}}\}$ and $\mathcal{P}^2 = \{j|(i,j) \in \hat{\mathcal{M}}\}$ denote the subset of considered pixels in each image and τ is a temperature hyper-parameter. Note that this matching objective is essentially a cross-entropy *classification* loss: contrary to regression in Eq. (3), the network is only rewarded if it gets the correct pixel right, not a nearby pixel. This strongly encourages the network to achieve high-precision matching. Finally, both regression and matching losses are combined to get the final training objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{conf}} + \beta\mathcal{L}_{\text{match}} \tag{9}$$

3.3 Fast reciprocal matching

Given two predicted feature maps $D^1, D^2 \in \mathbb{R}^{H \times W \times d}$, we aim to extract a set of reliable pixel correspondences, *i.e.* mutual nearest neighbors of each others:

$$\mathcal{M} = \{(i,j) \mid j = \mathrm{NN}_2(D_i^1) \text{ and } i = \mathrm{NN}_1(D_j^2)\}, \tag{10}$$

$$\text{with } \mathrm{NN}_A(D_j^B) = \arg\min_i \left\| D_i^A - D_j^B \right\|. \tag{11}$$

Unfortunately, naive implementation of reciprocal matching has a high computational complexity of $O(W^2H^2)$, since every pixel from an image must be compared to every pixels in the other image. While optimizing the nearest-neighbor (NN) search is possible, *e.g.* using K-d trees [1], this kind of optimization becomes typically very inefficient in high dimensional feature space and, in all cases, orders of magnitude slower than the inference time of MASt3R to output D^1 and D^2.

Fast matching. We therefore propose a faster approach based on sub-sampling. It is based on an iterated process that starts from an initial sparse set of k pixels $U^0 = \{U_n^0\}_{n=1}^k$, typically sampled regularly on a grid in the first image I^1. Each pixel is then mapped to its NN on I^2, yielding V^1, and the resulting pixels are mapped back again to I^1 in the same way:

$$U^t \longmapsto [\mathrm{NN}_2(D_u^1)]_{u \in U^t} \equiv V^t \longmapsto [\mathrm{NN}_1(D_v^2)]_{v \in V^t} \equiv U^{t+1} \tag{12}$$

The set of reciprocal matches (those which form a cycle, *i.e.* $\mathcal{M}_k^t = \{(U_n^t, V_n^t) \mid U_n^t = U_n^{t+1}\}$) are then collected. For the next iteration, pixels that already converged are filtered out, *i.e.* updating $U^{t+1} := U^{t+1} \setminus U^t$. Likewise, starting from $t = 1$ we also verify and filter V^{t+1}, comparing it with V^t in a similar fashion. As illustrated in Fig. 2 (left), this process is then iterated a fixed number of times, until most correspondences converge to stable (reciprocal) pairs. In Fig. 2 (center), we show that the number of un-converged point $|U^t|$ rapidly decreases to zero after a few iterations. Finally, the output set of correspondences consists of the concatenation of all reciprocal pairs $\mathcal{M}_k = \bigcup_t \mathcal{M}_k^t$.

Theoretical guarantees. The overall complexity of the fast matching is $O(kWH)$, which is $WH/k \gg 1$ times faster than the naive approach denoted *all*, as illustrated in Fig. 2 (right). It is worth pointing out that our fast matching algorithm extracts a *subset* of the full set $\mathcal{M}$, which is bounded in size by $|\mathcal{M}_k| \leq k$. We study in the supplementary material the convergence guarantees of this algorithm and how it evinces outlier-filtering properties, which explains why the end accuracy is actually *higher* than when using the full correspondence set $\mathcal{M}$, see Fig. 2 (right).

3.4 Coarse-to-fine matching

Due to the quadratic complexity of attention w.r.t. the input image area ($W \times H$), MASt3R only handles images of 512 pixels in their largest dimension. Larger images would require significantly more compute power to train, and ViTs do not

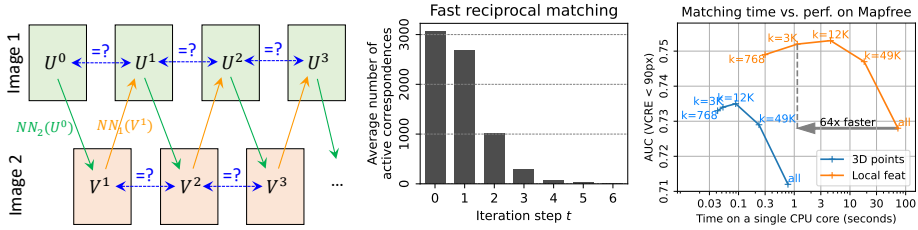

Fig. 2: Fast reciprocal matching. **Left**: Illustration of the fast matching process, starting from an initial subset of pixels U^0 and propagating it iteratively using NN search. Searching for cycles (blue arrows) detect reciprocal correspondences and allows to accelerate the subsequent steps, by removing points that converged. **Center**: Average number of remaining points in U^t at iteration $t = 1 \dots 6$. After only 5 iterations, nearly all points have already converged to a reciprocal match. **Right**: Performance-versus-time trade-off on the Map-free dataset. Performance actually improves, along with matching speed, when performing moderate levels of subsampling.

generalize yet to larger test-time resolutions [66, 69]. As a result, high-resolution images (*e.g.* 1M pixel) needs to be downscaled to be matched, afterwards the resulting correspondences are upscaled back to the original image resolution. This can lead to some performance loss, sometimes sufficient to cause substantial degradation in term of localization accuracy or reconstruction quality.

Coarse-to-fine matching is a standard technique to preserve the benefit of matching high-resolution images with a lower-resolution algorithm [70, 91]. We thus explore this idea for MASt3R. Our procedure starts with performing matching on downscaled versions of the two images. We denote the set of coarse correspondences obtained with subsampling k as $\mathcal{M}_k^0$. Next, we generate a grid of overlapping window crops W^1 and $W^2 \in \mathbb{R}^{w \times 4}$ on each full-resolution image independently. Each window crop measures 512 pixels in its largest dimension and contiguous windows overlap by 50%. We can then enumerate the set of all window pairs $(w_1, w_2) \in W^1 \times W^2$, from which we select a subset covering most of the coarse correspondences $\mathcal{M}_k^0$. Specifically, we add window pairs one by one in a greedy fashion until 90% of correspondences are covered. Finally, we perform matching for each window pair independently:

$$D^{w_1}, D^{w_2} = \text{MASt3R}(I^1_{w_1}, I^2_{w_2}) \tag{13}$$

$$\mathcal{M}_k^{w_1,w_2} = \text{fast\_reciprocal\_NN}(D^{w_1}, D^{w_2}) \tag{14}$$

Correspondences obtained from each window pair are finally mapped back to the original image coordinates and concatenated, thus providing dense full-resolution matches.

4 Experimental results

We detail in Sec. 4.1 the training procedure of MASt3R. Then, we evaluate on several tasks, each time comparing with the state of the art, starting with visual camera pose estimation on the Map-Free Relocalization Benchmark [5]

(Sec. 4.2), the CO3D and RealEstate datasets (Sec. 4.3) and other standard Visual Localization benchmarks in Sec. 4.4. Finally, we leverage MASt3R for Dense Multi-View Stereo (MVS) reconstruction in Sec. 4.5.

4.1 Training

Training data . We train our network with a mixture of 14 datasets: Habitat [79], ARKitScenes [20], Blended MVS [115], MegaDepth [50], Static Scenes 3D [59], ScanNet++ [117], CO3D-v2 [71], Waymo [88], Map-free [5], WildRgb [2], VirtualKitti [13], Unreal4K [96], TartanAir [107] and an internal dataset. These datasets feature diverse scene types: indoor, outdoor, synthetic, real-world, object-centric, etc. Among them, 10 datasets have metric ground-truth. When image pairs are not directly provided with the dataset, we extract them based on the method described in [108]. Specifically, we utilize off-the-shelf image retrieval and point matching algorithms to match and verify image pairs.

Training. We base our model architecture on the public DUSt3R model [106] and use the same backbone (ViT-Large encoder and ViT-Base decoder). To benefit the most from DUSt3R's 3D matching abilities, we initialize the model weights to the publicly available DUSt3R checkpoint. During each epoch, we randomly sample 650k pairs equally distributed between all datasets. We train our network for 35 epoch with a cosine schedule and initial learning rate set to 0.0001. Similar to [106], we randomize the image aspect ratio at training time, ensuring that the largest image dimension is 512 pixels. We set the local feature dimension to $d = 24$ and the matching loss weight to $\beta = 1$. It is important that the network sees different scales at training time, because coarse-to-fine matching starts from zoomed-out images to then zoom-in on details (see Sec. 3.4). We therefore perform aggressive data augmentation during training in the form of random cropping. Image crops are transformed with a homography to preserve the central position of the principal point.

Correspondence sampling. To generate ground-truth correspondences necessary for the matching loss (Eq. (7)), we simply find reciprocal correspondences between on the ground-truth 3D pointmaps $\hat{X}^{1,1} \leftrightarrow \hat{X}^{2,1}$. We then randomly subsample 4096 correspondences per image pairs. If we cannot find enough correspondences, we pad with random false correspondences so that the likelihood of finding a true match remains constant.

Fast nearest neighbors. For the fast reciprocal matching from Sec. 3.3, we implement the nearest neighbor function $\mathrm{NN}(x)$ from Eq. (11) differently depending on the dimension of x. When matching 3D points $x \in \mathbb{R}^3$, we implement $\mathrm{NN}(x)$ using K-d trees [58]. For matching local features with $d = 24$, however, K-d trees become highly inefficient due to the curse of dimensionality [26]. Therefore, we rely on the optimized FAISS library [25, 47] in this case.

4.2 Map-free localization

Dataset description. We start our experiments with the Map-free relocalization benchmark [5], an extremely challenging dataset aiming at localizing the camera

in metric space given a single reference image without any map. It comprises a training, validation and test sets of 460, 65 and 130 scenes resp., each featuring two video sequences. Following the benchmark, we evaluate in term of Virtual Correspondence Reprojection Error (VCRE) and camera pose accuracy, see [5] for details.

Impact of subsampling. We do not resort to coarse-to-fine matching for this dataset, as the image resolution is already close to MASt3R working resolution (720×540 vs. 512×384 resp.). As mentioned in Sec. 3.3, computing dense reciprocal matching is prohibitively slow even with optimized code for searching nearest neighbors. We therefore resort to subsampling the set of reciprocal correspondences, keeping at most k correspondences from the complete set $\mathcal{M}$ (Eq. (10)). Fig. 2 (right) shows the impact of subsampling in term of AUC (VCRE) performance and timing. Surprisingly, the performance significantly *improves* for intermediate values of subsampling. Using $k = 3000$, we can accelerate matching by a factor of 64 while significantly improving the performance. We provide insights in the supplementary material regarding this phenomenon. Unless stated otherwise, we keep $k = 3000$ for subsequent experiments.

Ablations on losses and matching modes. We report results on the validation set in Tab. 1 for different variants of our approach: DUSt3R matching 3D points (I); MASt3R also matching 3D points (II) or local features (III, IV, V). For all methods, we compute the relative pose from the essential matrix [38] estimated with the set of predicted matches (PnP performs similarly). The metric scene scale is inferred from the depth extracted with an off-the-shelf DPT finetuned on KITTI [69] (I-IV) or from the depth directly output by MASt3R (V).

First, we note that all proposed methods significantly outperforms the DUSt3R baseline, probably because MASt3R is trained longer and with more data. All other things being equal, matching descriptors perform significantly better than matching 3D points (II versus IV). This confirms our initial analysis that regression is inherently unsuited to compute pixel correspondences, see Sec. 3.2.

We also study the impact of training only with a single matching objective ($\mathcal{L}_{\text{match}}$ from Eq. (7), III). In this case, the performance overall degrades compared to training with both 3D and matching losses (IV), in particular in term of pose estimation accuracy (*e.g.* median rotation of 10.8° for (III) compared to 3.0° for (IV)). We point out that this is in spite of the decoder now having *more capacity to carry out a single task*, instead of two when performing 3D reconstruction simultaneously, indicating that grounding matching in 3D is indeed crucial to improve matching. Lastly, we observe that, when using metric depth directly output by MASt3R, the performance largely improves. This suggests that, as for matching, the depth prediction task is largely correlated with 3D scene understanding, and that the two tasks strongly benefit from each other.

Comparisons on the test set is reported in Tab. 2. Overall, MASt3R outperforms all state-of-the-art approaches by a large margin, achieving more than 93% in VCRE AUC. This is a 30% absolute improvement compared to the second best published method, LoFTR+KBR [86,87], that get 63.4% in AUC. Likewise, the median translation error is vastly reduced to 36cm, compared to approx. 2m for

Table 1: Results on the *validation* set of the Map-free dataset. (**First** and second best)

	match	depth	VCRE (<90px) Reproj. ↓	Prec. ↑	AUC ↑	Pose Error (<25cm, 5°) Median Err.↓	Precision ↑	AUC ↑
(I) DUSt3R	3d	DPT	125.8 px	45.2%	0.704	1.10m 9.4°	17.0%	0.344
(II) MASt3R	3d	DPT	112.0 px	49.9%	0.732	0.94m 3.6°	21.5%	0.409
(III) MASt3R-M	feat	DPT	107.7 px	51.7%	0.744	1.10m 10.8°	19.3%	0.382
(IV) MASt3R	feat	DPT	112.9 px	51.5%	0.752	0.93m 3.0°	23.2%	0.435
(V) MASt3R	feat	(auto)	**57.2** px	**75.9%**	**0.934**	**0.46m 3.0°**	**51.7%**	**0.746**

Table 2: Comparison with the state of the art on the *test* set of the Map-free dataset.

	depth	VCRE (<90px) Reproj. ↓	Prec. ↑	AUC ↑	Pose Error (<25cm, 5°) Median Err.↓	Precision ↑	AUC ↑
RPR [5]	DPT	147.1 px	40.2%	0.402	1.68m 22.5°	6.0%	0.060
SIFT [54]	DPT	222.8 px	25.0%	0.504	2.93m 61.4°	10.3%	0.252
SP+SG [78]	DPT	160.3 px	36.1%	0.602	1.88m 25.4°	16.8%	0.346
LoFTR [87]	KBR	165.0 px	34.3%	0.634	2.23m 37.8°	11.0%	0.295
FAR [75]	(auto)	137.0 px	44.2%	0.680	1.48m 17.2°	17.7%	0.392
RoMa [29]	DPT	128.8 px	45.6%	0.669	1.23m 11.1°	**22.8%**	**0.407**
Mickey [8]	(auto)	129.5 px	49.3%	**0.748**	1.66m 27.3°	13.3%	0.325
DUSt3R [106]	DPT	**116.0 px**	**50.3%**	0.697	**0.97m 7.1°**	21.6%	0.394
MASt3R	DPT	104.0 px	54.2%	0.726	0.80m **2.2°**	27.0%	0.456
MASt3R	(auto)	**48.7 px**	**79.3%**	0.933	**0.36m 2.2°**	**54.7%**	0.740
MASt3R (direct reg.)		53.2 px	79.1%	**0.941**	0.42m 3.1°	53.0%	**0.777**

the state-of-the-art methods. A large part of the improvement is of course due to MASt3R predicting metric depth, but note that our variant leveraging depth from DPT-KITTI (thus purely matching-based) outperforms all state-of-the-art approaches as well.

We also provide the results of direct regression with MASt3R, *i.e.* without matching, simply using PnP on the pointmap $X^{2,1}$ of the second image. These results are surprisingly on par with our matching-based variant, even though the ground-truth calibration of the reference camera is not used. As we show below, this does not hold true for other localization datasets, and computing the pose via matching (*e.g.* with PnP or essential matrix) with known intrinsics seems safer in general.

Qualitative results. We show in Fig. 3 some matching results for pairs with strong viewpoint change (up to 180°). We also highlight with insets some specific regions that are correctly matched by MASt3R in spite of drastic appearance changes. We believe these correspondences to be nearly impossible to get with 2D-based matching methods. In contrast, grounding the matching in 3D allows to solve the issue relatively straightforwardly.

4.3 Relative pose estimation

Datasets and protocol. Next, we evaluate for the task of relative pose estimation on the CO3Dv2 [71] and RealEstate10k [125] datasets. CO3Dv2 contains 6 million frames extracted from approximately 37k videos, covering 51 MS-COCO

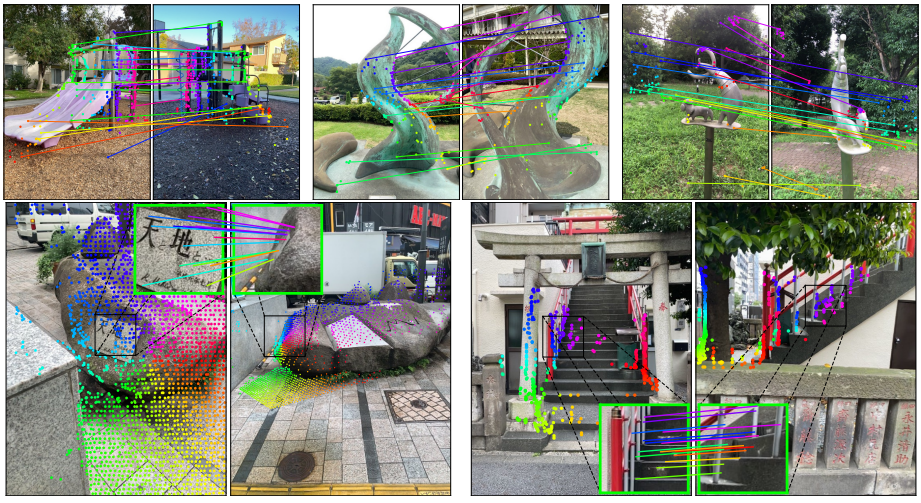

Fig. 3: Qualitative examples on the Map-free dataset. **Top row**: Pairs with strong viewpoint changes. Third one is a failure case. For clarity, we only draw a subset of all correspondences. **Bottom row**: We highlight interesting spots in close-up. These regions could hardly be matched by local keypoints. See text for details.

categories. Ground-truth camera poses are obtained using COLMAP [81] from 200 frames in each video. RealEstate10k is an indoor/outdoor dataset that features 80K video clips on YouTube totalling 10 million frames, camera poses being obtained via SLAM with bundle adjustment. Following [104], we evaluate MASt3R on 41 categories from CO3Dv2 and 1.8K video clips from the test set of RealEstate10k. Each sequence is 10 frames long, we evaluate relative camera poses between all possible 45 pairs, not using ground-truth focals.

Baselines and metrics. As before, matches obtained with MASt3R are used to estimate Essential Matrices and relative pose. Please note that our predictions are always done pairwise, contrary to all other methods that leverage multiple views (at the exception of DUSt3R-PnP). We compare to recent data-driven approaches like RelPose [120], RelPose++ [120], PoseReg and PoseDiff [104], the recent RayDiff [119] and DUSt3R [106]. We also report results for more traditional SfM methods like PixSFM [52] and COLMAP [82] extended with SuperPoint [21] and SuperGlue [78] (COLMAP+SPSG). Similar to [104], we report the Relative Rotation Accuracy (RRA) and Relative Translation Accuracy (RTA) for each image pair to evaluate the relative pose error and select a threshold $\tau = 15$ to report RTA@15 and RRA@15. Additionally, we calculate the mean Average Accuracy (mAA30), defined as the area under the accuracy curve of the angular differences at $min(\text{RRA@30}, \text{RTA@30})$.

Results. As shown in Tab. 3, SfM approaches tend to perform significantly worse on this task, mainly due to the poor visual support. This because images usually observe a small object, combined with the fact that many pairs have a wide baseline, sometimes up to 180°. On the contrary, 3D grounded approaches like RayDiffusion, DUSt3R and MASt3R are the two most competitive methods

Table 3: *Left*: Multi-view pose regression on the CO3Dv2 [71] and RealEstate10K [125] with 10 random frames. Parenthesis () denote methods that do not report results on the 10 views set, we report their best for comparison (8 views). We distinguish between (a) multi-view and (b) pairwise methods. *Right:* Dense MVS results on the DTU dataset, in *mm*. Handcrafted methods (c) perform worse than learning-based approaches (d) that train on this specific domain. Among the methods that operate in a zero-shot setting (e), MASt3R is the only one attaining reasonable performance.

	Methods	Co3Dv2			RealEstate10K
		RRA@15	RTA@15	mAA(30)	mAA(30)
(a)	Colmap+SG [21, 78]	36.1	27.3	25.3	45.2
	PixSfM [52]	33.7	32.9	30.1	49.4
	RelPose [120]	57.1	-	-	-
	PosReg [104]	53.2	49.1	45.0	-
	PoseDiff [104]	80.5	79.8	66.5	48.0
	RelPose++ [51]	(85.5)	-	-	-
	RayDiff [119]	(93.3)	-	-	-
	DUSt3R-GA [106]	**96.2**	86.8	76.7	67.7
(b)	DUSt3R [106]	94.3	88.4	77.2	61.2
	MASt3R	94.6	**91.9**	**81.8**	**76.4**

	Methods	Acc.↓	Comp.↓	Overall↓
(c)	Camp [14]	0.835	0.554	0.695
	Furu [32]	0.613	0.941	0.777
	Tola [95]	0.342	1.190	0.766
	Gipuma [33]	**0.283**	0.873	0.578
(d)	MVSNet [114]	0.396	0.527	0.462
	CVP-MVSNet [113]	0.296	0.406	0.351
	UCS-Net [18]	0.338	0.349	0.344
	CER-MVS [57]	0.359	0.305	0.332
	CIDER [111]	0.417	0.437	0.427
	PatchmatchNet [103]	0.427	0.277	0.352
	CasMVSNet [36]	0.325	0.385	0.355
	TransMVSNet [22]	0.321	0.289	0.305
	GeoMVSNet [122]	0.331	**0.259**	**0.295**
(e)	DUSt3R [106]	2.677	0.805	1.741
	MASt3R	0.403	0.344	0.374

on this dataset, the latter leading in translation and mAA on both datasets. Notably, on RealEstate our mAA score improves by at least 8.7 points over the best multi-view methods and 15.2 points over pairwise DUSt3R. This showcases the accuracy and robustness of our approach to few input view setups.

4.4 Visual localization

Datasets. We then evaluate MASt3R for the task of absolute pose estimation on the Aachen Day-Night [123] and InLoc [89] datasets. Aachen comprises 4,328 reference images taken with hand-held cameras, as well as 824 daytime and 98 nighttime query images taken with mobile phones in the old inner city of Aachen, Germany. InLoc [89] is an indoor dataset with challenging appearance variation between the 9,972 RGB-D + 6DOF pose database images and the 329 query images taken from an iPhone 7.

Metrics. We report report the percentage of successfully localized images within three thresholds: (0.25m, 2°), (0.5m, 5°) and (5m, 10°) for Aachen and (0.25m, 10°), (0.5m, 10°), (1m, 10°) for InLoc.

Results are reported in Table 4. We study the performance of MASt3R with variable number of retrieved images. As expected, a greater number of retrieved images (top40) yields better performance, achieving competitive performance on Aachen and significantly outperforming the state of the art on InLoc. Interestingly, our approach still performs very well even with a single retrieved image (top1), showcasing the robustness of 3D grounded matching. We also include direct regression results, which are rather poor, showing a striking impact of the dataset scale on the localization error, *i.e.* small scenes are much less affected (see results on Map-free in 4.2). This confirms the importance of feature matching to estimate reliable poses.

Table 4: Visual localization results on Aachen Day-Night and InLoc. We report our results for different number of retrieved database images (topN).

Methods	AachenDayNight [123]		InLoc [89]	
	Day	Night	DUC1	DUC2
Kapture+R2D2 [43]	**91.3/97.0/99.5**	**78.5/91.6/100**	41.4/60.1/73.7	47.3/67.2/73.3
SP+SuperGlue [78]	89.8/96.1/99.4	77.0/90.6/**100**	49.0/68.7/80.8	53.4/77.1/82.4
SP+LightGlue [53]	90.2/96.0/99.4	77.0/91.1/**100**	49.0/68.2/79.3	55.0/74.8/79.4
LoFTR [87]	88.7/95.6/99.0	**78.5**/90.6/99.0	47.5/72.2/84.8	54.2/74.8/85.5
DKM [28]	-	-	51.5/75.3/86.9	63.4/82.4/87.8
DUSt3R top1 [106]	72.7/89.6/98.1	59.7/80.1/93.2	36.4/55.1/66.7	27.5/42.7/49.6
DUSt3R top20 [106]	79.4/94.3/**99.5**	74.9/91.1/99.0	53.0/74.2/89.9	61.8/77.1/84.0
MASt3R top1	79.6/93.5/98.7	70.2/88.0/97.4	41.9/64.1/73.2	38.9/55.7/62.6
MASt3R top20	83.4/95.3/99.4	76.4/**91.6/100**	55.1/77.8/90.4	**71.0**/84.7/89.3
MASt3R top40	82.2/93.9/**99.5**	75.4/**91.6/100**	**56.1/79.3/90.9**	**71.0/87.0/91.6**
MASt3R direct reg. top1	1.5/4.5/60.7	1.6/4.2/47.6	13.1/32.3/58.1	10.7/26.0/38.2

4.5 Multiview 3D reconstruction

We finally perform MVS by triangulating the obtained matches. Note that the matching is performed in full resolution without prior knowledge of cameras, and the latter are only used to triangulate matches to 3D in ground-truth reference frame. To remove spurious 3D points, we simply apply geometric consistency post-processing [103].

Datasets and metrics. We evaluate our predictions on the DTU [3] dataset. Contrary to all competing learning methods, we apply our network in a zero-shot setting, *i.e.* we do not train nor finetune on the DTU train set and apply our model as is. In Tab. 3 we report the average accuracy, completeness and Chamfer distances error metrics as provided by the authors of the benchmarks. The accuracy for a point of the reconstructed shape is defined as the smallest Euclidean distance to the ground-truth, and the completeness of a point of the ground-truth as the smallest Euclidean distance to the reconstructed shape. The overall Chamfer distance is the average of both previous metrics.

Results. Data-driven approaches trained on this domain significantly outperform handcrafted ones, cutting the Chamfer error by half. To the best of our knowledge, we are the first to draw such conclusion in a zero-shot setting. MASt3R not only outperforms the DUSt3R baseline but also compete with the best methods, all without leveraging camera calibration nor poses for matching, neither having seen this camera setup before.

5 Conclusion

Grounding image matching in 3D with MASt3R significantly raised the bar on camera pose and localization tasks on many public benchmarks. We successfully improved DUSt3R with matching, getting the best of both worlds: enhanced robustness, while attaining and even surpassing what could be done with pixel matching alone. We introduced a fast reciprocal matcher and a coarse to fine approach for efficient processing, allowing users to balance between accuracy and speed. MASt3R is able to perform in few-view regimes (even in top1), that we believe will greatly increase versatility of localization.

References

1. Scipy. https://docs.scipy.org/doc/scipy
2. RGBD Objects in the Wild: Scaling Real-World 3D Object Learning from RGB-D Videos (2024), http://arxiv.org/abs/2401.12592, arXiv:2401.12592 [cs]
3. Aanæs, H., Jensen, R.R., Vogiatzis, G., Tola, E., Dahl, A.B.: Large-scale data for multiple-view stereopsis. IJCV (2016)
4. Addison, H., Eduard, T., etru1927, Kwang Moo, Y., old ufo, Sohier, D., Yuhe, J.: Image matching challenge 2022 (2022), https://kaggle.com/competitions/image-matching-challenge-2022
5. Arnold, E., Wynn, J., Vicente, S., Garcia-Hernando, G., Monszpart, Á., Prisacariu, V.A., Turmukhambetov, D., Brachmann, E.: Map-free visual relocalization: Metric pose relative to a single image. In: ECCV (2022)
6. Ashley, C., Eduard, T., HCL-Jevster, Kwang Moo, Y., lcmrll, old ufo, Sohier, D., tanjigou, WastedCode, Weiwei, S.: Image matching challenge 2023 (2023), https://kaggle.com/competitions/image-matching-challenge-2023
7. Balntas, V., Lenc, K., Vedaldi, A., Mikolajczyk, K.: HPatches: A benchmark and evaluation of handcrafted and learned local descriptors. In: CVPR (2017)
8. Barroso-Laguna, A., Munukutla, S., Prisacariu, V.A., Brachmann, E.: Matching 2d images in 3d: Metric relative pose from metric correspondences. In: CVPR (2024)
9. Barroso-Laguna, A., Riba, E., Ponsa, D., Mikolajczyk, K.: Key.Net: Keypoint Detection by Handcrafted and Learned CNN Filters. In: ICCV (2019)
10. Bhalgat, Y., Henriques, J.F., Zisserman, A.: A light touch approach to teaching transformers multi-view geometry. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023 (2023)
11. Bhowmik, A., Gumhold, S., Rother, C., Brachmann, E.: Reinforced feature points: Optimizing feature detection and description for a high-level task. In: CVPR (2020)
12. Bökman, G., Kahl, F.: A case for using rotation invariant features in state of the art feature matchers. In: CVPRW (2022)
13. Cabon, Y., Murray, N., Humenberger, M.: Virtual KITTI 2. CoRR **abs/2001.10773** (2020)
14. Campbell, N.D.F., Vogiatzis, G., Hernández, C., Cipolla, R.: Using multiple hypotheses to improve depth-maps for multi-view stereo. In: ECCV (2008)
15. Campos, C., Elvira, R., Rodríguez, J.J.G., M. Montiel, J.M., D. Tardós, J.: Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. IEEE Transactions on Robotics (2021)
16. Chaplot, D.S., Gandhi, D., Gupta, S., Gupta, A., Salakhutdinov, R.: Learning to explore using active neural slam. arXiv preprint arXiv:2004.05155 (2020)
17. Chen, H., Luo, Z., Zhou, L., Tian, Y., Zhen, M., Fang, T., McKinnon, D., Tsin, Y., Quan, L.: Aspanformer: Detector-free image matching with adaptive span transformer. European Conference on Computer Vision (ECCV) (2022)
18. Cheng, S., Xu, Z., Zhu, S., Li, Z., Li, L.E., Ramamoorthi, R., Su, H.: Deep stereo using adaptive thin volume representation with uncertainty awareness. In: CVPR (2020)
19. Csurka, G., Dance, C., Humenberger, M.: From Handcrafted to Deep Local Invariant Features. arXiv **1807.10254** (2018)

20. Dehghan, A., Baruch, G., Chen, Z., Feigin, Y., Fu, P., Gebauer, T., Kurz, D., Dimry, T., Joffe, B., Schwartz, A., Shulman, E.: ARKitScenes: A diverse real-world dataset for 3d indoor scene understanding using mobile RGB-D data. In: NeurIPS Datasets and Benchmarks (2021)
21. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised Interest Point Detection and Description. In: CVPR (2018)
22. Ding, Y., Yuan, W., Zhu, Q., Zhang, H., Liu, X., Wang, Y., Liu, X.: Transmvsnet: Global context-aware multi-view stereo network with transformers. In: CVPR (2022)
23. Dong, Q., Cao, C., Fu, Y.: Rethinking optical flow from geometric matching consistent perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
24. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)
25. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.E., Lomeli, M., Hosseini, L., Jégou, H.: The faiss library (2024)
26. Duda, R., Hart, P., G.Stork, D.: Pattern Classification (01 2001)
27. Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T.: D2-net: A trainable CNN for joint description and detection of local features. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 8092–8101. Computer Vision Foundation / IEEE (2019). https://doi.org/10.1109/CVPR.2019.00828, http://openaccess.thecvf.com/content_CVPR_2019/html/Dusmanu_D2-Net_A_Trainable_CNN_for_Joint_Description_and_Detection_of_CVPR_2019_paper.html
28. Edstedt, J., Athanasiadis, I., Wadenbäck, M., Felsberg, M.: DKM: Dense kernelized feature matching for geometry estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (2023)
29. Edstedt, J., Sun, Q., Bökman, G., Wadenbäck, M., Felsberg, M.: RoMa: Robust Dense Feature Matching. arXiv preprint arXiv:2305.15404 (2023)
30. Efe, U., Ince, K.G., Alatan, A.: Dfm: A performance baseline for deep feature matching. In: CVPRW (2021)
31. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6), 381–395 (1981). https://doi.org/10.1145/358669.358692, https://doi.org/10.1145/358669.358692
32. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multiview stereopsis. PAMI (2010)
33. Galliani, S., Lasinger, K., Schindler, K.: Massively parallel multiview stereopsis by surface normal diffusion. In: ICCV (June 2015)
34. Germain, H., Bourmaud, G., Lepetit, V.: S2DNet: Learning image features for accurate sparse-to-dense matching. In: ECCV (2020)
35. Gomes, L., Bellon, O.R.P., Silva, L.: 3d reconstruction methods for digital preservation of cultural heritage: A survey. Pattern Recognit. Lett. (2014)
36. Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P.: Cascade cost volume for high-resolution multi-view stereo and stereo matching. In: CVPR (2020)
37. Hammarstrand, L., Kahl, F., Maddern, W., Pajdla, T., Pollefeys, M., Sattler, T., Sivic, J., Stenborg, E., Toft, C., Torii, A.: Long-Term Visual Localization Benchmark. https://www.visuallocalization.net/

38. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press (2004). https://doi.org/10.1017/CBO9780511811685, https://doi.org/10.1017/cbo9780511811685
39. He, K., Lu, Y., Sclaroff, S.: Local descriptors optimized for average precision. In: CVPR (2018)
40. He, Y., Yan, R., Fragkiadaki, K., Yu, S.: Epipolar transformers. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020 (2020)
41. Hendrycks, D., Gimpel, K.: Bridging nonlinearities and stochastic regularizers with gaussian error linear units. CoRR **abs/1606.08415** (2016), http://arxiv.org/abs/1606.08415
42. Huang, Z., Shi, X., Zhang, C., Wang, Q., Cheung, K.C., Qin, H., Dai, J., Li, H.: Flowformer: A transformer architecture for optical flow. In: ECCV (2022)
43. Humenberger, M., Cabon, Y., Guerin, N., Morat, J., Revaud, J., Rerole, P., Pion, N., de Souza, C., Leroy, V., Csurka, G.: Robust image retrieval-based visual localization using kapture (2020)
44. Jiang, S., Campbell, D., Lu, Y., Li, H., Hartley, R.I.: Learning to estimate hidden motions with global motion aggregation (2021)
45. Jiang, W., Trulls, E., Hosang, J., Tagliasacchi, A., Yi, K.M.: COTR: Correspondence Transformer for Matching Across Images. In: ICCV (2021)
46. Jin, Y., Mishkin, D., Mishchuk, A., Matas, J., Fua, P., Yi, K., Trulls, E.: Image Matching across Wide Baselines: From Paper to Practice. IJCV (2020)
47. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. IEEE Transactions on Big Data **7**(3), 535–547 (2019)
48. Junjie, N., Yijin, L., Zhaoyang, H., Hongsheng, L., Hujun, B., Zhaopeng, C., Guofeng, Z.: Pats: Patch area transportation with subdivision for local feature matching. In: CVPR (2023)
49. Kloepfer, D.A., Henriques, J.F., Campbell, D.: SCENES: Subpixel Correspondence Estimation With Epipolar Supervision (2024), http://arxiv.org/abs/2401.10886
50. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: CVPR. pp. 2041–2050 (2018)
51. Lin, A., Zhang, J.Y., Ramanan, D., Tulsiani, S.: Relpose++: Recovering 6d poses from sparse-view observations. CoRR **abs/2305.04926** (2023)
52. Lindenberger, P., Sarlin, P., Larsson, V., Pollefeys, M.: Pixel-perfect structure-from-motion with featuremetric refinement. In: ICCV (2021)
53. Lindenberger, P., Sarlin, P., Pollefeys, M.: Lightglue: Local feature matching at light speed. In: ICCV (2023)
54. Lowe, D.: Distinctive Image Features from Scale-invariant Keypoints. IJCV (2004)
55. Luo, Z., Zhou, L., Bai, X., Chen, H., Zhang, J., Yao, Y., Li, S., Fang, T., Quan, L.: Aslfeat: Learning local features of accurate shape and localization. In: CVPR (2020)
56. Ma, J., Jiang, X., Fan, A., Jiang, J., Yan, J.: Image matching from handcrafted to deep features: A survey. IJCV (2021)
57. Ma, Z., Teed, Z., Deng, J.: Multiview stereo with cascaded epipolar raft. In: ECCV (2022)
58. Maneewongvatana, S., Mount, D.M.: Analysis of approximate nearest neighbor searching with clustered point sets. In: DIMACS. DIMACS Series in Discrete Mathematics and Theoretical Computer Science (1999)

59. Mayer, N., Ilg, E., Häusser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: CVPR (2016)
60. Melekhov, I., Tiulpin, A., Sattler, T., Pollefeys, M., Rahtu, E., Kannala, J.: DGC-Net: Dense geometric correspondence network. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV) (2019)
61. Mishkin, D., Matas, J., Perdoch, M., Lenc, K.: Wxbs: Wide baseline stereo generalizations. In: Xie, X., Jones, M.W., Tam, G.K.L. (eds.) BMVC (2015)
62. Mishkin, D., Radenovic, F., Matas, J.: Repeatability is not enough: Learning affine regions via discriminability. In: ECCV (2018)
63. Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: Orb-slam: a versatile and accurate monocular slam system. IEEE transactions on robotics (2015)
64. Na, Y., Kim, W.J., Han, K.B., Ha, S., Yoon, S.E.: Uforecon: Generalizable sparse-view surface reconstruction from arbitrary and unfavorable sets. In: CVPR (2024)
65. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. CoRR **abs/1807.03748** (2018), `http://arxiv.org/abs/1807.03748`
66. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P., Li, S., Misra, I., Rabbat, M.G., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
67. Özyeşil, O., Voroninski, V., Basri, R., Singer, A.: A survey of structure from motion*. Acta Numerica **26**, 305–364 (2017)
68. Peppa, M., Mills, J., Fieber, K., Haynes, I., Turner, S., Turner, A., Douglas, M., Bryan, P.: Archaeological feature detection from archive aerial photography with a sfm-mvs and image enhancement pipeline. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences **42**, 869–875 (2018)
69. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV (2021)
70. Ranjan, A., Black, M.J.: Optical flow estimation using a spatial pyramid network. In: CVPR (2017)
71. Reizenstein, J., Shapovalov, R., Henzler, P., Sbordone, L., Labatut, P., Novotný, D.: Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In: ICCV (2021)
72. Revaud, J., Weinzaepfel, P., Harchaoui, Z., Schmid, C.: EpicFlow: Edge-Preserving Interpolation of Correspondences for Optical Flow. In: CVPR (2015)
73. Revaud, J., Weinzaepfel, P., Harchaoui, Z., Schmid, C.: DeepMatching: Hierarchical deformable dense matching. IJCV (2016)
74. Revaud, J., Weinzaepfel, P., de Souza, C.R., Humenberger, M.: R2D2: repeatable and reliable detector and descriptor. In: NIPS (2019)
75. Rockwell, C., Kulkarni, N., Jin, L., Park, J.J., Johnson, J., Fouhey, D.F.: Far: Flexible, accurate and robust 6dof relative camera pose estimation (2024)
76. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.R.: ORB: an efficient alternative to SIFT or SURF. In: ICCV (2011)
77. Sarlin, P.E., Cadena, C., Siegwart, R., Dymczyk, M.: From coarse to fine: Robust hierarchical localization at large scale. In: CVPR (2019)
78. Sarlin, P., DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperGlue: Learning feature matching with graph neural networks. In: CVPR (2020)

79. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., Parikh, D., Batra, D.: Habitat: A platform for embodied ai research. In: ICCV (2019)
80. Schönberger, J.L., Hardmeier, H., Sattler, T., Pollefeys, M.: Comparative Evaluation of Hand-Crafted and Learned Local Features. In: CVPR (2017)
81. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
82. Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M.: Pixelwise view selection for unstructured multi-view stereo. In: ECCV (2016)
83. Sethi, I.K., Jain, R.C.: Finding trajectories of feature points in a monocular image sequence. IEEE TPAMI (1987)
84. Shi, X., Huang, Z., Bian, W., Li, D., Zhang, M., Cheung, K.C., See, S., Qin, H., Dai, J., Li, H.: Videoflow: Exploiting temporal cues for multi-frame optical flow estimation. In: ICCV (2023)
85. Shi, X., Huang, Z., Li, D., Zhang, M., Cheung, K.C., See, S., Qin, H., Dai, J., Li, H.: Flowformer++: Masked cost volume autoencoding for pretraining optical flow estimation. In: CVPR (2023)
86. Spencer, J., Russell, C., Hadfield, S., Bowden, R.: Kick back & relax++: Scaling beyond ground-truth depth with slowtv & cribstv. In: ArXiv Preprint (2024)
87. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: LoFTR: Detector-free local feature matching with transformers. CVPR (2021)
88. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset. In: CVPR (June 2020)
89. Taira, H., Okutomi, M., Sattler, T., Cimpoi, M., Pollefeys, M., Sivic, J., Pajdla, T., Torii, A.: InLoc: Indoor Visual Localization with Dense Matching and View Synthesis. PAMI (2019)
90. Tang, S., Zhang, J., Zhu, S., Tan, P.: Quadtree attention for vision transformers. ICLR (2022)
91. Teed, Z., Deng, J.: RAFT: recurrent all-pairs field transforms for optical flow. In: ECCV (2020)
92. Thrun, S.: Probabilistic robotics. Communications of the ACM **45**(3), 52–57 (2002)
93. Tian, Y., Yu, X., Fan, B., Wu, F., Heijnen, H., Balntas, V.: Sosnet: Second order similarity regularization for local descriptor learning. In: CVPR (2019)
94. Toft, C., Turmukhambetov, D., Sattler, T., Kahl, F., Brostow, G.J.: Single-image depth prediction makes feature matching easier. In: ECCV (2020)
95. Tola, E., Strecha, C., Fua, P.: Efficient large-scale multi-view stereo for ultra high-resolution image sets. Mach. Vis. Appl. (2012)
96. Tosi, F., Liao, Y., Schmitt, C., Geiger, A.: Smd-nets: Stereo mixture density networks. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
97. Truong, P., Danelljan, M., Gool, L.V., Timofte, R.: Learning accurate dense correspondences and when to trust them. In: CVPR (2021)
98. Truong, P., Danelljan, M., Timofte, R.: GLU-Net: Global-local universal network for dense flow and correspondences. In: CVPR (2020)
99. Truong, P., Danelljan, M., Timofte, R., Gool, L.V.: Pdc-net+: Enhanced probabilistic dense correspondence network. IEEE TPAMI (2023)

100. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) NeurIPS (2017)
101. Verdie, Y., Yi, K.M., Fua, P., Lepetit, V.: TILDE: A temporally invariant learned detector. In: CVPR (2015)
102. Wang, B., Chen, C., Cui, Z., Qin, J., Lu, C.X., Yu, Z., Zhao, P., Dong, Z., Zhu, F., Trigoni, N., Markham, A.: P2-net: Joint description and detection of local features for pixel and point matching. In: ICCV (2021)
103. Wang, F., Galliani, S., Vogel, C., Speciale, P., Pollefeys, M.: Patchmatchnet: Learned multi-view patchmatch stereo. In: CVPR. pp. 14194–14203 (2021)
104. Wang, J., Rupprecht, C., Novotny, D.: PoseDiffusion: Solving pose estimation via diffusion-aided bundle adjustment (2023)
105. Wang, Q., Zhou, X., Hariharan, B., Snavely, N.: Learning Feature Descriptors using Camera Pose Supervision. In: ECCV (2020)
106. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy (2023)
107. Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., Scherer, S.: Tartanair: A dataset to push the limits of visual slam (2020)
108. Weinzaepfel, P., Lucas, T., Leroy, V., Cabon, Y., Arora, V., Brégier, R., Csurka, G., Antsfeld, L., Chidlovskii, B., Revaud, J.: CroCo v2: Improved Cross-view Completion Pre-training for Stereo Matching and Optical Flow. In: ICCV (2023)
109. Wu, C.: VisualSFM: A Visual Structure from Motion System. `http://ccwu.me/vsfm/` (2011)
110. Wu, H., Sankaranarayanan, A.C., Chellappa, R.: Cvpr (2007)
111. Xu, Q., Tao, W.: Learning inverse depth regression for multi-view stereo with correlation cost volume. In: AAAI (2020)
112. Yang, G., Malisiewicz, T., Belongie, S.J.: Learning data-adaptive interest points through epipolar adaptation. In: CVPR Workshops (2019)
113. Yang, J., Mao, W., Álvarez, J.M., Liu, M.: Cost volume pyramid based depth inference for multi-view stereo. In: CVPR. pp. 4876–4885 (2020)
114. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. In: ECCV (2018)
115. Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., Quan, L.: Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In: CVPR (2020)
116. Yao, Y., Jafarian, Y., Park, H.S.: MONET: multiview semi-supervised keypoint detection via epipolar divergence. In: ICCV (2019)
117. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: Proceedings of the International Conference on Computer Vision (ICCV) (2023)
118. Yifan, W., Doersch, C., Arandjelovic, R., Carreira, J., Zisserman, A.: Input-level inductive biases for 3d reconstruction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022 (2022)
119. Zhang, J.Y., Lin, A., Kumar, M., Yang, T.H., Ramanan, D., Tulsiani, S.: Cameras as rays: Pose estimation via ray diffusion. In: International Conference on Learning Representations (ICLR) (2024)
120. Zhang, J.Y., Ramanan, D., Tulsiani, S.: Relpose: Predicting probabilistic relative rotation for single objects in the wild. In: ECCV (2022)

121. Zhang, X., Yu, F.X., Karaman, S., Chang, S.: Learning discriminative and transformation covariant local feature detectors. In: CVPR (2017)
122. Zhang, Z., Peng, R., Hu, Y., Wang, R.: Geomvsnet: Learning multi-view stereo with geometry perception. In: CVPR (2023)
123. Zhang, Z., Sattler, T., Scaramuzza, D.: Reference pose generation for long-term visual localization via learned features and view synthesis. IJCV (2021)
124. Zhou, Q., Sattler, T., Leal-Taixe, L.: Patch2pix: Epipolar-guided pixel-level correspondences. In: CVPR (2021)
125. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. SIGGRAPH (2018)
126. Zhu, S., Liu, X.: Pmatch: Paired masked image modeling for dense geometric matching. In: CVPR (2023)